Inversion breakpoints in Atlantic cod chromosomes – fixed or variable within and between individuals?

Tara Jane Daughton

Master of Science Thesis

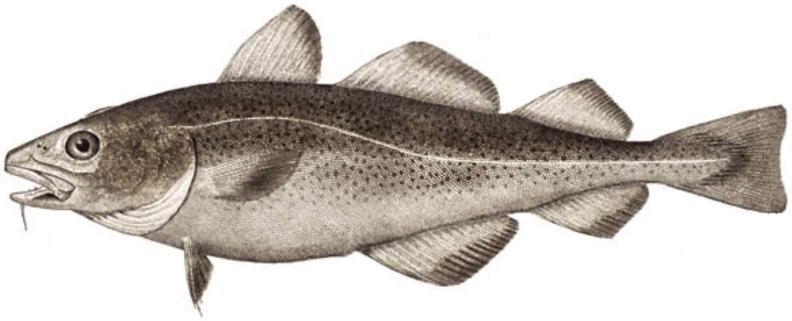
60 credits

Centre for Ecological and Evolutionary Synthesis
Department of Biosciences
Faculty of Mathematics and Natural Sciences
UNIVERSITY OF OSLO

August 2022

Inversion breakpoints in Atlantic cod chromosomes – fixed or variable within and between individuals?

Tara Jane Daughton



Inversion breakpoint in Atlantic cod chromosomes – fixed or variable within and between individuals?
Tara Jane Daughton
https://www.duo.uio.no/
Print: Reprosentralen, Universitetet i Oslo
Front page illustration credit: https://no.wikipedia.org/wiki/Fil:Atlantic_cod.jpg

Acknowledgments

First and foremost, I want to share my most profound appreciation for my supervisors. Professor Kjetill Sigurd Jakobsen, for giving me a warm welcome to the Centre of Ecological and Evolutionary Synthesis (CEES) and introducing me to the exciting world of research. Anna Zofia Komisarczuk, for teaching me the way of the lab and for all your insight and invaluable feedback. Helle Tessand Baalsrud, for guiding the way with uplifting talks and your motivation. Marine Brieuc, for lending me a helping hand with your knowledge and skills.

I also want to thank Ole Kristian Tørresen for giving me a crash course in bioinformatics and always being available to help. And Ave Tooming-Klunderud from the Norwegian Sequencing Centre, for lighting up my day with your positivity and always cheering me on.

I want to extend my gratitude to everyone at CEES. Thank you for all the kindness you have shown me over the last two years and for all the laughs. And a special thanks to my office buddies in room 3124, I am proud of you all, and I am looking forward to many years of friendship.

And finally, the most important people in my life. My family and closest friends. My dearest dad, *og min kjæreste mor*, I can do anything in this world because I have your love and support. My brothers, Daniel and Eliot, who have no idea what I´ve been up to but still manage to hype me up anytime, anywhere. I promise to start watching One Piece again and catch up. And, of course, my best friends, who I cherish with all my heart. You have all my love and gratitude. Thank you.

The sequencing service was provided by the Norwegian Sequencing Centre (www.sequencing.uio.no), a national technology platform hosted by the University of Oslo and supported by the "Functional Genomics" and "Infrastructure" programs of the Research Council of Norway and the Southeastern Regional Health Authorities". Sequencing costs were covered by the Department of Biosciences (UiO)

TONE

Abstract

Genomic variation is the key to speciation and evolution. Genomic variants underlie how species change, adapt, and evolve and range from single-point mutations to large chromosomal rearrangements. Inversions are an example of the latter and can profoundly affect local adaptation, such as with the Atlantic cod. The Atlantic cod that inhabit the northernmost coast of Norway consists of two ecotypes: the stationary Northern Coastal Cod (NCC) and the migratory Northeast Arctic Cod (NEAC), which feeds offshore in the Barents Sea and only returns to the Northern coast to breed. NCC and NEAC breed at the same time and locations along the coast of Norway, from Møre in the south to Sørøya in the north, with the largest spawning aggregation taking place around the Lofoten archipelago. Even though they spawn simultaneously, phenotypic, and genetic differences are maintained between the two populations. From previous studies, four megabase-scale supergenes have been linked to migratory lifestyle and environmental adaptations, and I set out to locate the breakpoints and determine if they vary between individuals.

To do so, I developed a PCR protocol for amplifying both inverted and non-inverted alleles to sequence the breakpoint regions within the Atlantic cod. In total, 79 breakpoint regions were sequenced using HiFi-sequencing, which creates accurate long reads known as HiFi reads. In doing so, I could use multiple, accurate, HiFi sequencing reads to (1) align each breakpoint read to the gadmor3 and coastal reference genome and then (2) make *de novo* haplotype assemblies for the breakpoint regions for different individuals with the inverted and non-inverted allele. The haplotype assemblies were used in a multiple sequence alignment, and I could pinpoint where the breakpoints are and that they are conserved and fixed within and between populations.

Innholdsfortegnelse

1. Introduction

What is the cause of phenotypic differences between individuals, populations, ecotypes, and sub-species? Within any given species there is a large span of phenotypic variation, which is a product of genomic variation and environmental variation. In some species, different populations are locally adapted to their environment. Gene flow between populations tends to erode genomic differences between them and thus hinder local adaptation (Woodruff, 2001). Despite this, there are examples of local adaptation evolving in the face of gene flow, such as the co-existing ecotypes of Atlantic cod (*Gadus morhua)* that inhabit the north coast of Norway (Rodríguez-Ramilo et al., 2019). The Atlantic cod that inhabit the northernmost coast of Norway consists of two ecotypes: the stationary Northern Coastal Cod (NCC) that lives by the coast all its life (coastal), and the migratory Northeast Arctic Cod (NEAC) that feeds offshore in the Barents Sea and only returns to the Northern coast to breed. NCC and NEAC breed at the same time and locations along the coast of Norway, from Møre in the south to Sørøya in the north, with the largest spawning aggregation taking place around the Lofoten archipelago (Figure 3). There is a high-level of gene-flow and therefore a low level of genome-wide divergence between the two ecotypes, which has puzzled biologists for decades. In their 2016 study, Berg and colleagues (Berg et al., 2016) revealed that the local adaptation between stationary (NCC) and migratory (NEAC) in cod is driven by chromosomal inversions, which form islands of genomic divergence in a sea of genomic connectivity between these populations.

1.1 Chromosomal inversions

Genomic variation is the key to speciation and evolution. It is how species change, adapt, and evolve (Sætre & Ravinet, 2019). A mechanism behind genomic variation are mutations, and mutations occur at random. Most mutations are neutral or deleterious, but some are beneficial (Futuyma & Kirkpatrick, 2017). Mutations can range from single nucleotide polymorphisms (SNPs) to structural variations (SVs) that are a 50bp or larger genomic alterations, such as insertions, deletions, duplications, inversions, and other rearrangements (Futuyma & Kirkpatrick, 2017). These genomic variations are the reasons behind the different phenotypes we observe, and the fate of any new phenotype is determined by natural selection and genetic drift. The pattern of genomic variation within a species can tell a lot about the forces of mutation, recombination, genetic drift, and natural selection, as well as the demographic history of a population. Sometimes we see an emergence of a phenotypic trait within a species that cannot be explained by one single point mutation. This sudden mutation that has resulted in a distinct phenotypic trait could be the result an inversion.

Chromosomal inversions are large-scale genomic mutations that result in a 180° rotation of a chromosome segment, and this rotation hinders recombination between the inverted and the non-inverted segment in heterozygous individuals. The lack of recombination leads to the inverted segment (derived) being isolated from the ancestral segments, and the inverted sequence is therefore protected from gene flow (Sætre & Ravinet, 2019) Inversions can thus maintain genomic regions highly differentiated between populations despite low overall genetic structure, genome-wide, due to high levels of gene flow (Noor et al., 2007) . By linking together co-adapted alleles, inversions can work as a driver for local adaptation. They can also have a crucial role in developing complex phenotypes caused by multiple genes by acting as 'super-genes' (Matschiner et al., 2022; Wellenreuther & Bernatchez, 2018) which may further differentiate populations and their ecotypes (i.e. promote local adaptation). Super-genes are a cluster of physically linked genes that are inherited as a single unit (Black & Shuker, 2019). Through selection and genetic drift, the inverted and non-inverted variants will diverge, which can, in turn, have important evolutionary implications. It is important to note that chromosomal rearrangement is still possible when double crossing-over events occur, but at the center of the inversion rather than near the breakpoints (Villoutreix et al., 2021). However, double crossing-over events are rare. The lack of recombination between the inverted and non-inverted variants will keep beneficial allelic combinations together, creating super-genes (Matschiner et al., 2022). Even though chromosomal inversion has been known for nearly a century, it is quite a challenge to understand their origin. Super-genes are loci that can create differences in colour, morphology, sexual compatibility, etc, within the same species. There are two hypotheses for how and why chromosomal inversions occur and super-genes are evolved (Villoutreix et al., 2021). One, inversions are selected for due to their effect on suppressing recombination between sets of epistatic or locally adapted genes. Two, inversions create adaptive mutations at their breakpoints, leading to their rise in frequency via selection on breakpoint variants. These hypotheses may seem similar, but they imply different evolutionary histories for supergenes.

When an inversion occurs, there are two breakpoints on each side of the sequence, and the middle segment is flipped and reinserted (Figure 1). There are different ways inversions can lead to phenotypic differences. One is that the breakpoint could disrupt a gene, altering its expression, and if the mutation is beneficial, it can be spread by positive selection (Futuyma & Kirkpatrick, 2017). Another way of inversion being spread in a population is through heterozygous recombination (Kirkpatrick & Barrett, 2015). When an inversion is heterozygous, it blocks recombination in the inverted area, which results in favorable gene combinations of alleles being inherited together. This ensures that the favorable combination of alleles is passed on together. Lastly, inversion can spread through genetic drift (Barth et al., 2019).

Recent studies indicate that inversions play a crucial role in eco-evolutionary processes from mating choice to social behavior and environmental adaptations (Wellenreuther & Bernatchez, 2018). The same species can have different banding patterns of genes in the chromosome as a result from inversions (Futuyma & Kirkpatrick, 2017). In other words, different individuals within the same species will have different genes in proximity that are inherited together. This can lead to phenotypic differences, such as with the male ruff (*Calidris pugnax*) (Küpper et al., 2016) and in the seaweed lies (*Coleopa frigida*) (Berdan et al., 2021). The male ruff has three strikingly different mating morphs; the aggressive independent behavior, the semi cooperative satellites and the female-mimic faeders (Lank et al., 2013). These distinct different mating behaviors are linked with an inversion on chromosome 11 that contain about 100 genes, where the breakpoint of the inversion disrupts the reading frame of an important gene (CENP-N) (Küpper et al., 2016). Those who are homozygotes for the inversion die at an early age, while those who are heterozygous are either satellites or faeders (Küpper et al., 2016). Each of these distinct behaviours have their own benefit when it comes to reproductive strategy and has therefore been maintained by frequency-dependent selection. Likewise, the seaweed fly (*Coelopa frigida)* harbours a large chromosomal inversion system called Cf-Inv(1) that is made up of three overlapping inversions. This inversion influence body size, development time, and viability (Black & Shuker, 2019). Of the traits mentioned, size is the trait where the inversion has the strongest effect (Berdan et al., 2021) . Cf-Inv(1) has two highly divergent arrangements, termed as α and β. The males who are homozygous for α are approximately threefold heavier than those who are homozygous for β. Consequently, αα males takes significantly longer to reach adulthood than ββ males (Butlin et al., 1982). The female seaweed fly seems mostly unaffected by karyotype, other than a small effect on size. The female and male seaweed fly largely share the same genome, indicating a particular sex-specific role for gene expression on the inversion (Berdan et al., 2021).

When an inversion is introgressed (i.e. transferred from one species or population into the gene pool of another species or population), it can be a powerful mechanism for range expansion, and one speculates is that introgression of an inversion can accelerates adaptations by crossing species boundaries (Kirkpatrick & Barrett, 2015). An example of where introgression of a chromosomal inversion has had a distinct phenotypic impact is the Amazonian butterfly (*Helicounis numata)*. Within this species there are seven different wing-pattern morphs that coexist, each one matching to near perfection the colour and shaped of the toxic Lepidoptera (*Helicniinae, Danainae, Pericopiinae)* (Jay et al., 2018).

Whilst the potential for gene flow is especially high for marine organisms, as there are few or no significant physical barriers, we do see evidence of genetic population structure being driven by structural variation, such as inversions, which could work as a driver for local environmental adaptations (Wellenreuther & Bernatchez, 2018). A prime example of this is the Atlantic cod (Barth et al., 2019; Berg et al., 2017; Berg et al., 2016).

1.2 Atlantic cod

The Atlantic cod has played a significant role for the Norwegian economy. Before it was ever called cod, it was just known as fish as there was no need to specify it any more than that. No other fishes have had quite the impact on the world as the Atlantic cod as it is a fish of all seasons, for all people.

We can find the Atlantic cod across the continental shelves of the North Atlantic, as well as the Baltic Sea. Consequently, the cod distributed in these areas experience a huge variance in temperature, all from -1,3 to 19,4 °C (Rose, 2019). Atlantic cod are eurythermal, meaning they can withstand large temperature shifts. This is necessary for the Atlantic cod as they can experience temperature varying with 10 °C in a single day. However, the cod physiology is profoundly influenced by temperature, meaning that a cod living at 0°C is quite different from a cod living at 12°C.

Atlantic cod consists of several ecotypes, and in this thesis, I focus on the local, stationary Norwegian Coastal Cod (NCC) and the migratory North-East Arctic Cod (NEAC – 'skrei') (Figure 1). Both NCC and NEAC spawn in the same areas each winter/early spring along the northern Norwegian coast, as far south as Møre, with the waters off the Lofoten islands being the largest spawning grounds (Michalsen et al., 2008). In addition to migrating from the cold feedings ground of the Barents Sea to the warmer spawning areas along the coast of Norway, the NEAC perform vertical movements down to depths of about 500 meters, while the stationary NCC stay in shallow waters. Even though they both spawn simultaneously, phenotypic and genetic differences is maintained between the two populations (Berg et al., 2016). From previous studies, four megabase-scale supergenes (Figure 2) have been linked to migratory lifestyle and environmental adaptations (Berg et al., 2017; Berg et al., 2016; Matschiner et al., 2022).

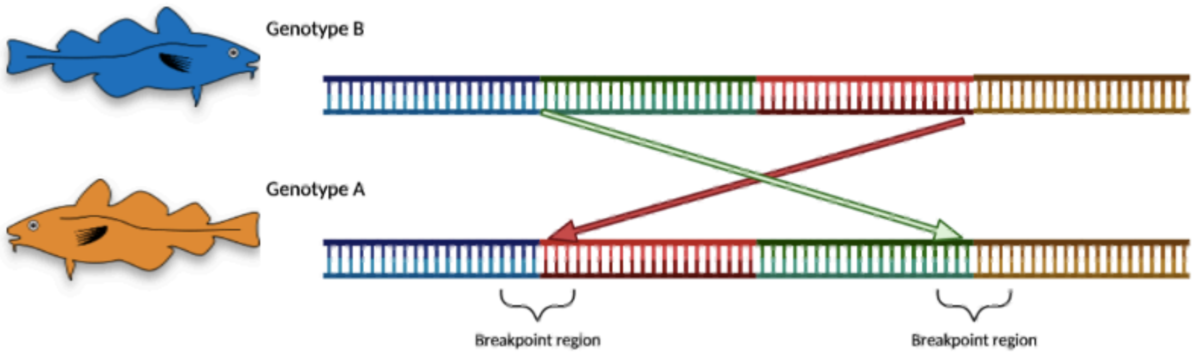
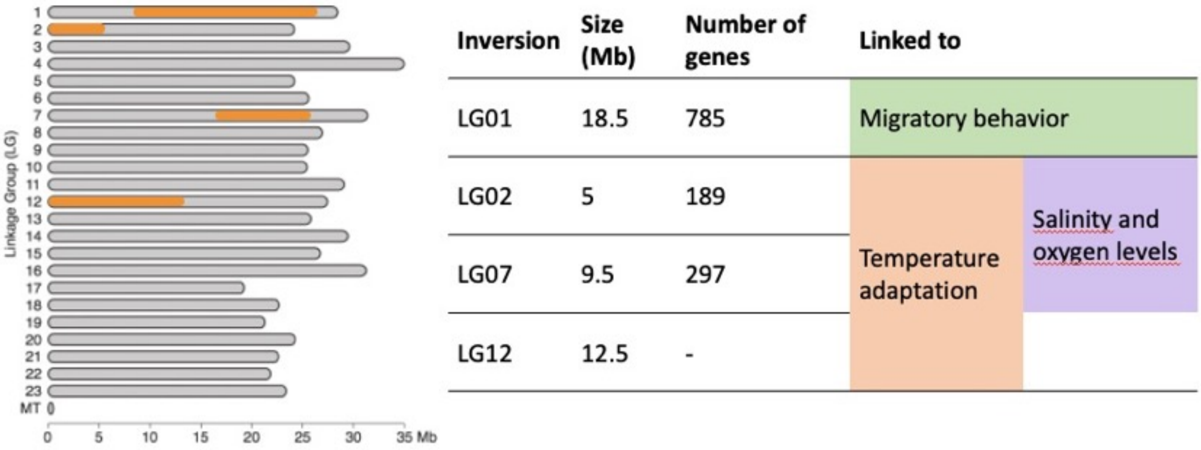


Figure 1*. Genotype B describes the inversion allele B that is more common in NEAC (migratory), while genotype A describes the inverted allele A that is more prevalent in NCC (stationary). There has been a 180 ° flip of the sequence from the ancestral allele B, resulting in the derived inverted allele A. As a result, there are two breakpoints on the inverted allele.*

Four large chromosomal inversions have been identified in linkage group (LG) LG 1, LG 2, LG 7 and LG 12 from studies showing large blocks of SNPs in high linkage disequilibrium (LD) and elevated F_{ST} values (in large continuous regions of each LG) in comparisons between NCC and NEAC (Rodríguez-Ramilo et al., 2019) (Figure 2). The Atlantic cod reference genome -denoted gadmor3 – is based on a NEAC individual (NCBI accession ID: GCF_902167405.1). From here on out, the inversion conformation in NEAC is known as B and the alternative allele is known as A. The inversions have been linked to adaptation to the local environment and behaviour of each ecotype (Barth et al., 2017; Berg et al., 2017; Berg et al., 2016). What is interesting is that chromosomal inversion can act as supergenes, that are associated with the migratory and stationary ecotypes between the NCC and NEAC. Furthermore, supergenes are associated with adaptations to different salinity levels and temperature change (Matschiner et al., 2022). A study from Barth and colleagues found through genomic analyses an overrepresentation of the chromosomal rearrangement in fjord cod on LG 2, known to contain genes linked to adaptation to low salinity (Barth et al., 2017). This suggests that through segregation of chromosomal rearrangements, recombination is suppressed, and essential functional genes are inherited together, which can be genes that are locally adapted for the fjord environment (Berg et al., 2017). More than 800 individual cod from across its geographical range have been sequenced using a short-read technology as part of the AquaGenome Project (https://www.aquagenome.uio.no), and the distributions of inversion genotypes have been identified. These inversions are old, with the oldest to be close to 900 000 years old. The individual LG varies in age from 900 000 to 600 000 – thus, they representing ancient independent evolutionary events in ancient populations of cod (Matschiner et al., 2022). However, we lack more detailed knowledge of the inversions, especially the inversion breakpoints. By comparing cod genomes with the genome of a closely related species, the haddock (*Melanogrammus aeglefinus*), Matchiner and colleagues were able to determine whether the A or B allele is the ancestral variant. Given the different age and ancestral states of the inversions at LG 1, LG 2 and LG 7 it seems plausible to suggest that these are independent events (Matschiner et al., 2022).

It has previously been shown that the double inversion on LG 1 on the migratory NEAC and the stationary NCC cods hinder recombination within heterozygotes which prevents introgression (Kirubakaran et al., 2016). The breakpoints appear in regions rich in repeats, which are difficult to resolve using short (Illumina) reads – at the population level. At present, we do not know if the inversion breakpoints are conserved across individuals and populations or if they are variable. This question is essential to answer because variable inversion breakpoints can affect recombination between individuals and suggest crossing over events occurring at different rates at the population level. Using highly accurate long-sequencing to generate high fidelity reads (PacBio-HiFi-sequencing) (https://www.pacb.com/technology/hifi-sequencing), we can create a more reliable resolution around the breakpoints, which can shed light on whether the exact breakpoints are variable, or not, in cod populations.



Inversion	Size (Mb)	Number of genes	Linked to	
LG01	18.5	785	Migratory behavior	
LG02	5	189	Temperature adaptation	Salinity and oxygen levels
LG07	9.5	297		
LG12	12.5	-		

Figure 2*. Previous studies have linked the inversions on the different linkage groups to different phenotypes. Here is an overview of the inversion, coloured in orange, with their respective size, number of genes, and what phenotype they are linked to.*

Locally adapted species are likely irreplaceable, yet some local Atlantic cod populations are collapsing with increased human activity (Mieszkowska et al., 2009). During the last century, there has been a dramatic decline in abundance across their biogeographic range. One can debate how much climatic warming and overfishing drive these changes. Nevertheless, there is a need for population genetics data combined with population ecology for evaluating the effects of climate change and commercial harvesting. Even though the Atlantic cod has for hundreds of years been a critical determinant of the wealth of human populations on both sides of the North Atlantic, it was not until the 1990s we started studying how environmental variables impacted the Atlantic cod (*Atlantic cod : the bio-ecology of the fish*, 2019; Rose, 2019) . There was a huge lack of knowledge regarding how environmental conditions affected the population dynamics, as none of the previous population statistics and models anticipated the collapse of the western Atlantic stocks (Chouinard & Fréchet, 1994).

Some Atlantic cod populations, such as the southern Norwegian and Swedish cod, suffer from overexploitation, which results in their population decline as well as a significant shift and imbalance of the ecosystem (Jonsson et al., 2016). Therefore, it is crucial to identify and clarify the potential and occurrence of local adaptation in such high gene flow species. Furthermore, it is important to improve our understanding of the genetic mechanisms for local adaptation (and speciation processes) to conserve genetic resources in a globally changing world. Chromosomal inversions can be an excellent tool for future cod-management, as they will give rise to genetic information about the difference in behaviour ecology in different cod populations.

1.3 Aim of this study

The aim of this study is to identify the breakpoints flanking the inversion on LG 1, 2 and 7 in individuals of cod from different populations to estimate the potential variation in the breakpoint region. This identification will also provide insight into what extent the breakpoints disrupt any gene or regulatory region hitherto undetected from the short read population genomes. To achieve this, we will develop a PCR protocol and design primers flanking the breakpoints and use PacBio HiFi sequencing technology which gives long, highly accurate reads (99,98%) which is ideal to span the repetitive regions in the inversion breakpoints.

2. Materials and method

2.1 Sample collection and DNA extraction

73 tissue samples were obtained from three wild populations of Atlantic cod: Lofoten (30), Averøya (28) and the celtic sea (15) (Figure 3) (Appendix Table 3A). These samples were originally sampled for the Aqua Genome project (Barth et al., 2017; Barth et al., 2019; Matschiner et al., 2022; Pinsky et al., 2021). Additionally, three individuals were obtained from the Atlantic cod breeding program, for aquaculture at Nofima located in Tromsø. The PCR protocol was established and optimized using the three individuals from Nofima and all the genomic DNA was stored at -80 °C.

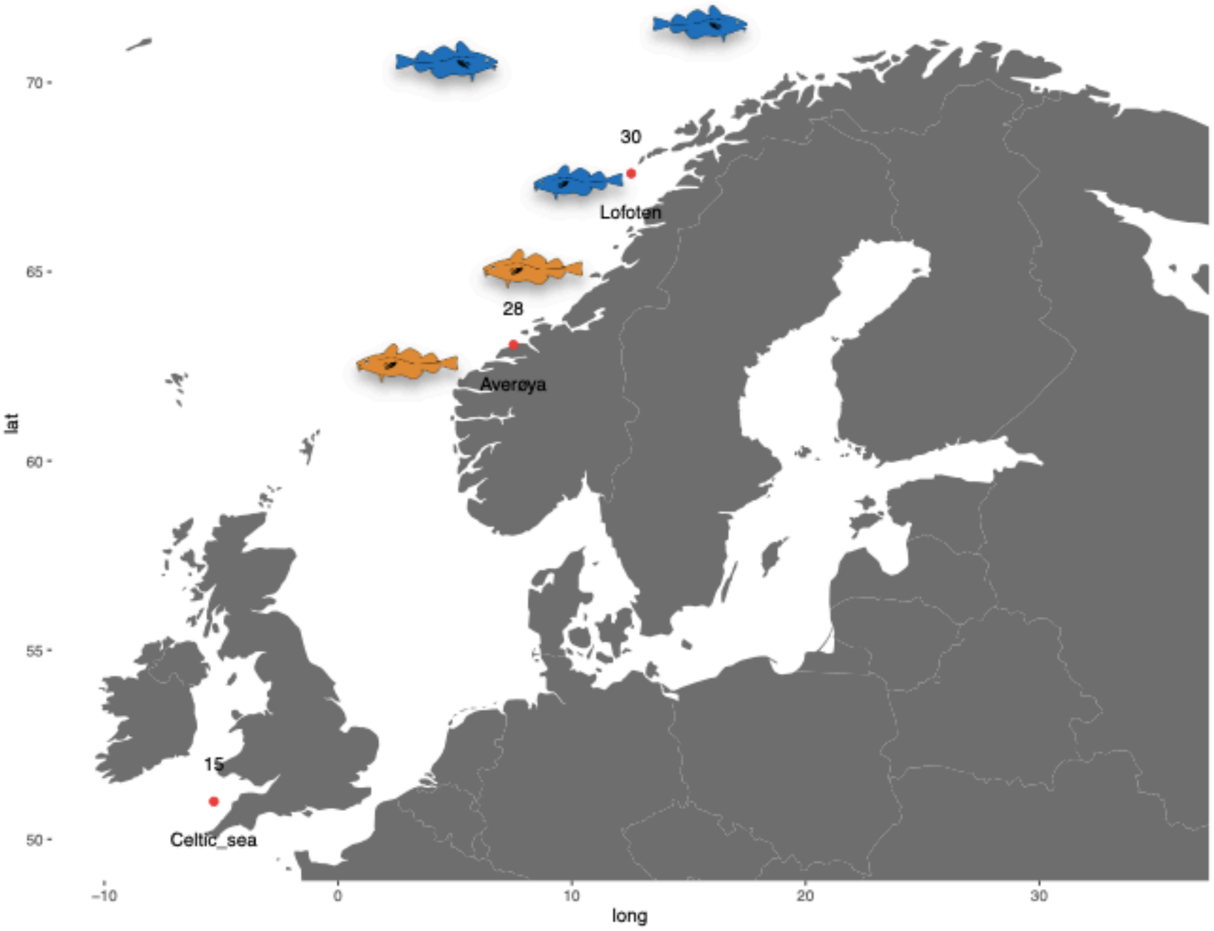


Figure 3*. A map of northern Europe showing in red dots where the three wild populations are located, and how many individuals from each of these populations were used for gDNA extraction. We received 15 individuals from the Celtic Sea, 28 from Averøya, 30 from Lofoten and 3 from Nofima. The NEAC individuals are represented as a blue cod figure, while NCC individuals are represented as an orange cod figure.*

DNA was extracted from 0.025g tissue samples using the DNeasy Blood and tissue mini kit by QIAGEN. All samples were extracted according to the manufacturer´s instructions in "DNeasy® Blood & Tissue Handbook, July 2020" .The concentration and purity of the DNA was estimated using both NanoDrop (Thermo Fisher Scientific, Waltham, MA, USA) and a Qubit fluorometer (Invitrogen, Thermo Fisher Scientific, Waltham, MA, USA). In total, DNA from 73 individuals was extracted and used for sequencing.

2.2 Characterization of inversion breakpoints

Inversion breakpoints were identified by comparing two high-quality genome assemblies of Atlantic cod which have different inversion status on LG 1, LG 2, LG 7 and LG 12. One of the genome assemblies was the chromosome level reference genome of Atlantic cod known as gadMor3 (NCBI accession ID: GCF_ 902167405.1) which was developed using long-read sequencing data produced from a NEAC individual, and the second one is a genome assembly for an individual from the NCC ecotype (Hoff et.al. in prep), that from here on will be referred to as the coastal genome.

The approximate location of the breakpoints of LG 1, LG 2, LG 7 and LG12 inversions were determined by aligning contigs from coastal genome to the gadMor3 genome and investigating where contigs are split in two and map at different locations in the NEAC genome (Brieuc et al in prep.). Approximate positions for the LG 2 inversion breakpoints are shown in figure 4.

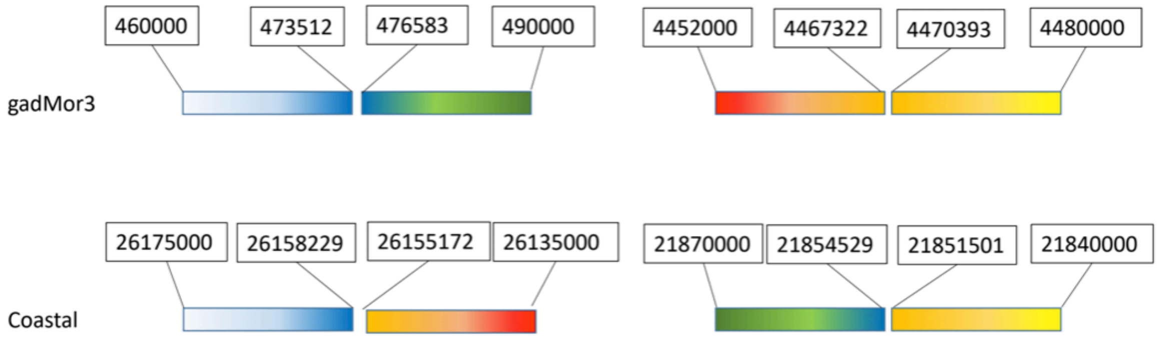


***Figure 4**. The approximate locations given, in basepairs (bp) of the breakpoint regions of LG 2 inversion on gadMor3 and coastal. In the gadMor3 reference genome, the first breakpoint is estimated to be on allele A between 473513-476583 and the second breakpoint to be between 4467322-4470393. On the coastal reference genome, the first breakpoint is estimated to on allele B between 26158229-26155172 and the second breakpoint is to be between 21854529-21851501.*

2.3 Polymerase chain reaction

Polymerase chain reaction (PCR) is a well-used technology that helps to amplify DNA sequences and it revolutionized molecular biology (Saiki et al., 1985). With only a small amount of DNA, the specific and unique sequence can be amplified and used in various downstream applications: cloning, sequencing or other (Clark & Pazdernik, 2013). For establishing the PCR protocol and primer assays, extracted gDNA from three Nofima fish were used as DNA templates. These samples were used for optimization because they were fresh samples that yielded high-quality DNA and were easily accessible. The segment that will be amplified, the targeted sequence, is a region that flanks the breakpoints. To initiate the synthesize, we need enzyme DNA polymerase, and the procedure involves several high-temperature steps, which means our polymerase must be able to endure high temperatures. Throughout the procedure of optimizing the PCR protocol, three different polymerases were tested. The Q5® High-Fidelity DNA polymerase, the Advantage 2 polymerase (A2P) and the KOD Hot start DNA polymerase. We concluded that a two-step PCR reaction was necessary using the Q5 polymerase in the primary PCR run, followed by the A2P when doing nested PCR.

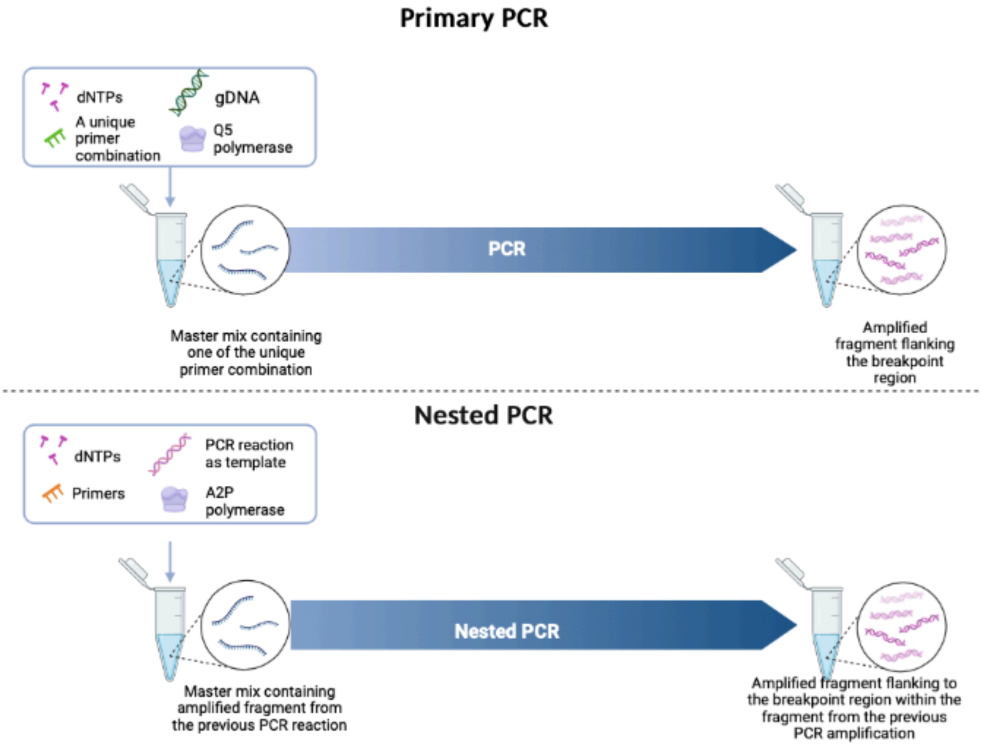

Primary PCR
dNTPs
gDNA
A unique primer combination
Q5 polymerase
PCR
Master mix containing one of the unique primer combination
Amplified fragment flanking the breakpoint region
Nested PCR
dNTPs
PCR reaction as template
Primers
A2P polymerase
Nested PCR
Master mix containing amplified fragment from the previous PCR reaction
Amplified fragment flanking to the breakpoint region within the fragment from the previous PCR amplification

Figure 5. *Q5 polymerase would produce specific, but weak fragments. However, we only need a few specific fragments for nested PCR where we would be able to amplify many fragments flanking the breakpoint region using A2P polymerase. This image was created with BioRender.com*

2.3.1 Primer design

The first objective of this study was to design primers to amplify the breakpoint regions on LG 1, LG 2 and LG 7. The primers were designed based on the position of the inversion breakpoints identified on gadmor3 (NCBI accession ID: GCF_ 902167405.1). The designed primers were in the sequence flanking the largest estimated range of the breakpoint (see Figure 5) so that the relevant sequence would be captured for library preparation and PacBio HiFi sequencing. A set of potential primers were designed by BLAST primer design, a local alignement tool, available on https://www.ncbi.nlm.nih.gov/tools/primer-blast/, and evaluated by NetPrimer (https://www.premierbiosoft.com/netprimer/). Only the three best primers of each flank were chosen from the pool (Table 1). BLAST compares nucleotide or protein sequences to datasets obtained from research and calculates the statistical significance. Conducting a BLAST search of the designed primers against gadmor3 and NCC assembly, revealed that some of them might bind to different locations in each genome, with lower specificity. However, in silico PCR revealed that each pair of primers was unique as there was no other location on the genome where both primers anneal at the same time, in the same efficiency, and produce similar in size fragment. Therefore, all designed primers were used in assay establishment as can be seen in table 1.

To analyze the inversion and their breakpoint region, the PCR protocol was optimized, and the most unique and efficient primers were selected for each breakpoint. We have tested all relevant PCR combinations and three various polymerase kits that were predicted to yield long fragments: Q5® High-Fidelity DNA Polymerase (New England BioLabs), Advantage 2 Polymerase (TaKaRa Bio) and KOD Hot Start. Various PCR parameters were tested: annealing temperature, elongation times, primer, and template concentrations, as well as all possible combinations of primary and nested PCR.

2.3.2 Developing PCR protocol for Linkage group 2

The Nofima individuals were used as a DNA template for developing the PCR protocol since these individuals were easily accessible. Furthermore, after trial and error in testing different primers and PCR parameters on LG 1, LG 2 and LG 7, results were produced for LG 2.

There are two breakpoint regions on LG 2 on the gadmor3 reference genome. Each breakpoint region has three potential primers on each side that could make a unique primer combination for the future breakpoint assay. As of now, these breakpoints will be referred to as the A/B and the C/D breakpoint, as seen in figure 6.

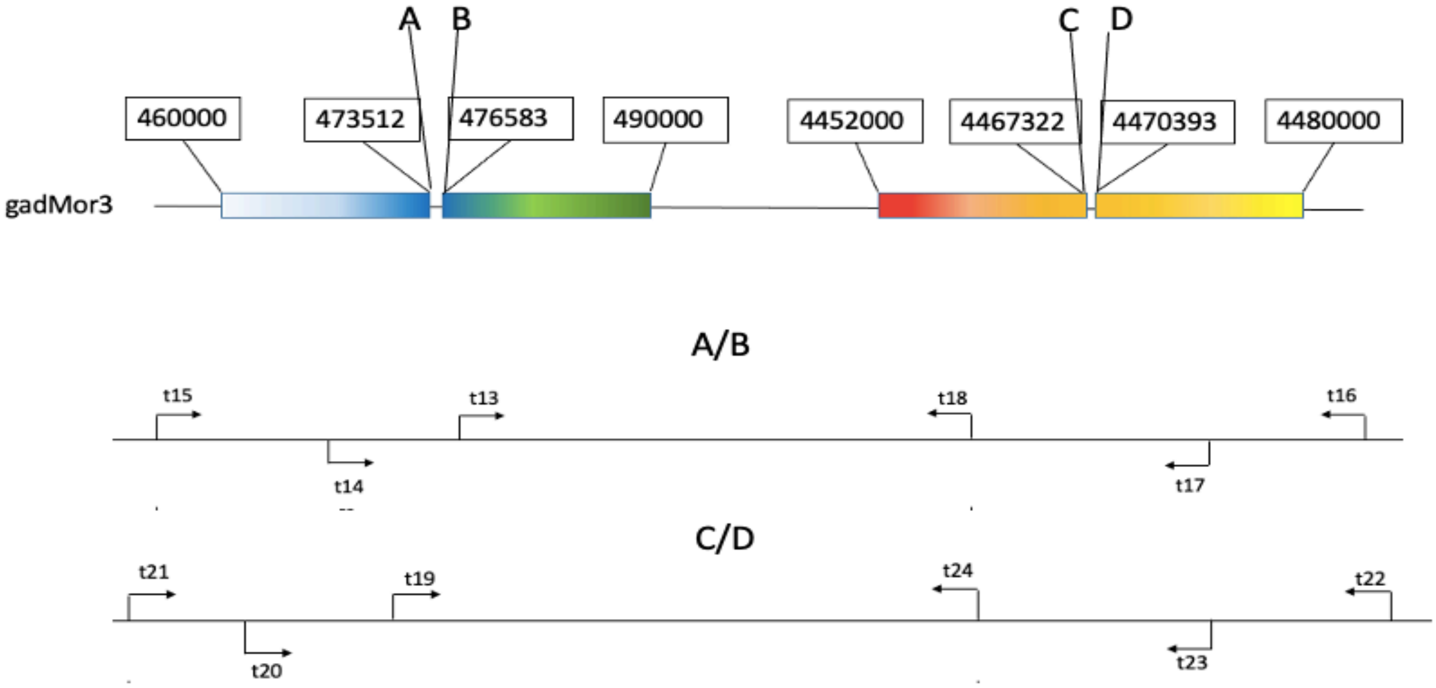


***Figure 6**. Linkage group 2 on gadMor3. A/B and C/D region is where we suspect the breakpoint regions are for the inverted segment on coastal-like individuals. Three potential primers were designed to sequence from each side on the breakpoint region.*

Table 1. An overview of all the primers used when making the AA- and BB-genotype protocols and their position in base pairs on their respective reference genome.

Name of Primer	Sequence	***Position on gadMor3***	***Position on coastal***
t13	CAGCTCAACTGCTCATGGGA	473439 - 473458	26158301 - 26158281
t14	CCAGAGCACAGCTATCGGAG	472583 - 472602	26159166 - 26159146
t15	GAGAACATCACCCCATGCGA	471551 - 471570	26159523 - 26159543
t16	CTCATTGGCTAGCCCACACA	479132 - 479151	21857002 - 21857022
t17	GTGCGCGACACGTTCTTATC	478981 – 479000	21856851 - 21856871
t18	GGGTCGCTTTCTTAGCGGAT	476917 - 476936	21854864 - 21854884
t19	AGCCAGAGTCTTCTTGAGCG	4467247 – 4467266	26155098 - 26155118
t20	CCCGGATAACCCAAACCCTC	4465939 – 4465958	26153802 - 26153822
t21	TCCGACTTTCACCCAACCAG	4463571 – 4463590	26151490 - 21151510
t22	GGTGCTGATTGTGCACCTTG	4473927 – 4473946	21847947 - 21847947
t23	TCCTGCAAAGCCTGTTGTGA	4471670 - 4471689	21850223 - 21850243
t24	GAGTCGGTAGGCCTTTCACC	4470464 - 4470483	21851429 - 21851449

2.3.3 Primary PCR

Q5® High-Fidelity DNA polymerase (Q5) yielded best result in the primary PCR, and the final protocol can be seen in table 2 and 3. The Q5 polymerase, now when this thesis is written, has the highest fidelity amplification available and yields ultra-low error rates, and a 3′ → 5′ exonuclease activity (BioLabsinc). Q5 is a genetically engineered enzyme that consists of two elements: a novel DNA polymerase and a double-stranded DNA (dsDNA) binding protein, Sso7. Sso7 is a kDa nonspecific DNA-binding protein that naturally functions in chromatin remodeling, and it improves speed, fidelity, and reliability of performance. The Sso7d domain supports a robust DNA amplification by preventing the polymerase from dissociating from their DNA templates (Scientific). The Q5 polymerase has proven to be an ideal polymerase to use for long and difficult amplicons, regardless of GC content. Because the polymerase is very specific, it produces fewer bands than A2P and KOD hot start, and even sometimes empty wells. Q5 was therefore ideal to use in a two stage PCR protocol here, so we could amplify the already targeted amplicons and have therefore fewer unspecific fragments.

The annealing temperature was determined with the NEB Tm Calculator version 1.15.0 (https://tmcalculator.neb.com/#!/main)

Table 2. *Master mix for primary PCR using Q5 polymerase.*

Reagent	1 reaction (µl)
Q5 Reaction Buffer	3
Q5 GC Enhancer	3
10 µM Forward Primer	0,75
10 µM Reverse Primer	0,75
genomic DNA (500ng/ul)	0,6
2 mM dNTPs	1,5
Q5 Polymerase	0,15
Nuclease-Free Water	5,25
Total Volume	**15**

Table 3. *Primary PCR protocol when using Q5 polymerase.*

Step	Temperature	Time
Initial Denaturation	98°C	3 minutes
	98°C	10 Seconds
35 Cycles - 50 sec/1kb	67°C	10 Seconds
for long products	72°C	10 Minutes
Final Extension	72°C	2 Minutes
Hold	4°C	∞

2.3.4 Nested PCR

The nested PCR was performed using A2P based its higher yield of the specific fragments from the primary PCR. A2P has a higher yield of PCR product and was therefore ideal as a polymerase used in nested PCR. The final protocol can be seen in table 4 and 5.

A2P is an optimized blend of PCR enzymes that is less specific than Q5. However, it has a high yield and high fidelity that produces more PCR product and is therefore an effective polymerase for nested PCR. A2P is more sensitive and is effective for amplification of long templates up to 18 kB and complex genomic DNA up to 6kb (Takara). It has proven to produce efficient, accurate, and convenient amplification of DNA from any template. The Advantage 2 Polymerase mix contains TITANIUM^{TM} Taq DNA Polymerase and a minor amount of a proofreading polymerase.

Table 4. *Master mix for nested PCR using A2P as polymerase*

Reagent	1 reaction (15 µl)
10X Advantage 2 PCR Buffer	1,5
Primary PCR reaction as template	0,5
10 µM Forward Primer	0,5
10 µM Reverse Primer	0,5
2 mM dNTPs	1,5
50X Advantage 2 Polymerase Mix	0,2
Nuclease-Free water	10,3
Total Volume	**15**

Table 5. *Nested PCR protocol when using A2P as polymerase*

Step	Temperature	Time
Initial Denaturation	95°C	1 Minute
35 Cycles - 1 min/1kb for long products	95°C	20 Seconds
	60°C	20 Seconds
	68°C	10 Minutes
Final Extension	68°C	6 Minutes
Hold	4°C	∞

While Q5 polymerase in primary PCR produced lower concentrations of fragments than A2P and showed weak bands (often no bands) when using gel-electrophoresis, it also produced fewer unspecific fragments based on the bands size than A2P. Furtermore, in nested PCR reactions, A2P proved to be more efficient to produce high yield of only specific fragments. In conclusion, using Q5 polymerase in primary PCR and A2P in nested PCR one would get high yield of the targeted fragments. The DNA fragments in the nested PCR reactions were purified as described in the Wizard®SV Gel and PCR Clean-Up system protocol from Promega.

2.4 Sanger sequencing and selection of primer pairs

After optimizing the PCR protocols and purifying the PCR reactions that contained the targeted fragments, the next step was to verify using sanger sequencing that the fragments amplified were the targeted sequence. For the nested PCR reaction that had multiple bands, gel extraction of fragments was performed. This was done by pipetting the total volume of PCR reaction onto a 1% agarose gel. The bands extracted were those between the size of 3,5 kb and 9 kb. When the DNA had moved through the gel and all the DNA fragments had separated by size, the targeted fragments would be removed from the gel using a scalpel and placed in a 1,5 mL tube for further processing.

The nested PCR reactions and the gel extraction were purified using the Wizard® SV Gel and PCR clean-up system kit according to the manufacturer's instructions. From the now purified nested PCR reaction and gel extraction, 5 µl of the sample was extracted and pipetted to a clean 1,5 mL tube. Furthermore, 2,5 µl of one of the primer combinations was pipetted, as well as 2,5 µl of Milli-Q water, as described in figure 7. Each tube with fragments obtained in nested PCR was then labelled with a barcode and ID info and then sent to be analyzed using Sanger sequencing (EurofinsGenomics).

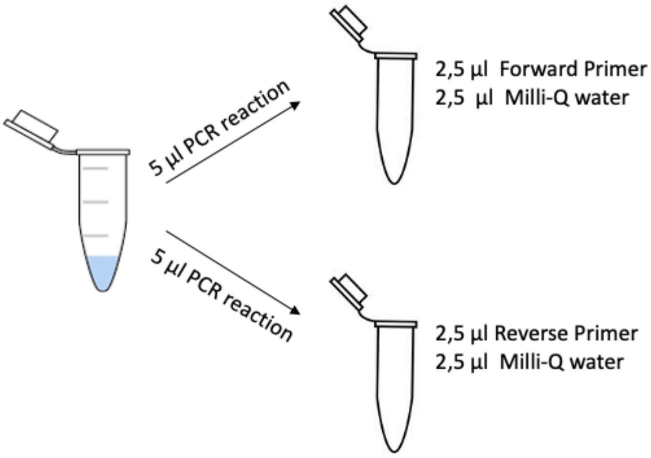
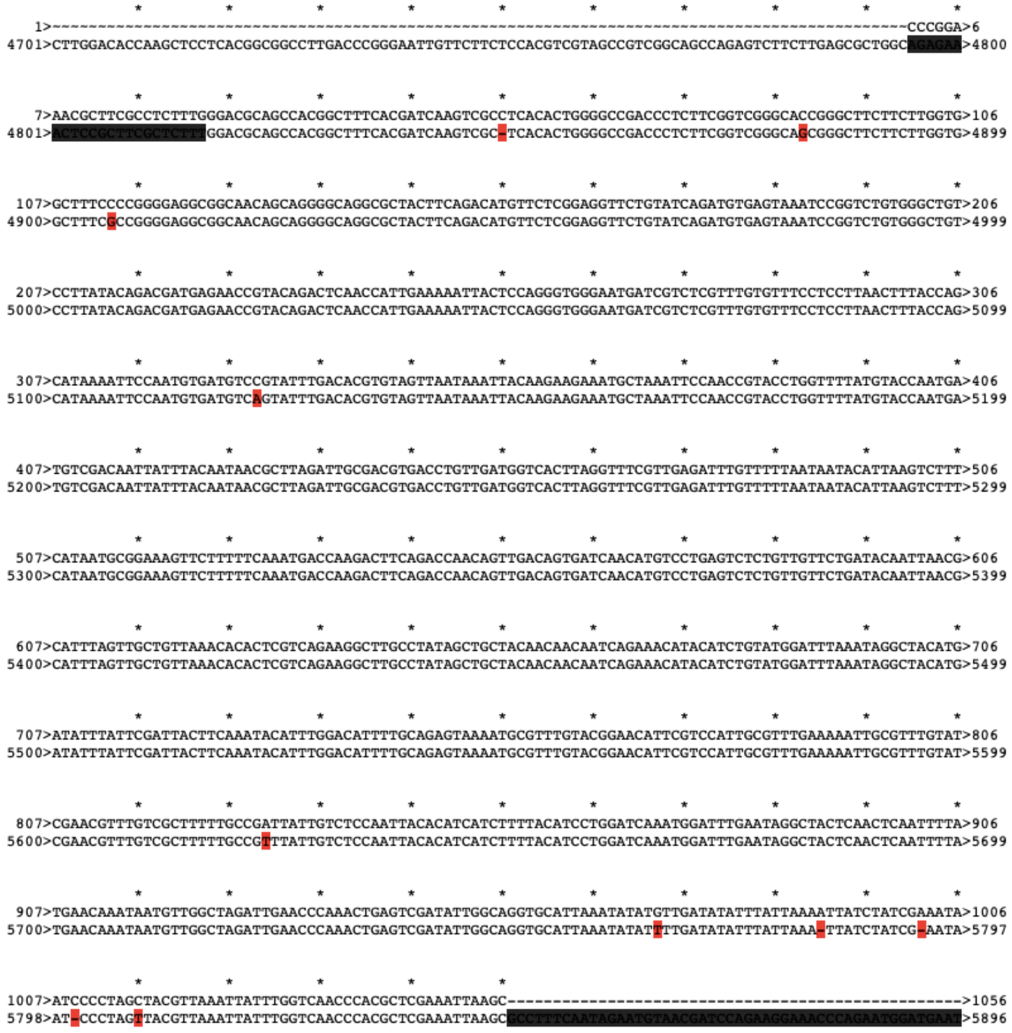


***Figure** 7. From the purified nested PCR reaction tube, 5 µl reaction was pipetted into a new 1,5 mL tube containing 2,5 µl of one of the two primers and 2,5 µl Milliq-Q water.*

After sending the targeted fragments for Sanger sequencing, datasets were received as raw reads in a FASTQ file format which could be aligned to gadMor3 reference-genome using BLAST. In doing so, we could confirm that the fragments we have amplified using PCR are the fragments flanking the breakpoint region. An example of an aligned fragment between the Sanger sequence and gadMor3 can be seen in figure 8.



```
              *         *         *         *         *         *         *         *         *         *
   1>~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~CCCGGA>6
4701>CTTGGACACCAAGCTCCTCACGGCGGCCTTGACCCGGGAATTGTTCTTCTCCACGTCGTAGCCGTCGGCAGCCAGAGTCTTCTTGAGCGCTGGCAGAGAA>4800

              *         *         *         *         *         *         *         *         *         *
   7>AACGCTTCGCCTCTTTGGGACGCAGCCACGGCTTTCACGATCAAGTCGCCTCACACTGGGGCCGACCCTCTTCGGTCGGGCACCGGGCTTCTTCTTGGTG>106
4801>ACTCCGCTTCGCTCTTTGGACGCAGCCACGGCTTTCACGATCAAGTCGC-TCACACTGGGGCCGACCCTCTTCGGTCGGGCAGCGGGCTTCTTCTTGGTG>4899

              *         *         *         *         *         *         *         *         *         *
 107>GCTTTCCCCGGGGAGGCGGCAACAGCAGGGGCAGGCGCTACTTCAGACATGTTCTCGGAGGTTCTGTATCAGATGTGAGTAAATCCGGTCTGTGGGCTGT>206
4900>GCTTTCGCCGGGGAGGCGGCAACAGCAGGGGCAGGCGCTACTTCAGACATGTTCTCGGAGGTTCTGTATCAGATGTGAGTAAATCCGGTCTGTGGGCTGT>4999

              *         *         *         *         *         *         *         *         *         *
 207>CCTTATACAGACGATGAGAACCGTACAGACTCAACCATTGAAAAATTACTCCAGGGTGGGAATGATCGTCTCGTTTGTGTTTCCTCCTTAACTTTACCAG>306
5000>CCTTATACAGACGATGAGAACCGTACAGACTCAACCATTGAAAAATTACTCCAGGGTGGGAATGATCGTCTCGTTTGTGTTTCCTCCTTAACTTTACCAG>5099

              *         *         *         *         *         *         *         *         *         *
 307>CATAAAATTCCAATGTGATGTCCGTATTTGACACGTGTAGTTAATAAATTACAAGAAGAAATGCTAAATTCCAACCGTACCTGGTTTTATGTACCAATGA>406
5100>CATAAAATTCCAATGTGATGTCAGTATTTGACACGTGTAGTTAATAAATTACAAGAAGAAATGCTAAATTCCAACCGTACCTGGTTTTATGTACCAATGA>5199

              *         *         *         *         *         *         *         *         *         *
 407>TGTCGACAATTATTTACAATAACGCTTAGATTGCGACGTGACCTGTTGATGGTCACTTAGGTTTCGTTGAGATTTGTTTTTAATAATACATTAAGTCTTT>506
5200>TGTCGACAATTATTTACAATAACGCTTAGATTGCGACGTGACCTGTTGATGGTCACTTAGGTTTCGTTGAGATTTGTTTTTAATAATACATTAAGTCTTT>5299

              *         *         *         *         *         *         *         *         *         *
 507>CATAATGCGGAAAGTTCTTTTTCAAATGACCAAGACTTCAGACCAACAGTTGACAGTGATCAACATGTCCTGAGTCTCTGTTGTTCTGATACAATTAACG>606
5300>CATAATGCGGAAAGTTCTTTTTCAAATGACCAAGACTTCAGACCAACAGTTGACAGTGATCAACATGTCCTGAGTCTCTGTTGTTCTGATACAATTAACG>5399

              *         *         *         *         *         *         *         *         *         *
 607>CATTTAGTTGCTGTTAAACACACTCGTCAGAAGGCTTGCCTATAGCTGCTACAACAACAATCAGAAACATACATCTGTATGGATTTAAATAGGCTACATG>706
5400>CATTTAGTTGCTGTTAAACACACTCGTCAGAAGGCTTGCCTATAGCTGCTACAACAACAATCAGAAACATACATCTGTATGGATTTAAATAGGCTACATG>5499

              *         *         *         *         *         *         *         *         *         *
 707>ATATTTATTCGATTACTTCAAATACATTTGGACATTTTGCAGAGTAAAATGCGTTTGTACGGAACATTCGTCCATTGCGTTTGAAAAATTGCGTTTGTAT>806
5500>ATATTTATTCGATTACTTCAAATACATTTGGACATTTTGCAGAGTAAAATGCGTTTGTACGGAACATTCGTCCATTGCGTTTGAAAAATTGCGTTTGTAT>5599

              *         *         *         *         *         *         *         *         *         *
 807>CGAACGTTTGTCGCTTTTTGCCGATTATTGTCTCCAATTACACATCATCTTTTACATCCTGGATCAAATGGATTTGAATAGGCTACTCAACTCAATTTTA>906
5600>CGAACGTTTGTCGCTTTTTGCCGTTTATTGTCTCCAATTACACATCATCTTTTACATCCTGGATCAAATGGATTTGAATAGGCTACTCAACTCAATTTTA>5699

              *         *         *         *         *         *         *         *         *         *
 907>TGAACAAATAATGTTGGCTAGATTGAACCCAAACTGAGTCGATATTGGCAGGTGCATTAAATATATGTTGATATATTTATTAAAATTATCTATCGAAATA>1006
5700>TGAACAAATAATGTTGGCTAGATTGAACCCAAACTGAGTCGATATTGGCAGGTGCATTAAATATATTTTGATATATTTATTAAA-TTATCTATCG-AATA>5797

              *         *         *         *         *
1007>ATCCCCTAGCTACGTTAAATTATTTGGTCAACCCACGCTCGAAATTAAGC-------------------------------------------------->1056
5798>AT-CCCTAGTTACGTTAAATTATTTGGTCAACCCACGCTCGAAATTAAGCGCCTTTCAATAGAATGTAACGATCCAGAAGGAAACCCAGAATGGATGAAT>5896
```

***Figure 8**. By using the linkage group 2 on gadMor3 as a template, I could align the sequenced fragment and see if there was a match. As seen here in the figure, the sequence fragment matched with the template. The sequenced fragment is the top query while the template is the bottom query. This is for one of the many primer combinations we sent in for sanger sequencing.*

When the sequences obtained from both ends of the PCR fragment matched the corresponding genomic DNA sequence in gadMor3, we assumed that the fragment was specific. We classified the primers as specific and suitable candidates for the final assay.

2.4.1 Primer combination for inversion

The same PCR approach was used to find the primer combinations for the Atlantic cod individuals' homozygote or heterozygote for the inverted allele A. Genomic DNA that was previously confirmed to have the inverted allele A was used as a template. As seen in figure 9, we assumed that the primers aligned and elongated on the Nofima individuals with the non-inverted allele B would also align and elongate on the individuals with the inverted allele A.

Out of the 73 gDNA samples that were extracted, 39 Atlantic cod individuals provided from the AquaGeome project had a successful PCR run and were used to establish breakpoint assay. These individuals can be seen in table 6, 7, and 8. The genotype is based on WGS short-read data from Brieuc et.al in prep.

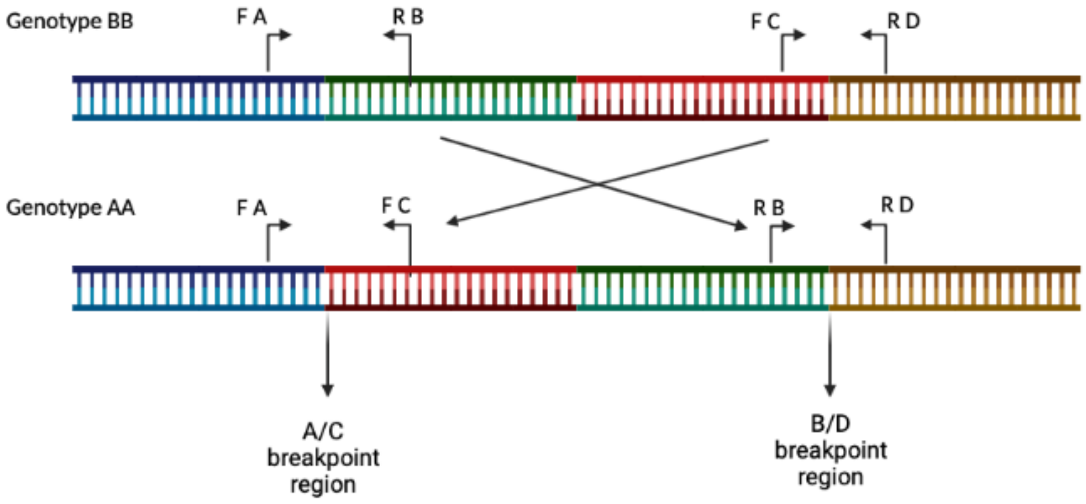


***Figure 9**. Genotype BB represent the Atlantic cod individuals that are homozygote for the non-inverted allele B. While ecotype AA represents Atlantic cod individuals who do are homozygote for inverted allele A on LG 2.*

Table 6*. An overview of the sample collection of the individual's homozygote for the inverted allele A (AA-genotype).*

Location	gDNA ID	Sample name (DNA template)
Lofoten – Coastal	LOF_A_14_03	3.3
Lofoten- Coastal	LOF_A_14_04	4.4
Lofoten – Coastal	LOF_A_14_05	5.5
Lofoten - Coastal	LOF_A_14_22	17
Lofoten – Coastal	LOF_A_14_22	21
Averøya - Coastal	AVE_M_14_02	24
Averøya – Coastal	AVE_M_14_05	25
Averøya – Coastal	AVE_M_14_20	35
Celtic sea	CelticSea_7 IC	36

Table 7*. An overview of the sample collection of the heterozygote individuals (AB genotyped).*

Location	gDNA ID	Sample name (DNA template)
Lofoten – Likely NEAC	LOF_M_14_53	4
Lofoten – NEAC	LOF_M_14_62	8
Lofoten – Likely NEAC	LOF_A_14_06	11
Lofoten – Coastal	LOF_A_14_08	12
Lofoten – Coastal	LOF_A_14_11	13
Lofoten – Coastal	LOF__14_09	14
Lofoten – Coastal	LOF_A_14_16	15
Lofoten – Coastal	LOF_A_14_17	16
Lofoten – Likely Coastal	LOF_A_14_20	19
Lofoten – Coastal	LOF_A_14_21	20
Lofoten – Coastal	LOF_a_14_23	22
Averøya - NEAC	AVE_M_14_09	28

Table 8. An overview over the sample collection of the homozygous BB genotyped individuals.

Location	gDNA ID	Sample name (DNA template)
Lofoten – NEAC	LOF_M_14_50	1
Lofoten – NEAC	LOF_M_14_51	2
Lofoten – NEAC	LOF_M_14_52	3
Lofoten – NEAC	LOF_M_14_54	5
Lofoten – NEAC	LOF_M_14_55	6
Lofoten – NEAC	LOF_M_14_56	7
Lofoten – NEAC	LOF_M_14_68	9
Lofoten – Coastal	LOF_A_14_01	10
Lofoten – NEAC	LOF_A_14_19	18
Aveøya – NEAC	AVE_M_14_01	23
Averøya - NEAC	AVE_M_14_06	26
Averøya – Likely NEAC	AVE_M_14_07	27
Lofoten – NEAC	LOF_M_14_27	27.27
Lofoten – NEAC	LOF_M_14_28	28.28
Lofoten – NEAC	LOF_M_14_29	29.29
Lofoten – NEAC	LOF_M_14_30	30.30
Averøya – Coastal	AVE_M_14_13	30
Averøya – NEAC	AVE_M_14_16	31

3.5 PacBio sequencing

Pacbio single molecule real-time (SMRT) sequencing provides a comprehensive view of genomes, transcriptomes, and epigenomes. Their sequencing technology has developed highly accurate long reads known as HiFi reads (Hon et al., 2020). As a result, this technology provides an accuracy of >99,8% (Wenger et al., 2019). In our case, it will give accurate long reads of all the targeted fragments to estimate where the breakpoint is and analyse if there is variation in the breakpoint region between and within populations. 20 mL of the nested PCR reaction with barcodes was pipetted into a PacBio sequencing plate provided by the Norwegian Sequencing Center (NSC) for PacBio Hifi sequencing. Each fragment sent to the sequencing centre had a barcode attached to the primers for identification. The barcodes were designed so the forward and reverse primer as a pair would be unique and identifiable. The NSC did pooling of the barcoded nested PCR reaction. The nested PCR product contained double-stranded DNA with the targeted fragment, and the sequencing centre ligated hairpin adapters to each end of the fragment, which formed a SMRTbell template. The sequencing polymerase binds to the SMRTbell, and the final sequencing library is prepared and loaded on the SMRT cell for sequencing (Figure 10). The SMRT cell contains millions of tiny wells called zero-mode waveguides (ZMWs), and each fragment is in one well. In these wells, light is emitted throughout the sequencing, and nucleotide incorporation is measured in real-time.

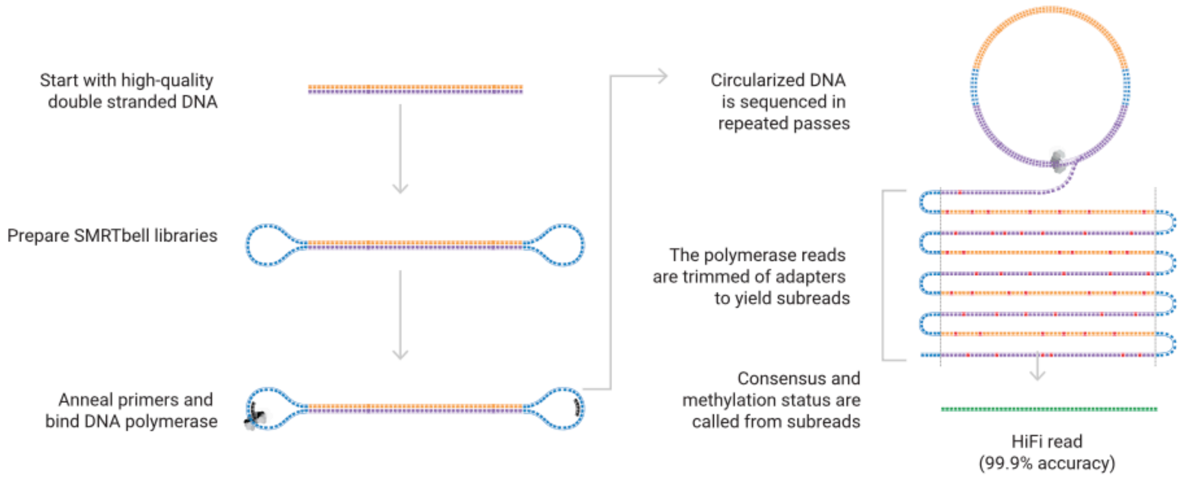

Start with high-quality double stranded DNA
Prepare SMRTbell libraries
Anneal primers and bind DNA polymerase
Circularized DNA is sequenced in repeated passes
The polymerase reads are trimmed of adapters to yield subreads
Consensus and methylation status are called from subreads
HiFi read (99.9% accuracy)

***Figure 10**. Circular consensus sequencing (CCS) improves the accuracy of single-molecule real-time (SMRT) sequencing (PacBio) and generates highly accurate long high-fidelity (HiFi) reads. The process involves ligating hairpin adapters to each end of the DNA fragments where the sequencing primers attach. We then get a circular template for the polymerase to navigate. Our samples were barcoded and multiplexed for increase in throughput. This image is from pacbi.com*

2.5 Sequence analyses

2.5.1 Sequence mapping

The data from PacBio HiFi sequencing was delivered from The NSC (www.sequencing.uio.no) as a FASTQ file format. The NSC demultiplexed the HiFi sequencing reads with Demultiplexing pipeline on SMRT Link v10.2.0.1333434. The circular consensus sequencing (CCS) reads were generated for demultiplexed polymerase reads and again demultiplexed using the barcoded primer sequences. By doing so, the HiFi sequencing reads were separated and indexed with the provide barcode ID. See appendix for table of the demultiplexing results.

FASTQ reads were mapped towards the reference genomes, both coastal and gadmor3 using minimap2 (Li, 2018). Minimap2 is a versatile sequence alignment tool and is available on GitHub at https://github.com/lh3/minimap2. The 79 outputs bam files were then merged into one bam file for each of the reference genomes, i.e. one for gadmor3 and one for coastal.

There are a lot of sequencing similarities between the breakpoint regions, which is why it is vital to confirm that the correct breakpoints were amplified. To do so, the mapped output bam files were examined as each read contains information about where on the genome it was mapped using SAMtools (Danecek et al., 2021). SAMtools would sort the mapping position between primary reads (which would be the correct reads), and alternative positions as secondary reads. Furthermore, the reads were confirmed to have sequenced our targeted region by uploading the mapped bam files and the reference genome on Integrative Genomics Viewer (IGV). Between using SAMtools and IGV, we could confirm that we had sequenced our targeted area.

2.5.2 Variant calling and phasing

We now have two datasets of sequence reads, each correctly aligned to one of the two reference genomes (gadMor 3 and coastal) stored as bam files and separated between these two ecotypes. The next step is then to call variants from these alignments using BCFtools mpileup (Danecek et al., 2021). We called variants - the process of analyzing the aligned bam files and identify positions that differ from the reference genome- for each dataset, gadMor3 and coastal, respectively. The resulting VCF files were used for haplotype phasing. The Atlantic cod is a diploid organism, meaning that it has two complete sets of chromosomes, one from each parent. Hence, after variant calling, the reads were separated into two chromosome sets. This process is called phasing, and was done by using WhatsHap v1.4 (Martin et al., 2016). Here one uses the bam files produced from mapping and run it with the VCF file produced from variant calling and one will have an output file with the phased reads (two haplotypes) for each reference genome. To better visualize each haplotype per read, and the variation within alleles per individual, the haplotypes were tagged, still using WhatsHap, before making an assembly using Flye v2.9 (Kolmogorov et al., 2020). The benefit of making *de novo* assemblies is that we can reconstruct the breakpoint region haplotypes without reference bias.

For visualizing the HiFi-sequencing reads the Integrative Genomics Viewer (IGV) was used. IGV is a high-performance tool for visual exploration which is free and easy to use. IGV enables intuitive real-time exploration of diverse, large-scale genomic data sets on standard desktop computers (Robinson et al., 2011)

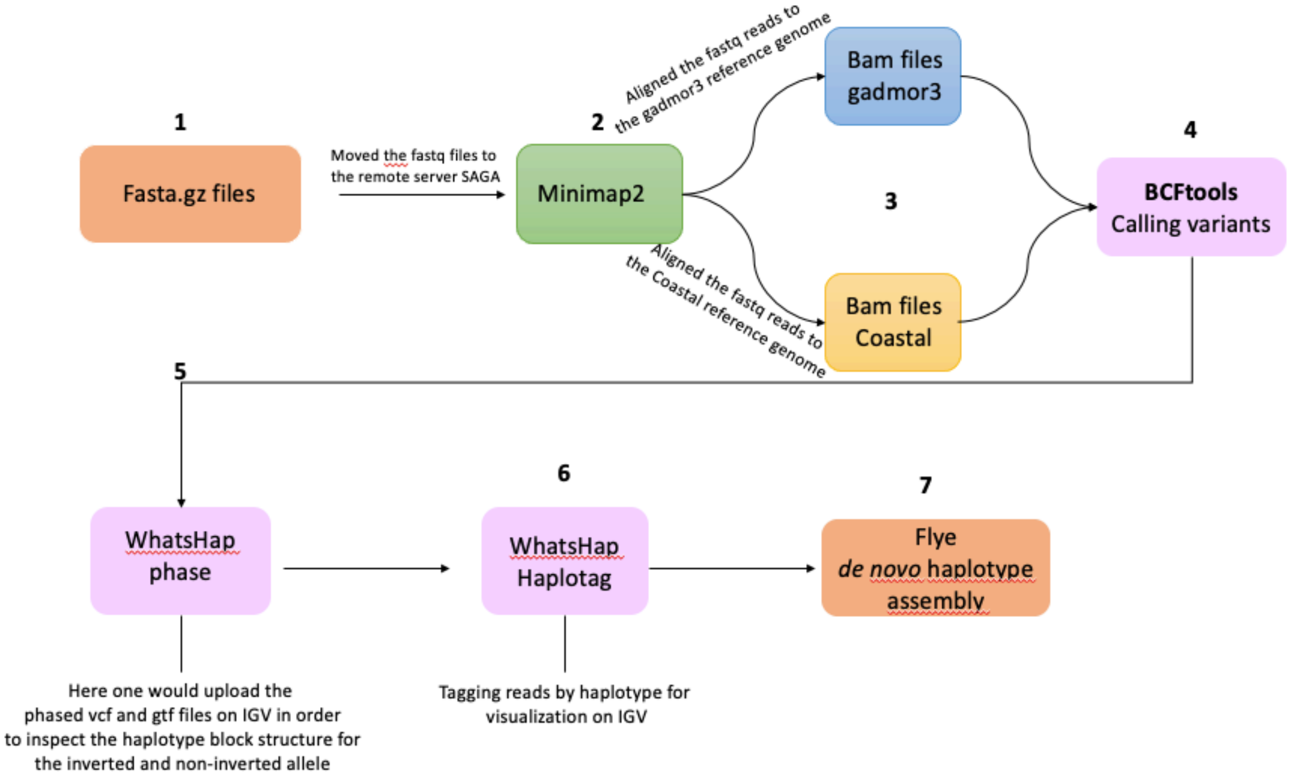

1
Fasta.gz files
Moved the fastq files to the remote server SAGA
2
Minimap2
Aligned the fastq reads to the gadmor3 reference genome
Bam files gadmor3
3
Aligned the fastq reads to the Coastal reference genome
Bam files Coastal
4
BCFtools Calling variants
5
WhatsHap phase
6
WhatsHap Haplotag
7
Flye de novo haplotype assembly
Here one would upload the phased vcf and gtf files on IGV in order to inspect the haplotype block structure for the inverted and non-inverted allele
Tagging reads by haplotype for visualization on IGV

***Figure 11**. A flowchart of the pipeline that was used for analyzing the HiFi sequences and create de novo haplotype assemblies.*

2.5.3 Multiple sequence alignment and masking repeats

The haplotype assemblies of the breakpoint region on LG 2 were aligned in a multiple sequence alignment in both MEGA11 and geneious prime. All four breakpoints were analyzed for variation within and between populations. Furthermore, all four breakpoints were screened for repeats using RepeatMasker 4.0.9, a free application available on http://www.repeatmasker.org. The RepeatMasker program screens the DNA sequences, uploaded in FASTA format, for interspersed repeats and low complexity DNA sequences.

3. Results

3.1 Design and optimization of PCR-protocols

Two PCR-protocols were made for amplifying targeted fragments through a continuing of testing different PCR parameters and primer combinations. One protocol would be used for homozygous individuals with the non-inverted allele B, called the BB-genotyped protocol. While the other was used for the homozygous individuals for the inverted allele A, called the AA-genotyped protocol. Both PCR protocols are very similar to each other; the only difference was the primer combination.

For the heterozygote individuals a PCR protocol which combined both AA- and BB-genotyped primer combination was developed. During the process of performing PCR with barcoded primers on the Atlantic cod samples, a reoccurring problem was encountered with some of the breakpoint regions. The PCR protocol was developed and optimized using Nofima individuals, however the Atlantic cod varies enough that some individuals were not responding well to the primer pair of choice. To find different primer combinations for the individuals that were lacking results the PCR protocol was optimized again by making a new breakpoint assay for these individuals (see chapter 4.1.3).

3.1.1 BB-genotype protocol (Ancestral allele)

All primers were paired up and tested to make a primary assay using genomic DNA from a Nofima fish called 4. One PCR reaction would contain one unique primer combination (table 8), and the expected size would vary depending on which primer pair it was. From primary PCR, the fragment size could vary from 3498 bp to 8119bp. All the primary PCR reactions samples were used as templates for nested PCR reaction. As seen in figure 13, only one band was visible from the primary PCR, but in the nested PCR gel, it is confirmed that almost every primary PCR reaction contained the targeted fragment. All the nested PCR reaction that contained fragments of the expected size was extracted for Sanger sequencing.

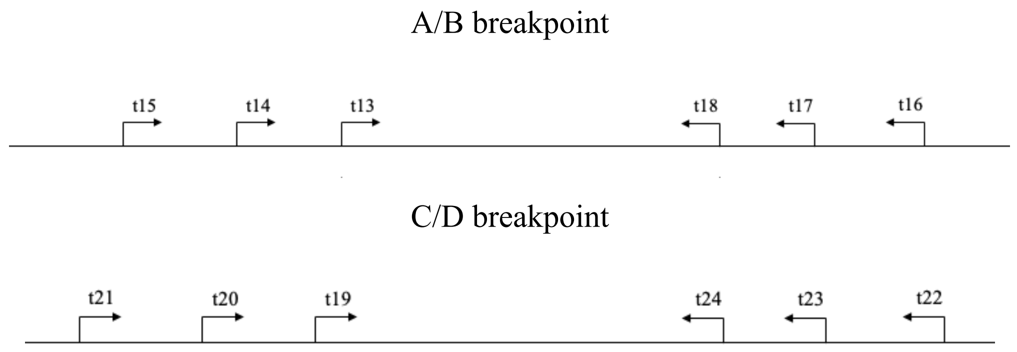


Figure 12. *An overview of the different primers and where they are located according to each other. Each break point region has three primers on each size, and in the process of finding which combination would yield best results, all combinations are tested for.*

Table 9. *An overview of each PCR reaction that was pipetted into the gel found on figure 13, using genomic DNA fish 4 as template. Each well describes one unique primer combination to find out which one can be used in the optimized protocol. These PCR reactions were then again used as template for nested PCR.*

Gene	**Sample Name**	**Primer**	**Exp.Size (bp)**
LG02 A/B	1 AB	t13 & t16	5562
LG02 A/B	2 AB	t13 & t17	5562
LG02 A/B	3 AB	t13 & t18	3498
LG02 A/B	4 AB	t14 & t16	6569
LG02 A/B	5 AB	t14 & t17	6418
LG02 A/B	6 AB	t14 & t18	4354
LG02 A/B	7 AB	t15 & t16	7601
LG02 A/B	8 AB	t15 & t17	7450
LG02 A/B	9 AB	t15 & 18	5386
LG02 C/D	1 CD	t19 & t22	6700
LG02 C/D	2 CD	t19 & t23	4443
LG02 C/D	3 CD	t19 & t24	3237
LG02 C/D	4 CD	t20 & t22	8008
LG02 C/D	5 CD	t20 & t23	5751
LG02 C/D	6 CD	t20 &t24	4545
LG02 C/D	7 CD	t21 & t22	10376
LG02 C/D	8 CD	t21 & t23	8119
LG02 C/D	9 CD	t21 & t24	6913

***Table 10**. An overview of each nested PCR reaction and its sample name.*

Sample Name	Primer	Template	Exp. Size	Sample Name	Primer	Template	Exp. Size
1	t13 & t17	1 AB	5562	28	t19 & t23	1 CD	4443
2	t13 & t18	1 AB	3498	29	t19 & t22	1 CD	6700
3	t13 & t18	2 AB	3498	30	t19 & t24	2 CD	3237
4	t14 & t17	4 AB	6418	31	t20 & t23	4 CD	5751
5	t14 & t16	4 AB	6569	32	t20 & t24	4 CD	4545
6	t13 & t16	4 AB	5713	33	t19 & t22	4 CD	6700
7	t13 & t17	4 AB	5562	34	t19 & t23	4 CD	4443
8	t13 & t18	4 AB	3498	35	t19 & t24	4 CD	3237
9	t14 & t17	5 AB	6418	36	t20 & t24	5 CD	4545
10	t13 & t18	5 AB	3498	37	t19 & t24	5 CD	3237
11	t13 & t17	5 AB	5562	38	t19 & t23	5 CD	4443
12	t13 & t18	6 AB	3498	39	t19 & t24	6 CD	3237
13	t15 & t17	7 AB	7450	40	t21 & t23	7 CD	8119
14	t15 & t18	7 AB	5386	41	t21 & t24	7 CD	6913
15	t14 & t16	7 AB	6569	42	t20 & t22	7 CD	8008
16	t14 & t17	7 AB	6418	43	t20 & t23	7 CD	5751
17	t14 & t18	7 AB	4354	44	t20 & t24	7 CD	4545
18	t13 & t16	7 AB	5713	45	t19 & t22	7 CD	6700
19	t13 & t17	7 AB	5562	46	t19 & t23	7 CD	4443
20	t13 & t18	7 AB	3498	47	t19 & t24	7 CD	3237
21	t15 & t18	8 AB	5386	48	t21 & t24	8 CD	6913
22	t14 & t17	8 AB	6418	49	t20 & t23	8 CD	5751
23	t14 & t18	8 AB	4354	50	t20 & t24	8 CD	4545
24	t13 & t18	8 AB	3498	51	t19 & t23	8 CD	4443
25	t13 & t17	8 AB	5562	52	t19 & t24	8 CD	3237
26	t14 & t18	9 AB	4354	53	t21 & t24	9 CD	6913
27	t13 & t18	9 AB	3498	54	t19 & t24	9 CD	3237

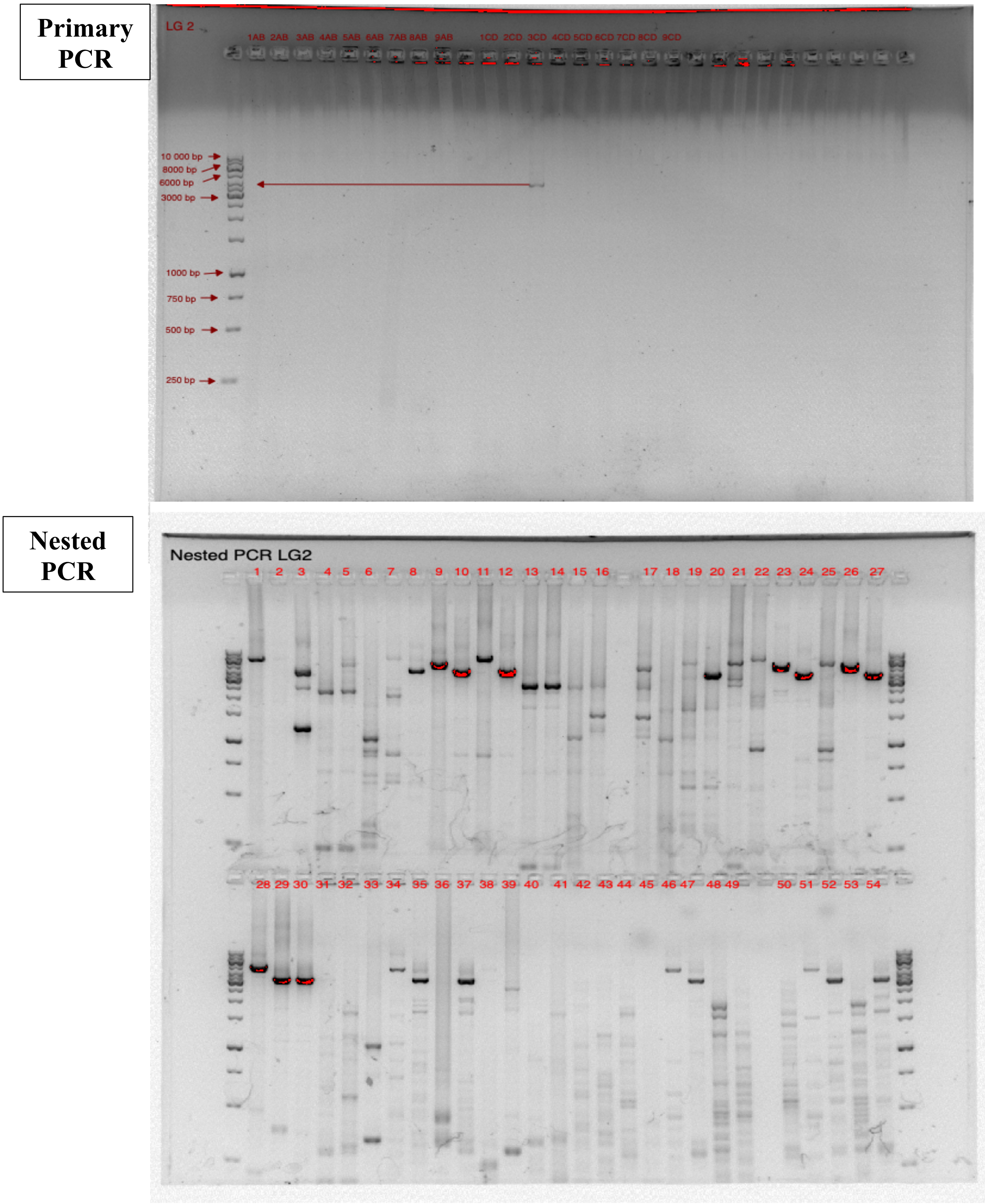


***Figure 13**. The top gel image is of the primary PCR reaction using the Nofima fish 4 as a template, called "fish 4". The bottom gel image is of the nested PCR using the primary PCR reactions as template. As seen in figure, the gel containing the primary PCR reactions is mostly empty, except for one band in well 3CD. Each primary PCR reaction was used as template for nested PCR, and the amplified fragments can be seen in the nested PCR gel. Here we see the targeted fragments that were then extracted and analyzed for sanger sequencing later.*

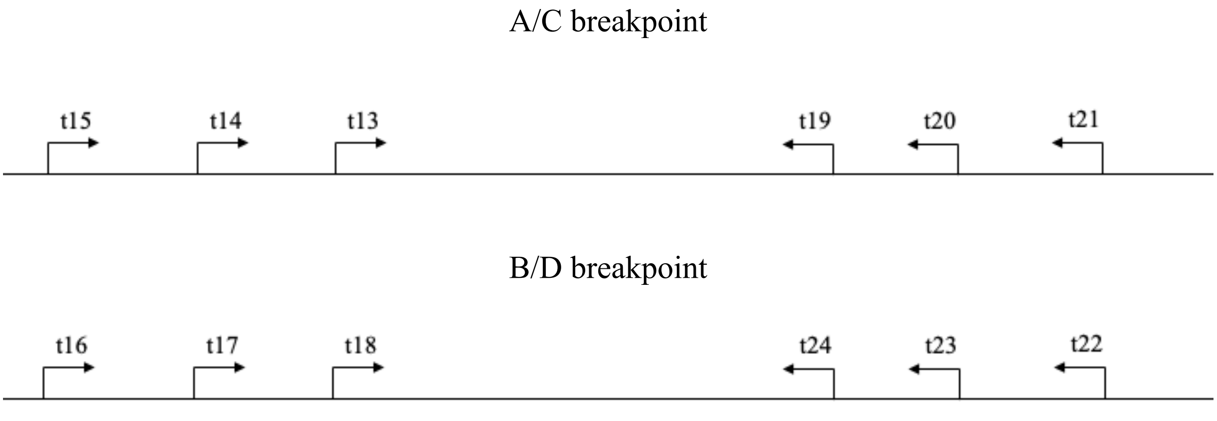
As seen in figure 13, many bands had the potential to be the targeted fragments. All fragments that were in the proximity of the expected size were sent for Sanger sequencing. After receiving the data in the form of a FASTA file, the fragments were mapped toward the reference genome, and we concluded that we would continue using these primer combinations for the BB-protocol:

Table 11. *An overview of the primer combinations for the BB-protocol*

	A/B	*C/D*	*Exp.Size (bp)*	
Primary	t14 & t16	t19 & t22	6569	6700
Nested	t14 & t17	t19 & t23	6418	4443

3.1.2 AA-genotype protocol (Derived allele)

All primers were paired up and tested to make a primary assay using genomic DNA from fish 29HC (LOF_M_14_29). One PCR reaction would contain one unique primer combination (table 12), and the expected size would vary depending on which primer pair it was. All the primary PCR reaction samples were used as templates for nested PCR reaction. As seen in figure 15, three bands were visible from primary PCR, but in the nested PCR gel, almost every primary PCR reaction contained the targeted fragment. All the nested PCR reaction that contained fragments of the expected size was extracted for Sanger sequencing.


A/C breakpoint

B/D breakpoint

Figure 14. *An overview of the different primers and where they are located according to each other. The figure shows that we use the same primers in both protocols, but the combinations differ. Each primer combination is unique, and in the optimization stage, we wanted to find the primer combination that flanked the breakpoint region.*

Table 12. *An overview of each PCR reaction that was pipetted into the gel found on figure 15, using fish 29 HC as template. Each well describes one unique primer combination for inverted segments to find out which one can be used in the optimized AA-protocol. These PCR reactions were then again used as template for nested PCR.*

Gene	Sample name	Primer	Exp.size (bp)
LG2 A/C	24	t14 & t21	7676
LG2 A/C	25	t15 & t19	4425
LG2 A/C	26	t15 & t20	5721
LG2 A/C	27	t15 & t21	8033
LG2 B/D	28	t16 & t22	9055
LG2 B/D	29	t16 & t23	6779
LG2 B/D	30	t16 & t24	5553
LG2 B/D	31	t17 & t22	8904
LG2 B/D	32	t17 & t23	6628
LG2 B/D	33	t17 & t24	5422
LG2 B/D	34	t18 & t22	6917
LG2 B/D	35	t18 & t23	4641
LG2 B/D	36	t18 & t24	3435

Table 13. *An overview of each nested PCR reaction and its sample name. The nested PCR reaction can be seen as bands on the gel in figure 15.*

Sample Name	Primer	Template	Exp.Size	Sample Name	Primer	Template	Exp. Size
4	t13 & t19	22	3203	30	t17 & t22	28	8904
5	t14 & t19	23	4068	31	t17 & t23	28	6628
6	t13 & t20	23	4499	32	t17 & t24	28	5422
7	t13 & t19	23	3203	33	t18 & t22	28	6917
8	t14 & t20	24	5364	34	t18 & t23	28	4641
9	t14 & t19	24	4068	35	t18 & t24	28	3435
10	t14 & t21	24	7676	36	t16 & 24	29	5553
11	t13 & t20	24	4499	37	t17 & t23	29	6628
12	t13 & t19	24	3203	38	t17 & t24	29	5422
13	t14 & t19	25	4068	39	t18 & t23	29	4641
14	t13 & t19	25	3203	40	t18 & t24	29	3435
15	t15 & t19	26	4425	41	t17 & t24	30	5422
16	t14 & t20	26	5364	42	t18 & t24	30	3435
17	t14 & t19	26	4068	43	t17 & t23	31	6628
18	t13 & t20	26	4499	44	t17 & t24	31	5422
19	t13 & t19	26	3203	45	t18 & t22	31	6917
20	t15 & t20	27	5721	46	t18 & t23	31	4641
21	t15 & t19	27	4425	47	t18 & t24	31	3435
22	t14 & t21	27	7676	48	t17 & t24	32	5422
23	t14 & t20	27	5364	49	t18 & t23	32	4641
24	t14 & t19	27	4068	50	t18 & t24	32	3435
25	t13 & t21	27	6811	51	t18 & t24	33	3435
26	t13 & t20	27	4499	52	t18 & t23	34	4641
27	t13 & t19	27	3203	53	t18 & t24	34	3435
28	t16 & t23	28	6779	54	t18 & t24	35	3435
29	t16 & t24	28	5553				

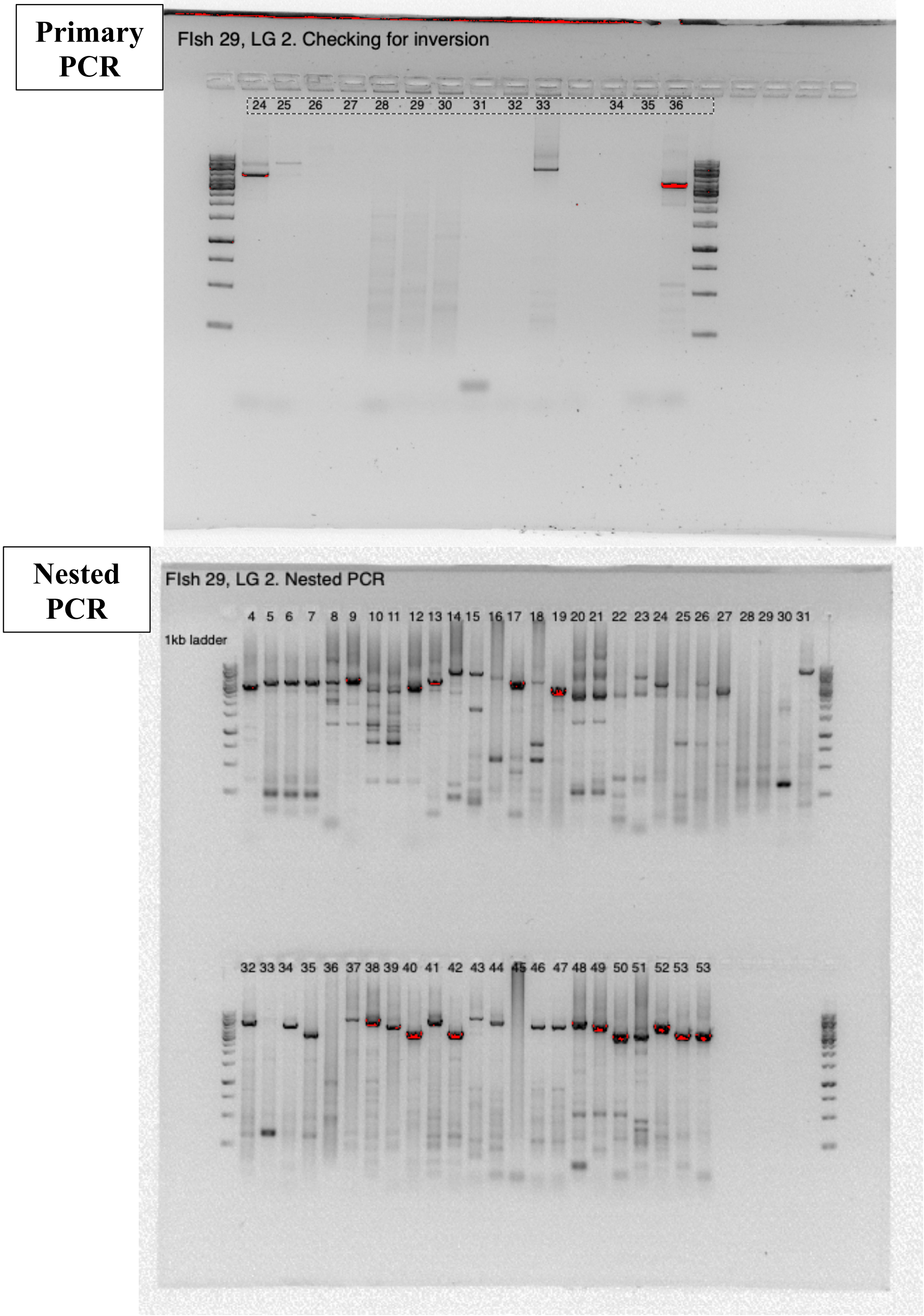


***Figure 15**. Gel image of the primary and nested PCR reactions of gDNA sample 29 HC. The primer reaction from primary PCR was used as template for nested PCR and as see in this figure, more amplified fragment was present. Some of the wells are still empty, which could be because the primer combination did not align properly to the sequence.*

The next step was to extract the fragments from the gel, clean the PCR reaction, and send it for Sanger sequencing to confirm that the fragment is the one flanking the breakpoint region. From figure X we could see many bands that had the potential to be our targeted fragments. After sending the fragments for Sanger sequencing, we concluded that we would continue using these primer combinations for the AA-protocol:

Table 12. An *overview of the primer combination for the AA-protocol*

	A/C	B/D	Exp.Size (bp)	
Primary	*t14 & t20*	*t16 & t22*	5364	9055
Nested	*t14 & t19*	*t17 & t23*	4068	6628

3.1.3 Optimization of the PCR-protocol

A setback occurred when performing PCR on all the Atlantic cod individuals using the primer combinations established after Sanger sequencing. Some PCR reactions were not yielding any results. Previously when optimizing the protocol, Nofima individuals were used as genomic template for the BB-protocol, and a NEAC individual for AA-protocol. However, the genetic variation in the rimer region between Atlantic cod populations may make the primers less suitable for some populations. There were two primer combinations that more often than other primers were lacking results. The A/B breakpoint region and the B/D breakpoint region, which most likely meant that the primers for B sequence were in a region not sufficiently conserved in cod population. This required a step back and to test out some primer combinations using gDNA sample that were proven to be of good quality from previous PCR reaction.

3.1.3.1 Optimizing the BB-protocol

For developing a new primary assay for the BB-protocol, two different Atlantic cod individuals were chosen. Both are NEAC individuals, genotype BB, and have the non-inverted allele, however, they are from two different populations. LOF_M_15_50 is from the Lofoten population, while AVE_M_14_01 is from Averøya. From running the primary and nested PCR, the new primer combination would be the one that gave high concentration of fragments for both individuals.

Table 13. Overview of the gDNA samples that were used for testing out new primer combinations for the optimized BB-protocol.

Species ID	gDNA sample ID	Species type	Genotype	Protocol used	Name of the protocol
LOF_M_14_50	1	NEAC	BB	BB	20220307.PCR.AB-TEST.LG2
AVE_M_14_01	23	NEAC	BB	BB	20220307.PCR.AB-TEST.LG2

Table 14. *Overview over the primary PCR reactions that would make the A/B breakpoint assay. Using all the primer combinations on two individuals we know have good quality DNA, to find a new primer combination for the A/B breakpoint region. The gel can be seen in figure X.*

Gene	Sample name	Primer	Template
LG2 A/B	1	t13 & t18	1
LG2 A/B	2	t13 & t17	1
LG2 A/B	3	t13 & t16	1
LG2 A/B	4	t14 & t18	1
LG2 A/B	5	t14 & t17	1
LG2 A/B	6	t14 & t16	1
LG2 A/B	7	t15 & t18	1
LG2 A/B	8	t15 & t17	1
LG2 A/B	9	t15 & t16	1
LG2 A/B	10	t13 & t18	23
LG2 A/B	11	t13 & t17	23
LG2 A/B	12	t13 & t16	23
LG2 A/B	13	t14 & t18	23
LG2 A/B	14	t14 & t17	23
LG2 A/B	15	t14 & t16	23
LG2 A/B	16	t15 & t18	23
LG2 A/B	17	t15 & t17	23
LG2 A/B	18	t15 & t16	23

Table 15. *An overview of the nested PCR reactions.*

Nested PCR sample name	Primer	Primary PCR template
1	t13 & t18	2
2	t13 & t18	3
3	t13 & t17	3
4	t13 & t18	4
5	t14 & t18	5
6	t13 & t18	5
7	t13 & t17	5
8	t14 & t17	6
9	t14 & t18	6
10	t13 & t16	6
11	t13 & t17	6
12	t13 & t18	6
13	t14 & t18	7
14	t13 & t18	7
15	t15 & t18	8
16	t14 & t17	8
17	t14 &t18	8
18	t15 & t17	9
19	t15 & t18	9
20	t14 & t16	9
21	t13 & t16	9
22	t13 & t18	11
23	t13 & t18	12
24	t13 & t17	12
25	t13 & t18	13
26	t14 & t18	14
27	t13 & t18	14
28	t13 & t17	14
29	t14 & t17	15
30	t14 & t18	15
31	t13 & t16	15
32	t13 & t17	15
33	t13 & t18	15
34	t14 & t18	16
35	t13 & t18	16
36	t15 & t18	17
37	t14 & t17	17
38	t14 & t18	17
39	t15 & t17	18
40	t15 & t18	18
41	t14 & t16	18
42	t13 & t16	18

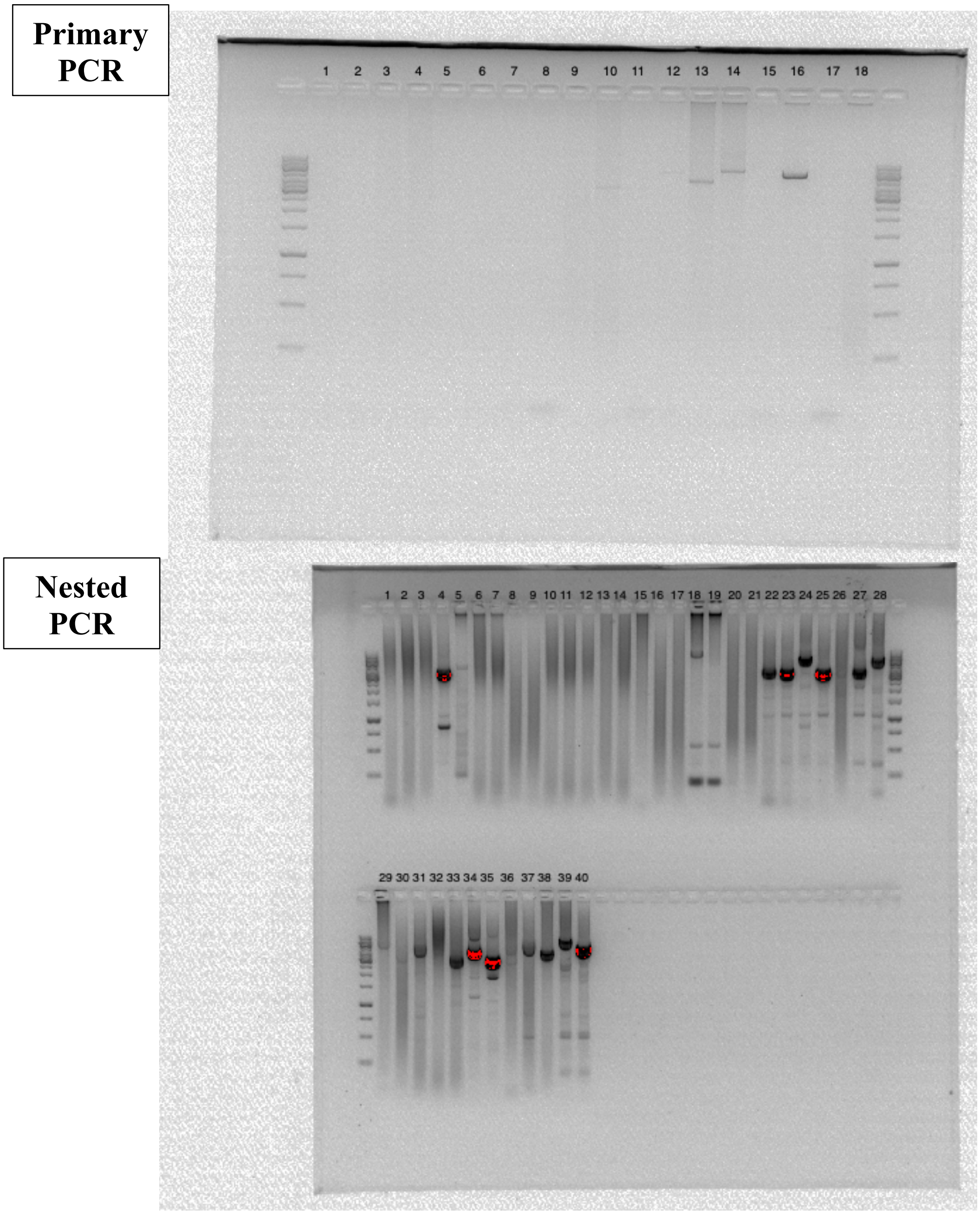


Figure 16. *Top gel is of the primary PCR reactions found in table X. The bottom gel is of the nested PCR reactions found in table X. As seen in the figure, using new primer combinations yielded amplified fragments that previously were now amplified using the primer combination established when using a Nofima individual as template.*

From making the optimized A/B breakpoint assay, new primer combinations would be used for the individuals with the non-inverted allele

Table 16. *An overview over the new primer combination for the optimized BB-protoccol*

	A/B	Exp. Size (bp)
Primary	*t14 & t18*	4354
Nested	*t13 & t18*	3498

3.1.3.2 Optimizing the AA-protocol

A stationary coastal individual from the Lofoten population (LOF_A_14_18) was used to develop a new primary assay for the AA-protocol (Table 17). It had previously yielded results for the A/B breakpoint fragment, which confirmed that the DNA quality was good, but the protocol needed an optimized primer combination for the B/D breakpoint.

Table 17. *Overview of the gDNA sample that was used for testing out new primer combinations for the optimized AA-protocol.*

Species ID	gDNA sample ID	Species type	Genotype	Protocol used	Name of the protocol
LOF_A_14_18	17	Coastal	AA	AA	220228-LG2BD-TEST

Table 18. *Overview over the primary PCR reactions for the B/D assay.*

Gene	Sample name	Primer	Template
LG 2 B/D	10	t16 & t22	17
LG 2 B/D	11	t16 & t23	17
LG 2 B/D	12	t16 & t24	17
LG 2 B/D	13	t17 & t22	17
LG 2 B/D	14	t17 & t23	17
LG 2 B/D	15	t17 & t24	17
LG 2 B/D	16	t18 & t22	17
LG 2 B/D	17	t18 & t23	17
LG 2 B/D	18	t18 & t24	17

Table 19. *Overview of the nested PCR reactions for the B/D assay*

Nested PCR sample name	Primer	Primary PCR template
28	t16 & t23	10
29	t16 & t24	10
30	t17 & t22	10
31	t17 & t23	10
32	t17 & t24	10
33	t18 & t22	10
34	t18 & t23	10
35	t18 & t24	10
36	t16 & t24	11
37	t17 & r23	11
38	t17 & t24	11
39	t18 & t23	11
40	t18 & t24	11
41	t17 & t24	12
42	t18 & t24	12
43	t17 & t23	13
44	t17 & t24	13
45	t18 & t22	13
46	t18 & t23	13
47	t18 & t24	13
48	t17 & t24	14
49	t18 & t23	14
50	t18 & t24	14
51	t18 & t24	15
52	t18 & t23	16
53	t18 & t24	16
54	t18 & t24	17

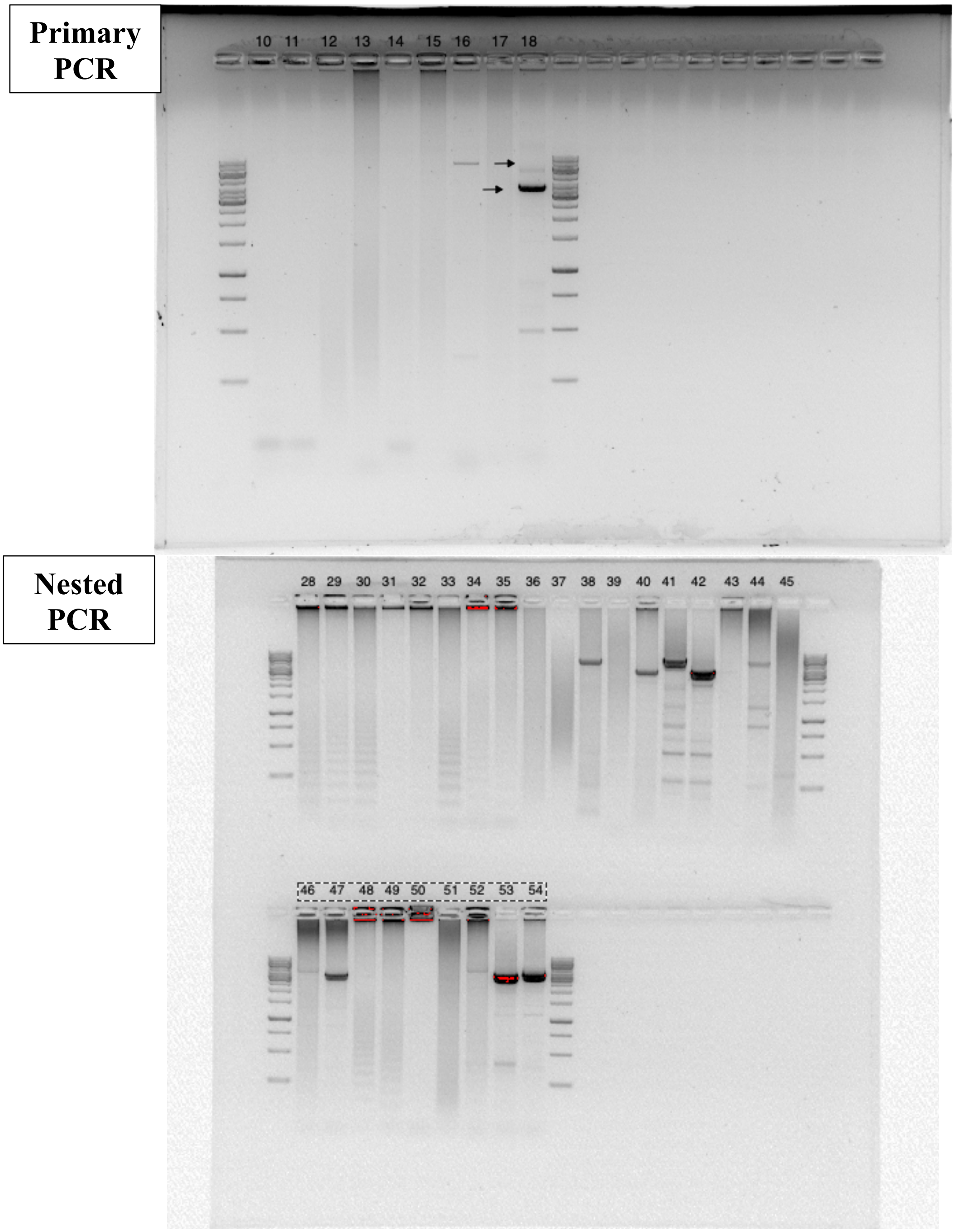


Figure 17*. Top gel contain the primary PCR reactions found in table X. The bottom gel contain the nested PCR reactions found in table X.*

From making the B/D breakpoint assay, we concluded that we would continue with these primer combinations:

Table 20. *An overview over the new primer combination for the optimized AA-protoccol*

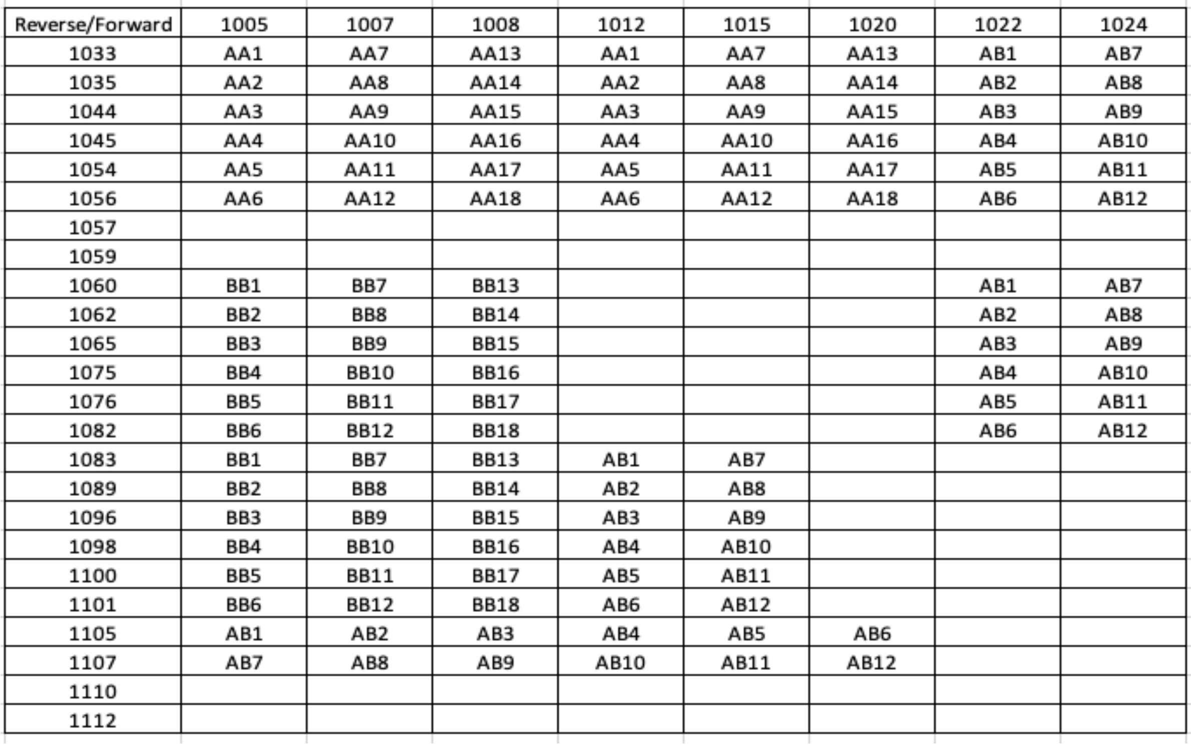
	B/D	*B/D*	*Exp.Size (bp)*	
Primary	t16 & t23	t17 & t22	6779	8904
Nested	t18 & t24	t18 & t24	3435	

3.2 PacBio HiFi-sequencing

3.2.1 Library preparation, CCS and demultiplexing results

For PacBio HiFi-sequencing, 83 breakpoint fragments were successfully barcoded during PCR. The barcoded primers were universal primers for multiplexing amplicons provided by the NSC, and the pair would be unique and identifiable (Table 19) for HiFi-sequencing. The 83 breakpoint fragments (see appendix for further info) were barcoded for identification and sent to the NSC for HiFi-sequencing. Library preparations were done by NSC on about 25% of 8M SMRT cells on Sequel II instrument using Sequel II Binding kit 2.0 and Sequencing chemistry v2.0. The number of reads from sequencing the entire SMRT cell was 4 237 680, and the average polymerase read length was 65.4 kb. Loading was performed by diffusion, and the movie time was 30 hours with pre-extinction. Polymerase reads were demultiplexed with Demultiplexing pipeline on SMRT Link v10.2.0.133434. CCS reads were generated for demultiplexed polymerase reads and demultiplexed again using barcoded primer sequences. The number of HiFI reads provided by NSC was 459 241, and the mean length of the reads was 4 592 bp.

Table 21. *An overview over barcodes that would be used for identifying individuals after PacBio sequencing. The combination between the forward primer with barcode, and the reverse primer with barcode would be unique and identifiable when analyzing the data from the PacBio HiFi sequencing.*



Reverse/Forward	1005	1007	1008	1012	1015	1020	1022	1024
1033	AA1	AA7	AA13	AA1	AA7	AA13	AB1	AB7
1035	AA2	AA8	AA14	AA2	AA8	AA14	AB2	AB8
1044	AA3	AA9	AA15	AA3	AA9	AA15	AB3	AB9
1045	AA4	AA10	AA16	AA4	AA10	AA16	AB4	AB10
1054	AA5	AA11	AA17	AA5	AA11	AA17	AB5	AB11
1056	AA6	AA12	AA18	AA6	AA12	AA18	AB6	AB12
1057								
1059								
1060	BB1	BB7	BB13				AB1	AB7
1062	BB2	BB8	BB14				AB2	AB8
1065	BB3	BB9	BB15				AB3	AB9
1075	BB4	BB10	BB16				AB4	AB10
1076	BB5	BB11	BB17				AB5	AB11
1082	BB6	BB12	BB18				AB6	AB12
1083	BB1	BB7	BB13	AB1	AB7			
1089	BB2	BB8	BB14	AB2	AB8			
1096	BB3	BB9	BB15	AB3	AB9			
1098	BB4	BB10	BB16	AB4	AB10			
1100	BB5	BB11	BB17	AB5	AB11			
1101	BB6	BB12	BB18	AB6	AB12			
1105	AB1	AB2	AB3	AB4	AB5	AB6		
1107	AB7	AB8	AB9	AB10	AB11	AB12		
1110								
1112								

3.3 HiFi sequencing results

Out of the 83 PCR products that were sent for HiFi-sequencing, 79 breakpoint regions were successfully sequenced, by the NSC, and analyzed. All 79 sequences were mapped against gadMor3 and coastal reference genomes, resulting in a total of 158 CCS sequences with 99% read accuracy that were used to analyze for variation in the breakpoint region between- and within population. From figure 18, one can see an overview of the genes that are located close to the breakpoint region on the gadmor3.0 reference genome from the NCBI website. With the accurate long-read HiFi-sequences of the breakpoint region, one can get a better estimate on where on the gadMor3 region the breakpoint is, and the local gene synteny surrounding the breakpoints.

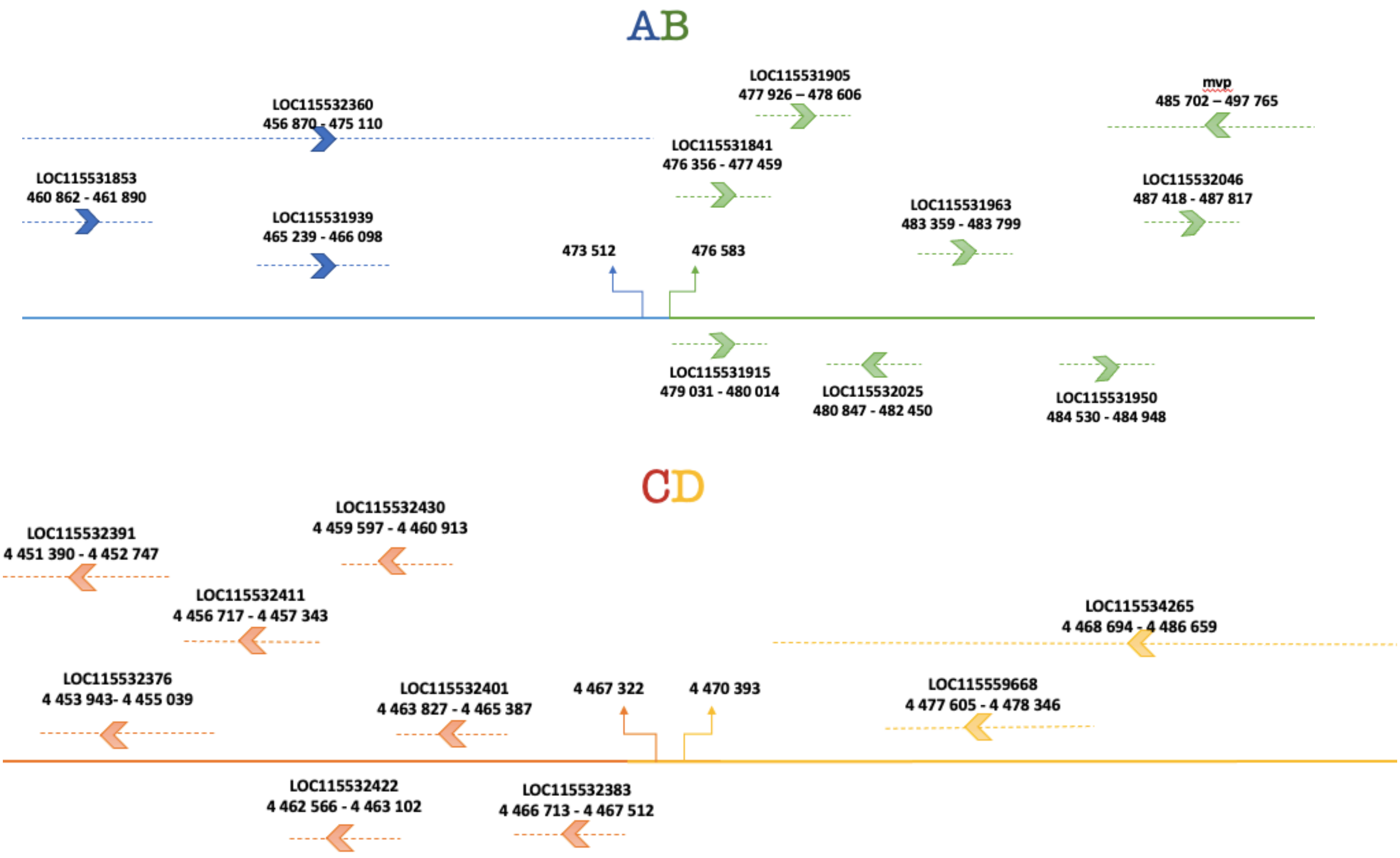
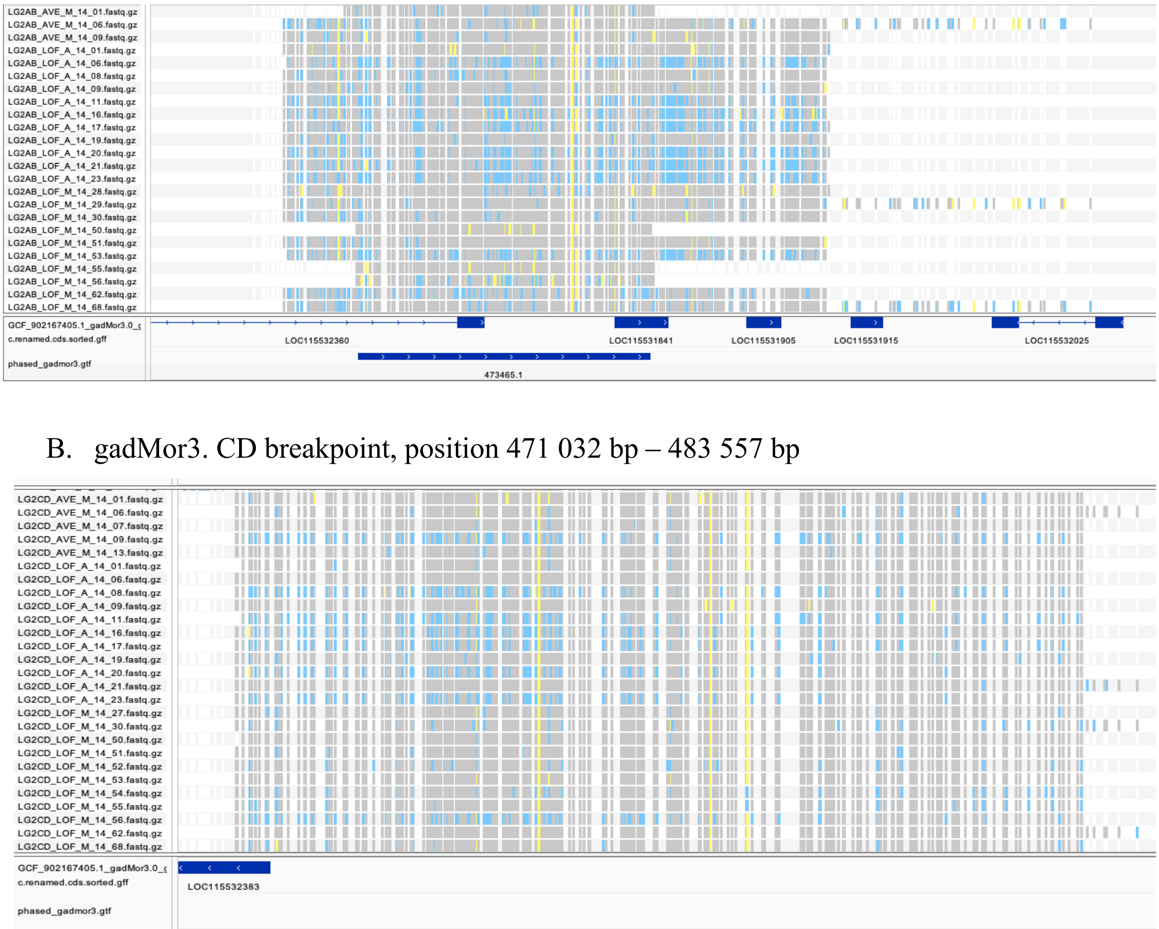


***Figure 18**. An overview over of the genes close to the breakpoint region on gadmor3 reference genome. The AB breakpoint region is located approximately between 473 512 bp and 476 583 bp, while the CD breakpoint region is between 4 467 322 bp and 4 470 393 bp. The dotted lines represent exons and introns, and the arrow indicates the direction*

3.3.1 Mapping to reference genomes and variant calling

Genomic visualization is essential for interpreting the HiFi-sequencing data, and for that IGV was used. Reads were divided between breakpoint regions and mapped towards either gadmor3 or coastal reference genome, depending on the breakpoint region (Figure 19 and 20). In addition to comparing the phased alignments to the reference genome, the annotation for each reference genome was also uploaded as a gff file to better visualize if variation was occurring close to a gene and/or regulatory areas. In doing so, the targeted regions were confirmed to have been sequenced and one could get an estimate of the variation within a breakpoint region compared to reference genome.

A. gadMor3. AB breakpoint, position 471 032 bp -483 557 bp

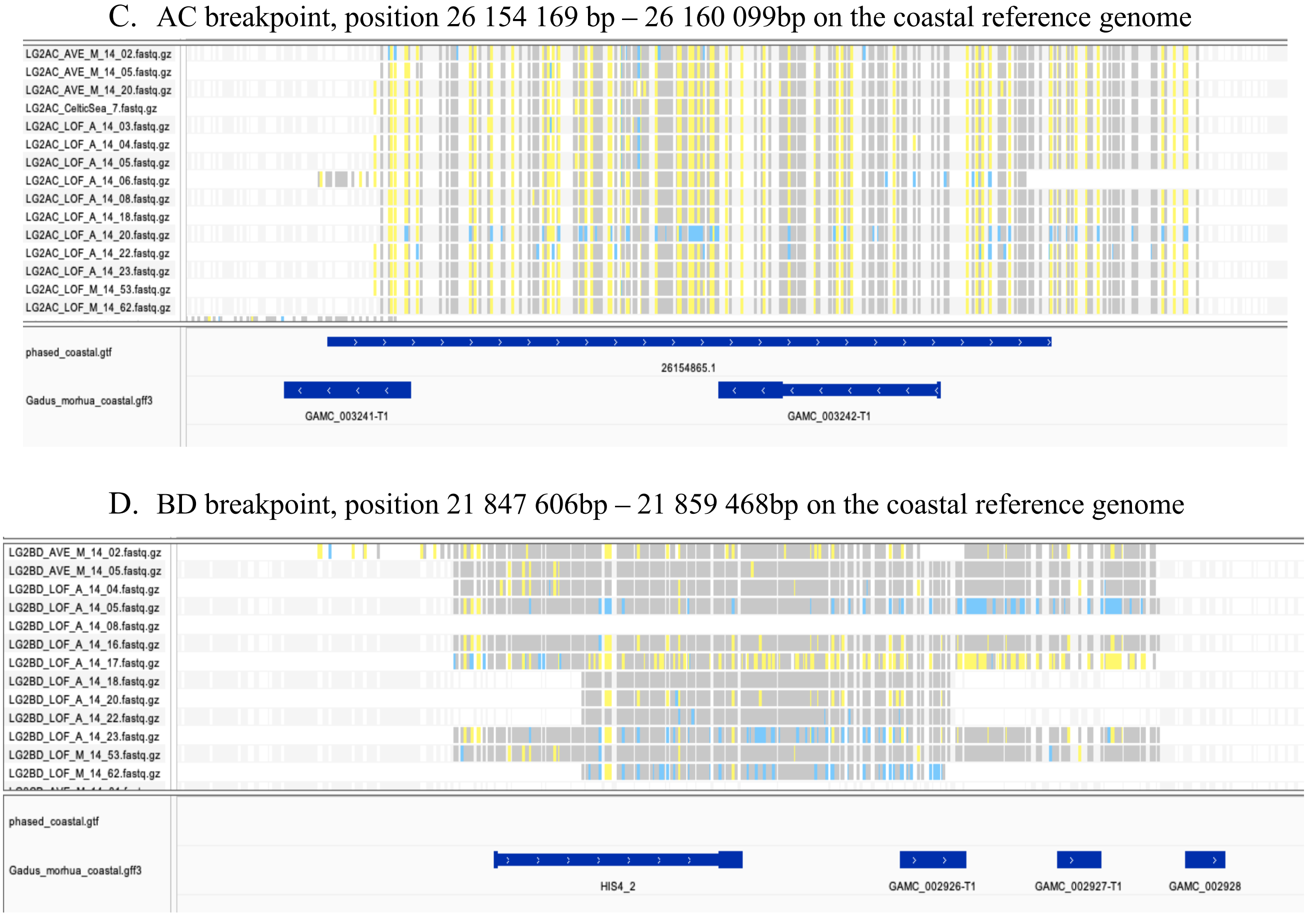


B. gadMor3. CD breakpoint, position 471 032 bp – 483 557 bp

LG2AB_AVE_M_14_01.fastq.gz
LG2AB_AVE_M_14_06.fastq.gz
LG2AB_AVE_M_14_09.fastq.gz
LG2AB_LOF_A_14_01.fastq.gz
LG2AB_LOF_A_14_06.fastq.gz
LG2AB_LOF_A_14_08.fastq.gz
LG2AB_LOF_A_14_09.fastq.gz
LG2AB_LOF_A_14_11.fastq.gz
LG2AB_LOF_A_14_16.fastq.gz
LG2AB_LOF_A_14_17.fastq.gz
LG2AB_LOF_A_14_19.fastq.gz
LG2AB_LOF_A_14_20.fastq.gz
LG2AB_LOF_A_14_21.fastq.gz
LG2AB_LOF_A_14_23.fastq.gz
LG2AB_LOF_M_14_28.fastq.gz
LG2AB_LOF_M_14_29.fastq.gz
LG2AB_LOF_M_14_30.fastq.gz
LG2AB_LOF_M_14_50.fastq.gz
LG2AB_LOF_M_14_51.fastq.gz
LG2AB_LOF_M_14_53.fastq.gz
LG2AB_LOF_M_14_55.fastq.gz
LG2AB_LOF_M_14_56.fastq.gz
LG2AB_LOF_M_14_62.fastq.gz
LG2AB_LOF_M_14_68.fastq.gz
GCF_902167405.1_gadMor3.0_g
c.renamed.cds.sorted.gff
phased_gadmor3.gtf
LOC115532360
LOC115531841
LOC115531905
LOC115531915
LOC115532025
473465.1
LG2CD_AVE_M_14_01.fastq.gz
LG2CD_AVE_M_14_06.fastq.gz
LG2CD_AVE_M_14_07.fastq.gz
LG2CD_AVE_M_14_09.fastq.gz
LG2CD_AVE_M_14_13.fastq.gz
LG2CD_LOF_A_14_01.fastq.gz
LG2CD_LOF_A_14_06.fastq.gz
LG2CD_LOF_A_14_08.fastq.gz
LG2CD_LOF_A_14_09.fastq.gz
LG2CD_LOF_A_14_11.fastq.gz
LG2CD_LOF_A_14_16.fastq.gz
LG2CD_LOF_A_14_17.fastq.gz
LG2CD_LOF_A_14_19.fastq.gz
LG2CD_LOF_A_14_20.fastq.gz
LG2CD_LOF_A_14_21.fastq.gz
LG2CD_LOF_A_14_23.fastq.gz
LG2CD_LOF_M_14_27.fastq.gz
LG2CD_LOF_M_14_30.fastq.gz
LG2CD_LOF_M_14_50.fastq.gz
LG2CD_LOF_M_14_51.fastq.gz
LG2CD_LOF_M_14_52.fastq.gz
LG2CD_LOF_M_14_53.fastq.gz
LG2CD_LOF_M_14_54.fastq.gz
LG2CD_LOF_M_14_55.fastq.gz
LG2CD_LOF_M_14_56.fastq.gz
LG2CD_LOF_M_14_62.fastq.gz
LG2CD_LOF_M_14_68.fastq.gz
GCF_902167405.1_gadMor3.0_g
c.renamed.cds.sorted.gff
phased_gadmor3.gtf
LOC115532383

Figure 19*. Here the HiFI-sequences of the AB- and CD breakpoint regions are aligned towards the gadMor3 reference genome. For these alignments blue represents heterozygous nonref, yellow represents homozygous nonref while gray represent homozygous ref. White indicates gaps.*

From calling variants using BCFtools mpileup one could get an estimate of the variation we find when mapping the reads toward the gadMor3 reference genome. The BCFtools mpileup was able to variant call all 79 samples and found 993 SNP, 68 indels, 44 multiallelic sites and 32 multiallelic SNP sites.


C. AC breakpoint, position 26 154 169 bp – 26 160 099bp on the coastal reference genome

LG2AC_AVE_M_14_02.fastq.gz
LG2AC_AVE_M_14_05.fastq.gz
LG2AC_AVE_M_14_20.fastq.gz
LG2AC_CelticSea_7.fastq.gz
LG2AC_LOF_A_14_03.fastq.gz
LG2AC_LOF_A_14_04.fastq.gz
LG2AC_LOF_A_14_05.fastq.gz
LG2AC_LOF_A_14_06.fastq.gz
LG2AC_LOF_A_14_08.fastq.gz
LG2AC_LOF_A_14_18.fastq.gz
LG2AC_LOF_A_14_20.fastq.gz
LG2AC_LOF_A_14_22.fastq.gz
LG2AC_LOF_A_14_23.fastq.gz
LG2AC_LOF_M_14_53.fastq.gz
LG2AC_LOF_M_14_62.fastq.gz
phased_coastal.gtf
26154865.1
Gadus_morhua_coastal.gff3
GAMC_003241-T1
GAMC_003242-T1

D. BD breakpoint, position 21 847 606bp – 21 859 468bp on the coastal reference genome

LG2BD_AVE_M_14_02.fastq.gz
LG2BD_AVE_M_14_05.fastq.gz
LG2BD_LOF_A_14_04.fastq.gz
LG2BD_LOF_A_14_05.fastq.gz
LG2BD_LOF_A_14_08.fastq.gz
LG2BD_LOF_A_14_16.fastq.gz
LG2BD_LOF_A_14_17.fastq.gz
LG2BD_LOF_A_14_18.fastq.gz
LG2BD_LOF_A_14_20.fastq.gz
LG2BD_LOF_A_14_22.fastq.gz
LG2BD_LOF_A_14_23.fastq.gz
LG2BD_LOF_M_14_53.fastq.gz
LG2BD_LOF_M_14_62.fastq.gz
phased_coastal.gtf
Gadus_morhua_coastal.gff3
HIS4_2
GAMC_002926-T1
GAMC_002927-T1
GAMC_002928

Figure 20. *Here the phased HiFi-sequences of the AC- and BD breakpoint region are aligned towards the cosatal reference genome. For these alignments blue represents heterozygous nonref, yellow represents homozygous nonref while gray represent homozygous ref. White indicates gaps.*

BCFtools mpilup was able to variant call all 79 alignments towards the coastal reference genome. The variation between the alignments and the coastal reference were found to be, using BCFtools statistics, 1395 SNP, 74 indels, 61 multiallelic sites and 49 multiallelic SNP sites. See appendix table 5A and 6A for further info regarding substitution types.

3.3.2 Phasing results

After the variant calling step, the sequences were phased using information from the vcf files – the output files from the variant calling step, as well as the BAM-files that were mapped towards gadMor3 (A/B and C/D) and the coastal (A/C and B/D) reference genome, to create *de novo* haplotypes assemblies. The phasing results can be seen in table 20, and both the phased gadMor3-based haplotypes and coastal-based haplotypes were used to reconstruct de *novo haplotype* assemblies. There were less single-nucleotide variants (SNVs) in the gadMor3 VCF file that contained the A/B and C/D haplotypes (992) than in the phased coastal VCF file that contained the A/C and B/D haplotypes (1344). Out of the 736 heterozygous SNVs detected in the gadMor3 VCF, 22 heterozygous variants are marked as phased while 714 are marked as unphased. While in the coastal VCF, out of the 1148 heterozygous SNVs detected in the input VCF, 23 were phased while 1125 were unphased.

Table 22*. A summary over the statistics from whathap phasing. The merged bam files of the breakpoint fragments were phased against gadmor3 reference genome and the coastal reference genome, resulting in two phased VCF files, which in this table is called gadMor3 and Coastal.*

	gadMor3	*Coastal*
Variants in VCF	992 SNV	1344 SNV
Heterozygous	736 (694 SNVs)	1148 (1100 SNVs)
Phased	22 (22 SNVs)	23 (23 SNVs)
Unphased	714	1125
Singletons	0	0
Blocks	1	1
Average block size (no. of variants	22. 00	23.00
Block lengths (bp)	3424	3567

3.4 Multiple sequence alignment

The breakpoint region haplotypes were aligned together using the Molecular Evolutionary Genomics Alignment version 11 (MEGA11) software (Tamura et al., 2021) and Geneious Prime, a software program downloaded from https://www.geneious.com. All the sequences were aligned using MUSCLE, a multiple sequence alignment tool, with default setting. All the multiple sequence alignments can be seen in appendix. To estimate the breakpoint regions and if the regions vary, each individual was compared with the putative breakpoint regions on the two reference genome assemblies, gadmor3 and coastal, respectfully.

3.4.1 Comparing the inverted and non-inverted haplotypes

3.4.1.1 AB/AC heterozygotes

As mentioned above, two programs were used for multiple sequence alignment. To get an estimation on where the breakpoint region occurs, on the gadmor3 reference genome, AB and AC sequences were aligned towards another, where A would align with A, but B and C would show variation. To better estimate the breakpoint region, heterozygous individuals were used. In figure 21 one can observe that variation occurs between A/B sequence (number 3 in the alignment) and the A/C sequence (number 2 in the alignment) at around site 4600 bp, on the consensus alignment, for the heterozygous individual LOF_A_14_09. In figure 21, one can see that the A/B, A/C, gadmor3- and coastal reference genome aligns with close to 100% identify until approximately site 4500 on the consensus, highlighted with a red arrow. After this site, more variation occurs between A/B and A/C. There is a conserved region that spans approximately 200 bp between all four alignments, between site 4000 and 4250 on the consensus alignment.

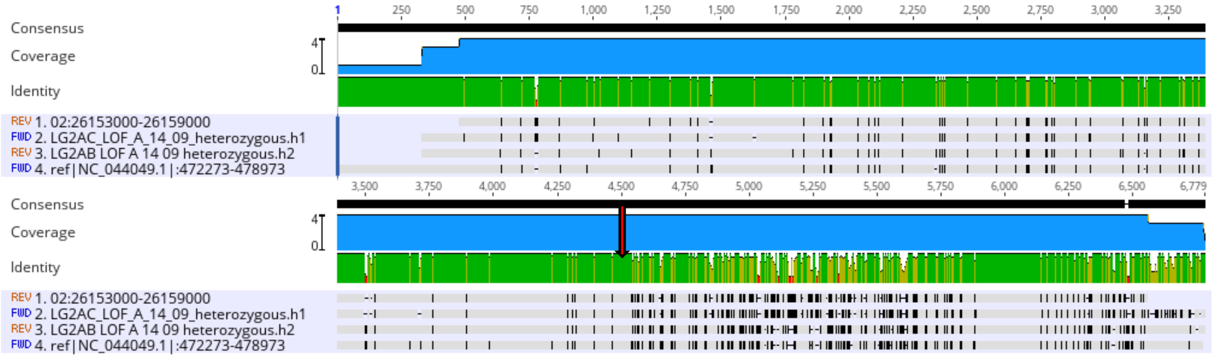
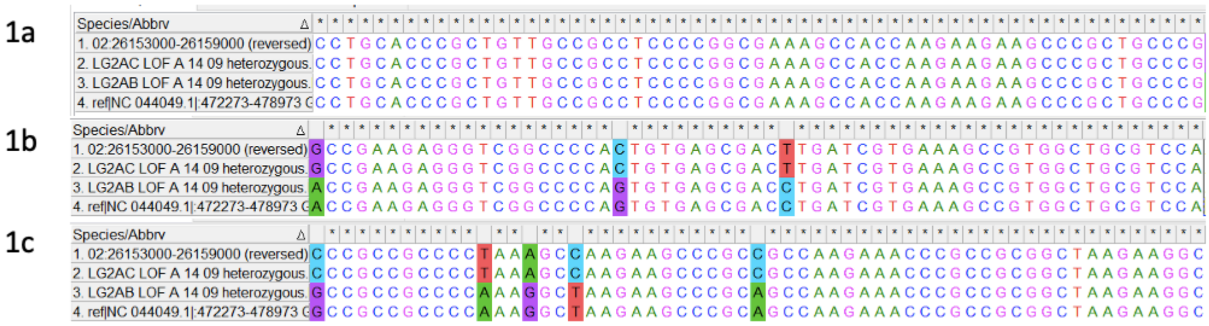


***Figure 21**. Alignment of breakpoint region A/B and A/C in the heterozygous individual, LOF_A_14_09 against the gadmor3, number 4 in the alignment, (472273bp-478973bp) and coastal, number 1 in the alignment, (26153000bp-26159000bp) reference genomes. The A/C breakpoint is number 2, while the A/B breakpoint is number 3.Green colour represent a 100% identity, green-brown 30-99% identity, and red represents < 30% identity. Red arrow indicates site 4500 on the consensus line, which is where C and B allele begins.*

The alignment from figure 21 was transferred to MEGA11 for multiple sequences alignment to visualize the conserved region that spans at approximately site 4000bp-4250bp on the consensus alignment, and the variation that occurs after (Figure 22). There is very little variation between LG2AB_LOF_A_14_09 and LG2AC_LOF_A_14_09 until site 4291 bp on the consensus alignment where AC and coastal genome starts to differ from both AB alignment and the gadmor3 genome. The AB breakpoint region is therefore estimated to be before site 476813 bp on the gadmor3 genome and site 26 154 795 bp on the coastal genome



```
Species/Abbrv
1. 02:26153000-26159000 (reversed) CCTGCACCCGCTGTTGCCGCCTCCCCGGCGAAAGCCACCAAGAAGAAGCCCGCTGCCCG
2. LG2AC LOF A 14 09 heterozygous. CCTGCACCCGCTGTTGCCGCCTCCCCGGCGAAAGCCACCAAGAAGAAGCCCGCTGCCCG
3. LG2AB LOF A 14 09 heterozygous. CCTGCACCCGCTGTTGCCGCCTCCCCGGCGAAAGCCACCAAGAAGAAGCCCGCTGCCCG
4. ref|NC 044049.1|:472273-478973 G CCTGCACCCGCTGTTGCCGCCTCCCCGGCGAAAGCCACCAAGAAGAAGCCCGCTGCCCG

Species/Abbrv
1. 02:26153000-26159000 (reversed) GCCGAAGAGGGTCGGCCCCACTGTGAGCGACTTGATCGTGAAAGCCGTGGCTGCGTCCA
2. LG2AC LOF A 14 09 heterozygous. GCCGAAGAGGGTCGGCCCCACTGTGAGCGACTTGATCGTGAAAGCCGTGGCTGCGTCCA
3. LG2AB LOF A 14 09 heterozygous. ACCGAAGAGGGTCGGCCCCAGTGTGAGCGACCTGATCGTGAAAGCCGTGGCTGCGTCCA
4. ref|NC 044049.1|:472273-478973 G ACCGAAGAGGGTCGGCCCCAGTGTGAGCGACCTGATCGTGAAAGCCGTGGCTGCGTCCA

Species/Abbrv
1. 02:26153000-26159000 (reversed) CCCGCCGCCCCTAAAGCCAAGAAGCCCGCCGCCAAGAAACCCGCCGCGGCTAAGAAGGC
2. LG2AC LOF A 14 09 heterozygous. CCCGCCGCCCCTAAAGCCAAGAAGCCCGCCGCCAAGAAACCCGCCGCGGCTAAGAAGGC
3. LG2AB LOF A 14 09 heterozygous. GCCGCCGCCCCAAAGGCTAAGAAGCCCGCAGCCAAGAAACCCGCCGCGGCTAAGAAGGC
4. ref|NC 044049.1|:472273-478973 G GCCGCCGCCCCAAAGGCTAAGAAGCCCGCAGCCAAGAAACCCGCCGCGGCTAAGAAGGC
```

***Figure 22**. 1a is site 4232bp-4290bp on the consensus line that shows the end of a conserved region between all alignments that spans approximately 200bp on all four sequences. The conserved region spans from 476260bp to 476486 bp on the gadmor3 reference genome and 26155007bp to 26154795 bp on the reversed coastal reference genome. 1b is the site 4291bp-4375bp on the consensus alignment and are the first signs of variation. 3c is site 4553bp-4611bp on the consensus alignment to further show that the A allele has ended.*

To continue analyzing if the A/B and A/C breakpoint region varies between individuals of the same population, two more heterozygous individuals were added to the multiple sequence alignment, LOF_A_14_11 and LOF_A_14_53 (Figure 24). First, we see the whole sequenced region is aligned towards each other with more than often 100% identity, which most likely is the A allele. Then there is a long stretch of a conserved region between site 4000bp and 4250 bp on the consensus line before there is significant variation between the A/B (number 3 in the alignment) and A/C (number 2 on in the alignment) that occurs at round 4500 bp on the consensus line. This region on the alignment is highlighted with a black box in figure 24. In these two additional individuals, the start of the diverging sequences corresponding to the B/C part of the alignments appears to be in the exact location. This would suggest that the breakpoint region might be conserved across these individuals. Additionally, in figure 26, this region appears remarkably preserved across all individuals.

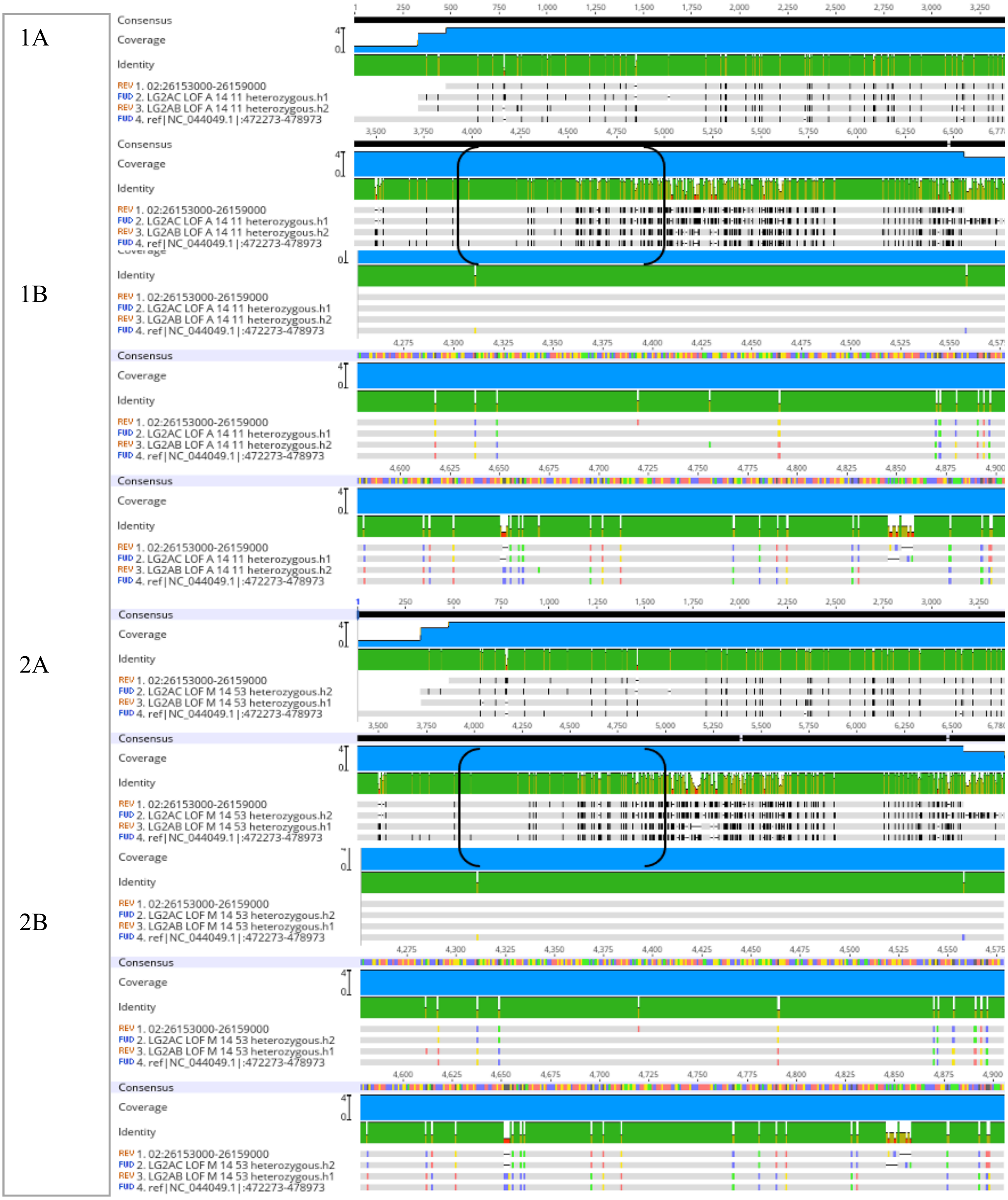


***Figure 24**. In 1A the two breakpoint regions A/B (number 3) and A/C (number 2) in the heterozygous individual LOF_A_14_11 is aligned towards the gadmor3 reference genome (number 4) and the coastal reference genome (number 1). In 1B, the area within the black square is zoomed in to better view the end of the A allele, and the start of the B and C allele, indicating that this is where the breakpoint is. As seen in 2A and 2B, this applies for the heterozygous individual LOF_A_14_53 as well*

3.4.1.2 AB/AC all individuals

The A/C breakpoint sequences within the homozygous individuals are shorter than the A/B sequences by roughly 2000bp. When comparing the two fragments, the A/C sequence spans at approximately 4000 bp, while the A/B sequence spans at around 6000 bp. However, the region corresponding to the B/D part of the alignment has a much higher divergence, which further establishes the breakpoint region. As seen in figure 26, the conserved region is not as large as when only looking at the three heterozygous individuals (Figure 24). However, as seen in figure 26, there is a region conserved within every individual at the area where the breakpoint has been estimated to be. From visual inspection, the heterozygous individuals for the inverted allele A (A/C breakpoint) shows two different haplotypes. While the individuals with the non-inverted allele B seems to have more random SNPs and some SNPs that is observed within all individuals.

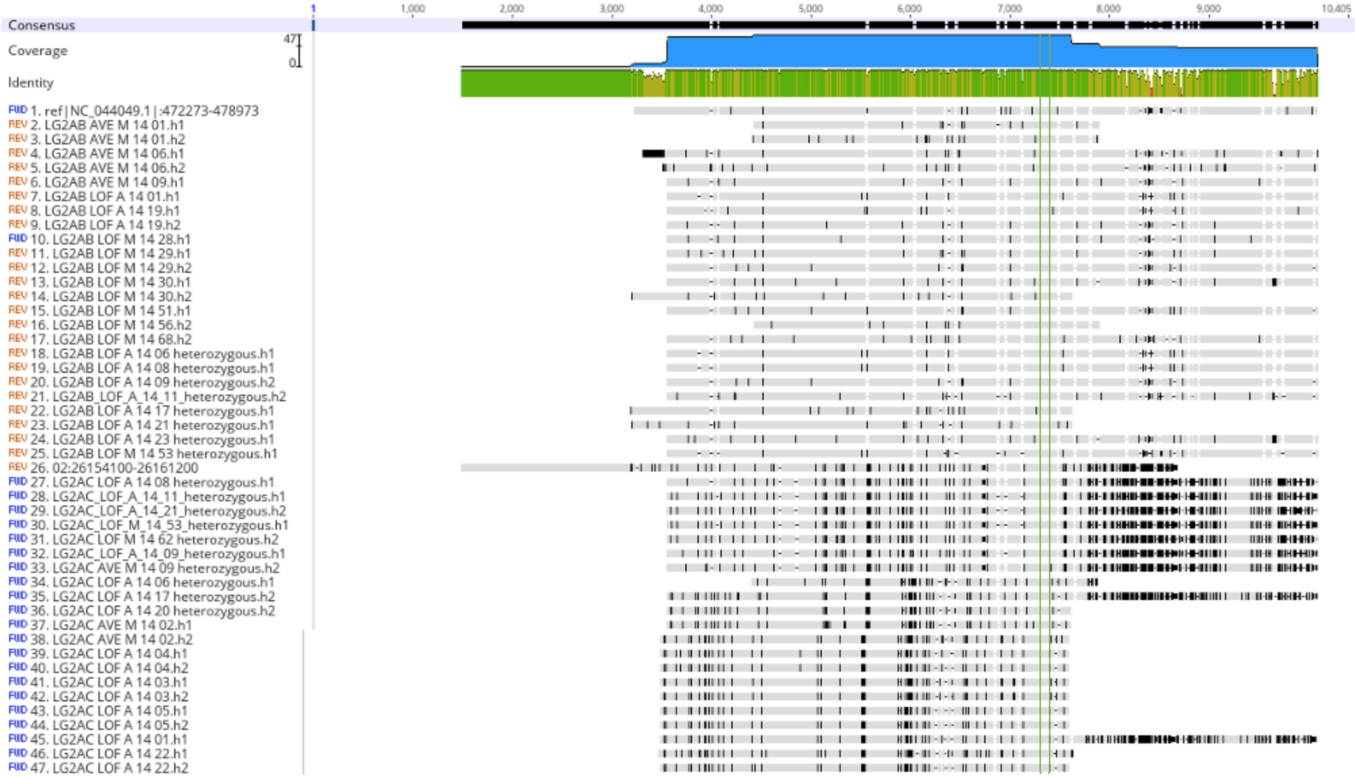

1
1,000
2,000
3,000
4,000
5,000
6,000
7,000
8,000
9,000
10,405
Consensus
Coverage
47
0
Identity
FWD 1. ref|NC_044049.1|:472273-478973
REV 2. LG2AB AVE M 14 01.h1
REV 3. LG2AB AVE M 14 01.h2
REV 4. LG2AB AVE M 14 06.h1
REV 5. LG2AB AVE M 14 06.h2
REV 6. LG2AB AVE M 14 09.h1
REV 7. LG2AB LOF A 14 01.h1
REV 8. LG2AB LOF A 14 19.h1
REV 9. LG2AB LOF A 14 19.h2
FWD 10. LG2AB LOF M 14 28.h1
REV 11. LG2AB LOF M 14 29.h1
REV 12. LG2AB LOF M 14 29.h2
REV 13. LG2AB LOF M 14 30.h1
REV 14. LG2AB LOF M 14 30.h2
REV 15. LG2AB LOF M 14 51.h1
REV 16. LG2AB LOF M 14 56.h2
REV 17. LG2AB LOF M 14 68.h2
REV 18. LG2AB LOF A 14 06 heterozygous.h1
REV 19. LG2AB LOF A 14 08 heterozygous.h1
REV 20. LG2AB LOF A 14 09 heterozygous.h2
REV 21. LG2AB_LOF_A_14_11_heterozygous.h2
REV 22. LG2AB LOF A 14 17 heterozygous.h1
REV 23. LG2AB LOF A 14 21 heterozygous.h1
REV 24. LG2AB LOF A 14 23 heterozygous.h1
REV 25. LG2AB LOF M 14 53 heterozygous.h1
REV 26. 02:26154100-26161200
FWD 27. LG2AC LOF A 14 08 heterozygous.h1
FWD 28. LG2AC_LOF_A_14_11_heterozygous.h1
FWD 29. LG2AC_LOF_A_14_21_heterozygous.h2
FWD 30. LG2AC_LOF_M_14_53_heterozygous.h1
FWD 31. LG2AC LOF M 14 62 heterozygous.h2
FWD 32. LG2AC_LOF_A_14_09_heterozygous.h1
FWD 33. LG2AC AVE M 14 09 heterozygous.h2
FWD 34. LG2AC LOF A 14 06 heterozygous.h1
FWD 35. LG2AC LOF A 14 17 heterozygous.h2
FWD 36. LG2AC LOF A 14 20 heterozygous.h2
FWD 37. LG2AC AVE M 14 02.h1
FWD 38. LG2AC AVE M 14 02.h2
FWD 39. LG2AC LOF A 14 04.h1
FWD 40. LG2AC LOF A 14 04.h2
FWD 41. LG2AC LOF A 14 03.h1
FWD 42. LG2AC LOF A 14 03.h2
FWD 43. LG2AC LOF A 14 05.h1
FWD 44. LG2AC LOF A 14 05.h2
FWD 45. LG2AC LOF A 14 01.h1
FWD 46. LG2AC LOF A 14 22.h1
FWD 47. LG2AC LOF A 14 22.h2

***Figure 26**. A multiple sequence alignment of the individuals with the non-inverted allele B (A/B) and the inverted allele A (A/C). The conserved region between the green lines, and highlighted, spans from 476260bp to 476486 bp on the gadMor3 reference genome and 26155007bp to 26154795 bp on the reversed coastal reference genome. This is the area, whether within or closer to 476490bp (gadmor3) the breakpoint is. The sequences after that are the B and C sequences.*

3.4.1.3 CD/BD heterozygotes

To get an estimation on where the CD/BD breakpoint region occurs, on the gadmor3 reference genome, CD and BD sequences were aligned towards another, where D would align with D, but C and B would show variation. Again, we used heterozygous individuals for this alignment to better estimate the breakpoint region. From figure 27, one can see that, compared to the AB/AC breakpoint, it was a bit more of a challenge to estimate the breakpoint region by looking at the alignment. For the heterozygous individual LOF_A_14_06, variation between the two alignments ends at approximately 5419 bp on the forward gadMor3 reference genome. For comparison, two more heterozygous individuals (LOF_A_14_11 and LOF_A_14_23) were aligned and can be seen in figure 28. All three individuals show the same pattern of multiple, but shorter, conservative regions with some SNPs in between these regions, before the D allele aligns. This makes it a bit more difficult to estimate when the D allele begins, indicating where the breakpoint region ends.

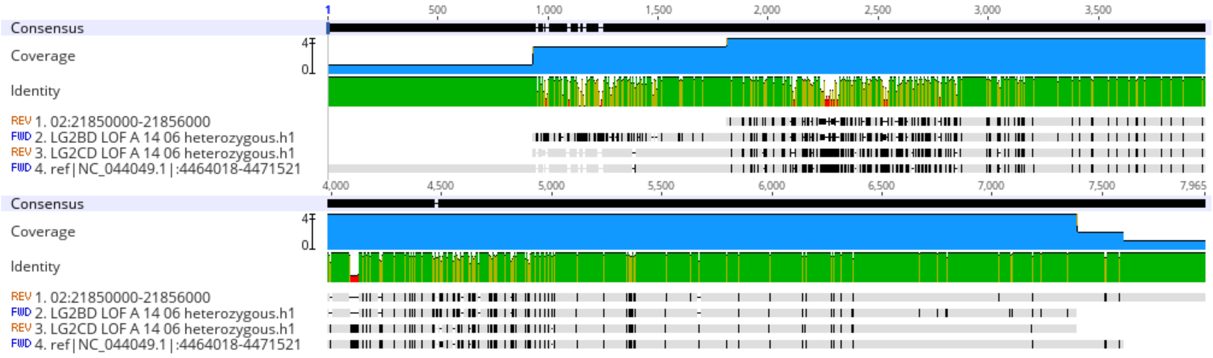

1
500
1,000
1,500
2,000
2,500
3,000
3,500
Consensus
Coverage
4
0
Identity
REV 1. 02:21850000-21856000
FWD 2. LG2BD LOF A 14 06 heterozygous.h1
REV 3. LG2CD LOF A 14 06 heterozygous.h1
FWD 4. ref|NC_044049.1|:4464018-4471521
4,000
4,500
5,000
5,500
6,000
6,500
7,000
7,500
7,965
Consensus
Coverage
4
0
Identity
REV 1. 02:21850000-21856000
FWD 2. LG2BD LOF A 14 06 heterozygous.h1
REV 3. LG2CD LOF A 14 06 heterozygous.h1
FWD 4. ref|NC_044049.1|:4464018-4471521

***Figure 27**. To estimate the CD/BD breakpoint region, a heterozygote individual was aligned against the gadmor3- and coastal reference genome. We expect to see the D allele align towards each other, while there´ll be variation between the C and B allele. The D allele alignment starts at approximately site 5000bp on the consensus alignment. The pairwise identity between the four alignments is 94,8% and 91,8% identical sites.*

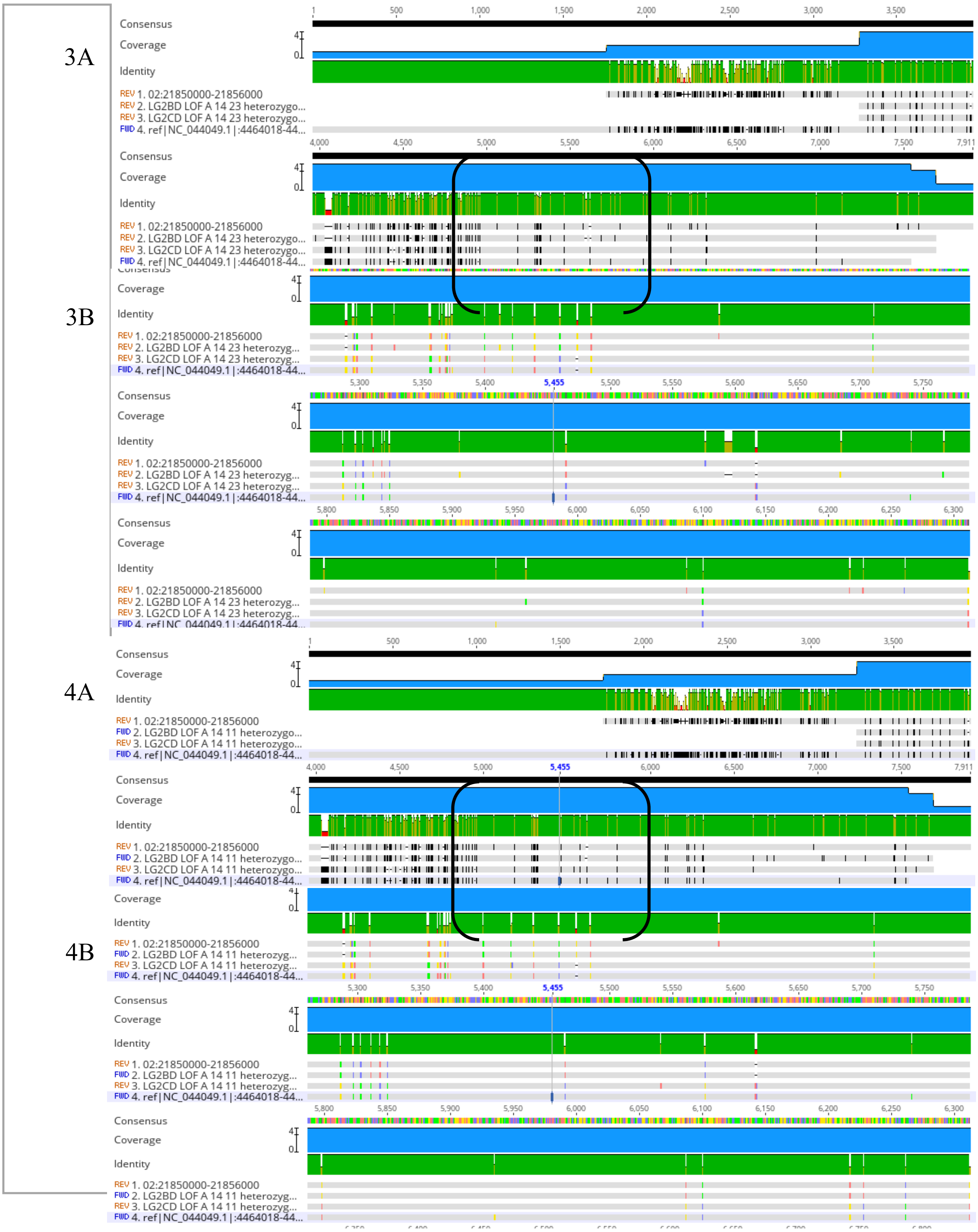
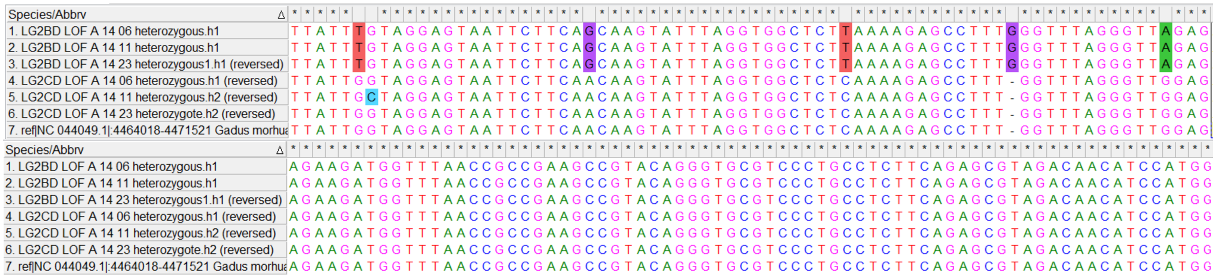


Figure 28. *In 3A the two breakpoint regions B/D (number 2) and C/D (number 3), in the heterozygous individual LOF_A_14_23 is aligned towards the gadmor3 reference genome (number 4) and the coastal reference genome (number 1). In 3B the, the area within the black square is zoomed in to better view the end of the B and C allele, and the start of the D allele, indicating that this is where the breakpoint is. As seen in 4A and 4B, this applies for the heterozygous individual LOF_A_14_11 as well.*



1a

```
Species/Abbrv
1. LG2BD LOF A 14 06 heterozygous.h1                   TTATTTGTAGGAGTAATTCTTCAGCAAGTATTTAGGTGGCTCTTAAAAGAGCCTTTGGGTTTAGGGTTAGAG
2. LG2BD LOF A 14 11 heterozygous.h1                   TTATTTGTAGGAGTAATTCTTCAGCAAGTATTTAGGTGGCTCTTAAAAGAGCCTTTGGGTTTAGGGTTAGAG
3. LG2BD LOF A 14 23 heterozygous1.h1 (reversed)       TTATTTGTAGGAGTAATTCTTCAGCAAGTATTTAGGTGGCTCTTAAAAGAGCCTTTGGGTTTAGGGTTAGAG
4. LG2CD LOF A 14 06 heterozygous.h1 (reversed)        TTATTGGTAGGAGTAATTCTTCAACAAGTATTTAGGTGGCTCTCAAAAGAGCCTTT-GGTTTAGGGTTGGAG
5. LG2CD LOF A 14 11 heterozygous.h2 (reversed)        TTATTGCTAGGAGTAATTCTTCAACAAGTATTTAGGTGGCTCTCAAAAGAGCCTTT-GGTTTAGGGTTGGAG
6. LG2CD LOF A 14 23 heterozygote.h2 (reversed)        TTATTGGTAGGAGTAATTCTTCAACAAGTATTTAGGTGGCTCTCAAAAGAGCCTTT-GGTTTAGGGTTGGAG
7. ref|NC 044049.1|:4464018-4471521 Gadus morhua       TTATTGGTAGGAGTAATTCTTCAACAAGTATTTAGGTGGCTCTCAAAAGAGCCTTT-GGTTTAGGGTTGGAG
```

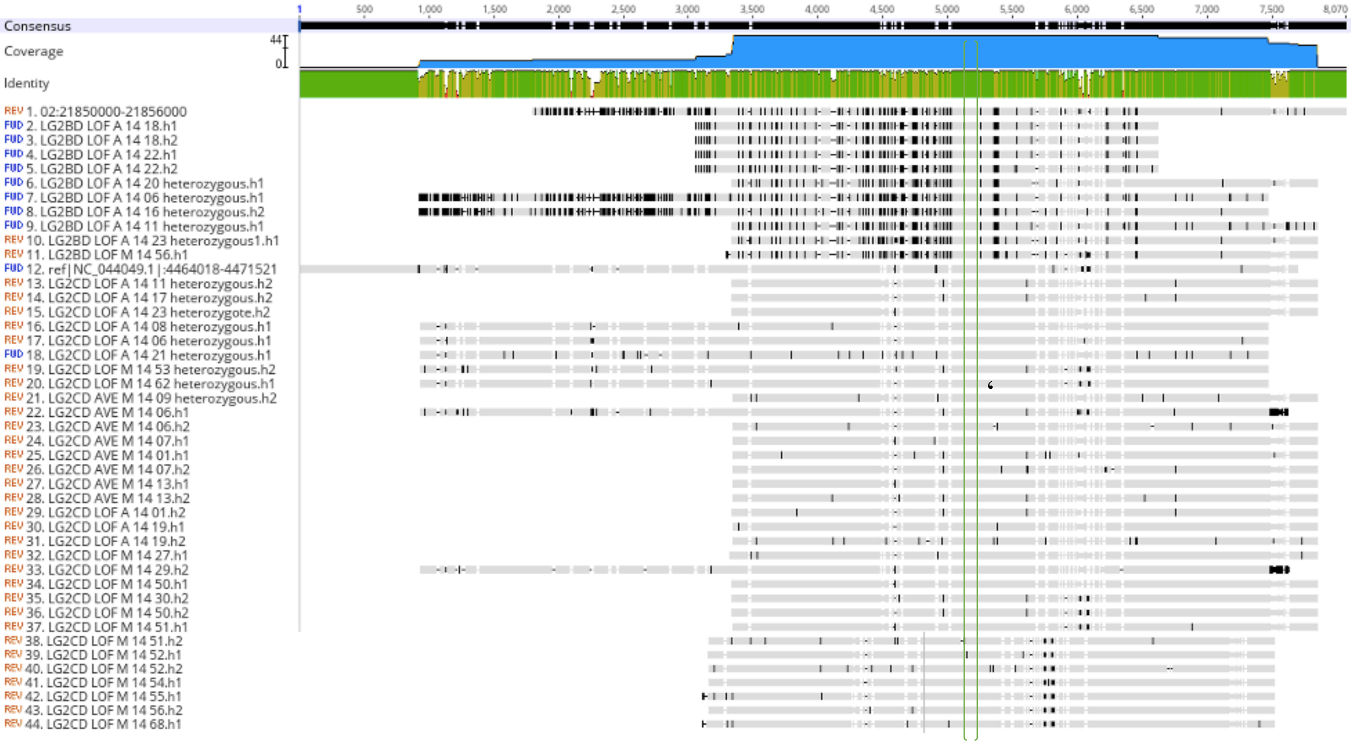
2b

```
Species/Abbrv
1. LG2BD LOF A 14 06 heterozygous.h1                   AGAAGATGGTTTAACCGCCGAAGCCGTACAGGGTGCGTCCCTGCCTCTTCAGAGCGTAGACAACATCCATGG
2. LG2BD LOF A 14 11 heterozygous.h1                   AGAAGATGGTTTAACCGCCGAAGCCGTACAGGGTGCGTCCCTGCCTCTTCAGAGCGTAGACAACATCCATGG
3. LG2BD LOF A 14 23 heterozygous1.h1 (reversed)       AGAAGATGGTTTAACCGCCGAAGCCGTACAGGGTGCGTCCCTGCCTCTTCAGAGCGTAGACAACATCCATGG
4. LG2CD LOF A 14 06 heterozygous.h1 (reversed)        AGAAGATGGTTTAACCGCCGAAGCCGTACAGGGTGCGTCCCTGCCTCTTCAGAGCGTAGACAACATCCATGG
5. LG2CD LOF A 14 11 heterozygous.h2 (reversed)        AGAAGATGGTTTAACCGCCGAAGCCGTACAGGGTGCGTCCCTGCCTCTTCAGAGCGTAGACAACATCCATGG
6. LG2CD LOF A 14 23 heterozygote.h2 (reversed)        AGAAGATGGTTTAACCGCCGAAGCCGTACAGGGTGCGTCCCTGCCTCTTCAGAGCGTAGACAACATCCATGG
7. ref|NC 044049.1|:4464018-4471521 Gadus morhua       AGAAGATGGTTTAACCGCCGAAGCCGTACAGGGTGCGTCCCTGCCTCTTCAGAGCGTAGACAACATCCATGG
```

Figure 29. *A multiple sequence alignment of the three heterozygous individuals: LOF_A_14_06, LOF_A_14_11 and LOF_A_14_23 from 4 468 955 bp (1a) to 4 469 106 bp (2b) on the gadmor3 reference genome. From this figure one can see that the variation is occurring at the same site and seems conserved.*

3.4.1.4 CD/BD all individuals

As mentioned above, the C/D and B/D breakpoint turned out to be a bit more difficult to locate as the region has multiple conserved sequences with some SNPs in between (see site 5 000 to 5 500 on figure 30). Using the heterozygous individuals, the breakpoint region was narrowed down to 4464018bp – 4471521bp on the gadMor3 reference genome and position 21850000bp – 21856000 bp on the reversed coastal reference genome (figure 28 and 29) and can be seen between the two green lines on figure 30.



```
Consensus
Coverage
Identity
REV 1. 02:21850000-21856000
FWD 2. LG2BD LOF A 14 18.h1
FWD 3. LG2BD LOF A 14 18.h2
FWD 4. LG2BD LOF A 14 22.h1
FWD 5. LG2BD LOF A 14 22.h2
FWD 6. LG2BD LOF A 14 20 heterozygous.h1
FWD 7. LG2BD LOF A 14 06 heterozygous.h1
FWD 8. LG2BD LOF A 14 16 heterozygous.h2
FWD 9. LG2BD LOF A 14 11 heterozygous.h1
REV 10. LG2BD LOF A 14 23 heterozygous1.h1
REV 11. LG2BD LOF M 14 56.h1
FWD 12. ref|NC_044049.1|:4464018-4471521
REV 13. LG2CD LOF A 14 11 heterozygous.h2
REV 14. LG2CD LOF A 14 17 heterozygous.h2
REV 15. LG2CD LOF A 14 23 heterozygote.h2
REV 16. LG2CD LOF A 14 08 heterozygous.h1
REV 17. LG2CD LOF A 14 06 heterozygous.h1
FWD 18. LG2CD LOF A 14 21 heterozygous.h1
REV 19. LG2CD LOF M 14 53 heterozygous.h2
REV 20. LG2CD LOF M 14 62 heterozygous.h1
REV 21. LG2CD AVE M 14 09 heterozygous.h2
REV 22. LG2CD AVE M 14 06.h1
REV 23. LG2CD AVE M 14 06.h2
REV 24. LG2CD AVE M 14 07.h1
REV 25. LG2CD AVE M 14 01.h1
REV 26. LG2CD AVE M 14 07.h2
REV 27. LG2CD AVE M 14 13.h1
REV 28. LG2CD AVE M 14 13.h2
REV 29. LG2CD LOF A 14 01.h2
REV 30. LG2CD LOF A 14 19.h1
REV 31. LG2CD LOF A 14 19.h2
REV 32. LG2CD LOF M 14 27.h1
REV 33. LG2CD LOF M 14 29.h2
REV 34. LG2CD LOF M 14 50.h1
REV 35. LG2CD LOF M 14 30.h2
REV 36. LG2CD LOF M 14 50.h2
REV 37. LG2CD LOF M 14 51.h1
REV 38. LG2CD LOF M 14 51.h2
REV 39. LG2CD LOF M 14 52.h1
REV 40. LG2CD LOF M 14 52.h2
REV 41. LG2CD LOF M 14 54.h1
REV 42. LG2CD LOF M 14 55.h1
REV 43. LG2CD LOF M 14 56.h2
REV 44. LG2CD LOF M 14 68.h1
```

Figure 30. *The multiple sequence alignment for the C/D and B/D breakpoint. It is not clear from this MSA where the breakpoint is. However, we do see conserved region between all individuals and one can hypothesize that the breakpoint is somewhere close (or in) the conserved area.*

3.5 Measures of sequence diversity within the breakpoints

The pairwise identity was detected (Table 23) for the different breakpoint haplotypes. The multiple sequence alignment for each breakpoint region was mapped towards their respective reference genome. The C/D breakpoint region is the most conserved region with highest pairwise identity of 99,5%. There were seven less sequences for the A/B breakpoint region and the pairwise identity was 99,2 %, making this breakpoint region also quite conserved.

Table 23. *Pairwise identities calculated after aligning multiple sequences of each breakpoint to the reference genome of the respected breakpoint.*

Breakpoint	Mapped R.G	Pairwise identity	Sequences
A/B	gadMor3 472026 - 479050	99,2 %	25
C/D	gadMor3 4465975-4472079	99,5 %	33
A/C	coastal 26155100-26159200	98,6 %	19
B/D	coastal 21851400-21854900	99,4 %	11
(A/B)/(A/C)	gadMor3 472026 - 479050	97,3 %	47
(C/D)/(B/D)	gadMor3 4465975-4472079	97,9 %	44

3.6 Phylogenetic analyses

The breakpoint alignments from the MSA (see chapter 3.4) were used to construct a phylogenetic tree using the Geneious tree builder (Figure 31 and 32). The genetic distance model used to build the trees was Tamura-Nei, and the tree building method was neighbour joining. There were no outgroups. From the phylogenetic tree of the inverted allele (Figure 32), we can see that there are two haplotypes for each breakpoint. There are not enough individuals to see a pattern on where these haplotypes differ (Figure 31 and 32). However, there seems to be a division between the heterozygote for the inverted allele, and those who are homozygote.

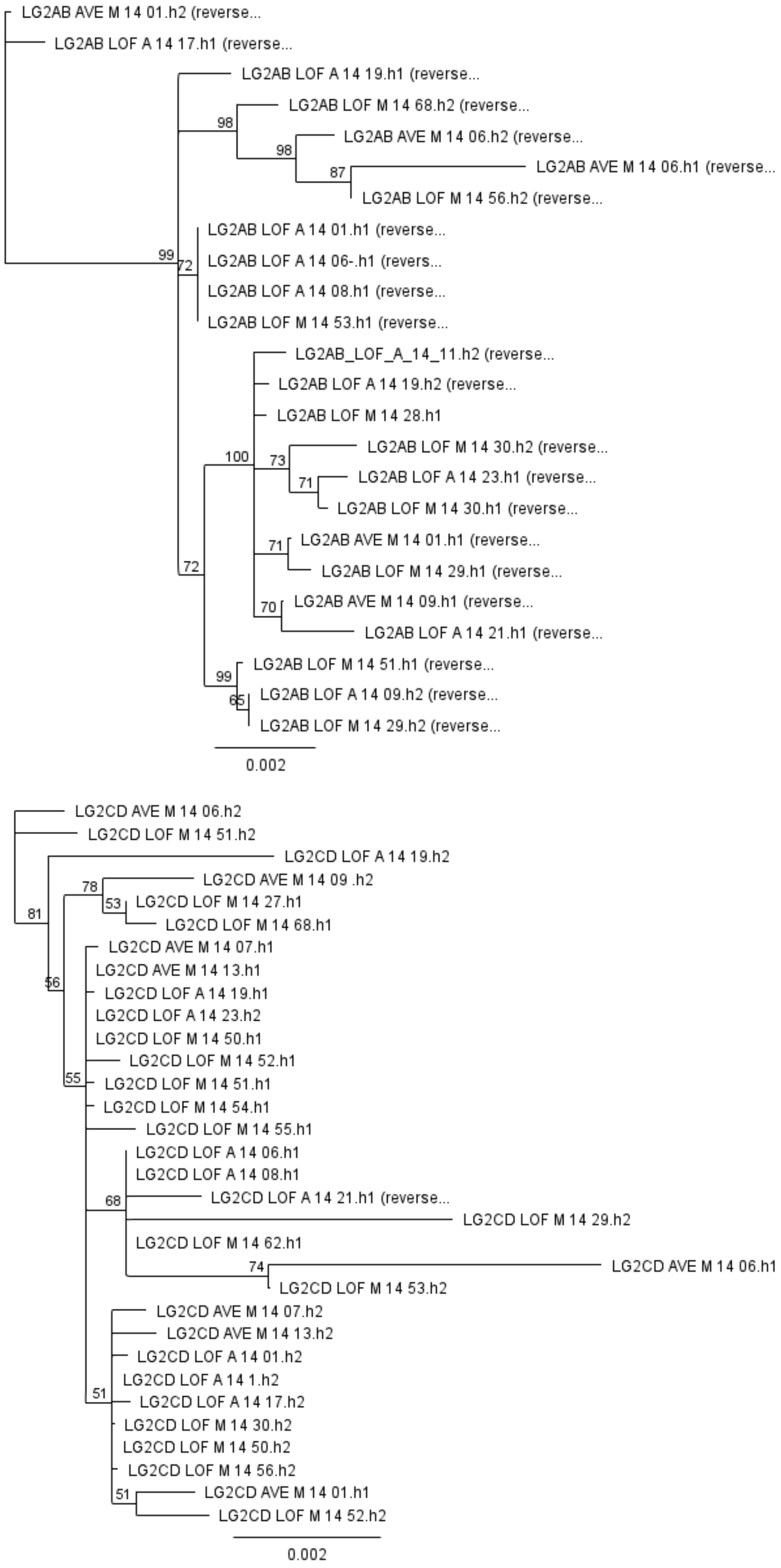

LG2AB AVE M 14 01.h2 (reverse...
LG2AB LOF A 14 17.h1 (reverse...
LG2AB LOF A 14 19.h1 (reverse...
98
LG2AB LOF M 14 68.h2 (reverse...
98
LG2AB AVE M 14 06.h2 (reverse...
87
LG2AB AVE M 14 06.h1 (reverse...
LG2AB LOF M 14 56.h2 (reverse...
LG2AB LOF A 14 01.h1 (reverse...
99
72
LG2AB LOF A 14 06-.h1 (revers...
LG2AB LOF A 14 08.h1 (reverse...
LG2AB LOF M 14 53.h1 (reverse...
LG2AB_LOF_A_14_11.h2 (reverse...
LG2AB LOF A 14 19.h2 (reverse...
LG2AB LOF M 14 28.h1
100
73
LG2AB LOF M 14 30.h2 (reverse...
71
LG2AB LOF A 14 23.h1 (reverse...
LG2AB LOF M 14 30.h1 (reverse...
71
LG2AB AVE M 14 01.h1 (reverse...
72
LG2AB LOF M 14 29.h1 (reverse...
70
LG2AB AVE M 14 09.h1 (reverse...
LG2AB LOF A 14 21.h1 (reverse...
99
LG2AB LOF M 14 51.h1 (reverse...
65
LG2AB LOF A 14 09.h2 (reverse...
LG2AB LOF M 14 29.h2 (reverse...
0.002
LG2CD AVE M 14 06.h2
LG2CD LOF M 14 51.h2
LG2CD LOF A 14 19.h2
78
LG2CD AVE M 14 09 .h2
81
53
LG2CD LOF M 14 27.h1
LG2CD LOF M 14 68.h1
LG2CD AVE M 14 07.h1
LG2CD AVE M 14 13.h1
56
LG2CD LOF A 14 19.h1
LG2CD LOF A 14 23.h2
LG2CD LOF M 14 50.h1
LG2CD LOF M 14 52.h1
55
LG2CD LOF M 14 51.h1
LG2CD LOF M 14 54.h1
LG2CD LOF M 14 55.h1
LG2CD LOF A 14 06.h1
LG2CD LOF A 14 08.h1
68
LG2CD LOF A 14 21.h1 (reverse...
LG2CD LOF M 14 29.h2
LG2CD LOF M 14 62.h1
74
LG2CD AVE M 14 06.h1
LG2CD LOF M 14 53.h2
LG2CD AVE M 14 07.h2
LG2CD AVE M 14 13.h2
LG2CD LOF A 14 01.h2
LG2CD LOF A 14 1.h2
51
LG2CD LOF A 14 17.h2
LG2CD LOF M 14 30.h2
LG2CD LOF M 14 50.h2
LG2CD LOF M 14 56.h2
51
LG2CD AVE M 14 01.h1
LG2CD LOF M 14 52.h2
0.002

***Figure 31**. A neighbour-joining phylogenetic three, with bootstrap support, of the A/B (top) and C/D (bottom) breakpoints*

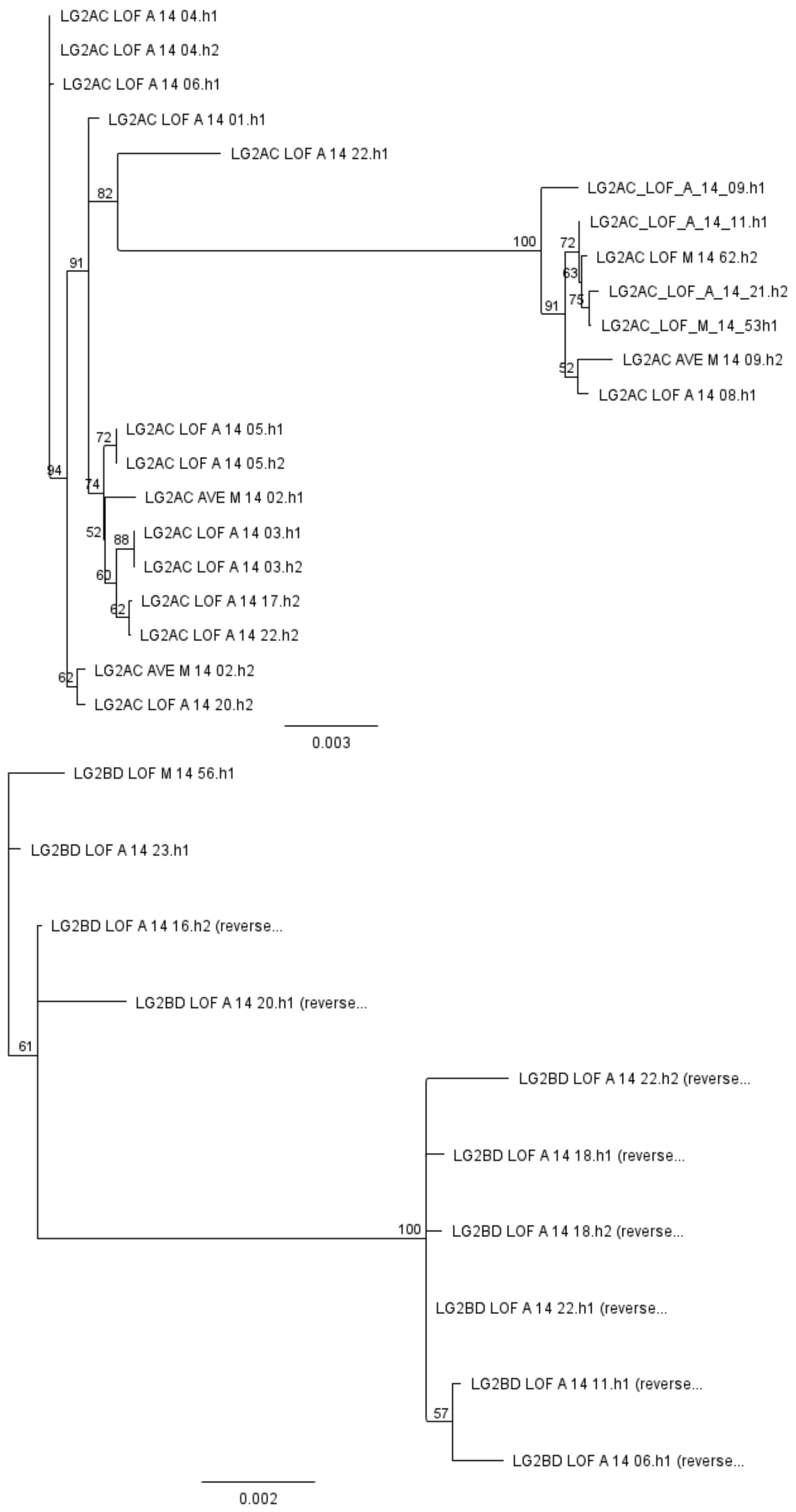


Figure 32*. A neighbour-joining phylogenetic three, with bootstrap support, of the A/C (top) and B/D (bottom) breakpoints.*

3.7 RepeatMasker

To analyzing the repeats that are in the breakpoint regions, the DNA sequences were screened for repeats using RepeatMasker. In each breakpoint region, long terminal repeat (LTR) retrotransposons were detected (Table 24). And in all the breakpoint regions, the LTR retrotransposons that were detected was the gipsy/DIRS1 group

***Table 24**. A summary of the repeats found on each breakpoint query. Each breakpoint was uploaded unto RepeatMaskern in FASTA format. The DNA sequences were screened for interspersed repeats and low complexity DNA sequences. The output file had a detailed annotation of the repeats that were present in the query sequences.*

	AB	*CD*	*AC*	*BD*
Sequences	26	37	22	8
GC level (%)	45,1	44,7	44,8	44,7
LTR: Gipsy/DIRS1	23	37	23	4
Simple repeats	3		0	2

4 Discussion

In my master thesis, I uncovered the breakpoints in inverted and non-inverted alleles in Atlantic cod to be conserved within and between populations. The A/B breakpoint has been narrowed down to be either within the conserved region of 476220 -476486 bp or closer to where variation occurs which is at position 476490 on the gadMor3 reference genome (Figure 24) The A/C-breakpoint on the inverted allele A has a 212bp conserved region between site 26 154 795 – 26 155 007 bp on the coastal reference genome 28. The breakpoint could be within this conserved region, or perhaps closer to where variation occurs which is on site 26 154 795 bp on the reversed sequence on the coastal reference genome.

The C/D and B/D breakpoints have more variation between populations on the breakpoint region, making it difficult to pinpoint where the breakpoint could be exactly. There is multiple, yet shorter, conserved regions that span from site 4 468 942bp on the forward gadmor3 reference genome. They are broken up with some SNPs that differ between those who have the inverted and non-inverted allele before the D sequence aligns for all individuals. Therefore, I estimate that the breakpoint region is somewhere between the conserved region, narrowing the area of breakpoint region to approximately 4 469 402bp – 4 469 513bp on the forward gadmor3 reference genome and at approximately 21 853 610bp – 21 853 498 bp on the reversed coastal reference genome.

To define the breakpoint region, I developed a PCR protocol for amplifying both inverted and non-inverted alleles to sequence the breakpoint regions within the Atlantic cod. In doing so, I could use multiple, accurate, HiFi sequencing reads to (1) align each breakpoint read to the gadmor3 and coastal reference genome and then (2) make *de novo* haplotype assemblies for the haplotypes surrounding the breakpoint regions for different individuals homozygous for the inverted or the non-inverted allele, or heterozygous, respectively. I could then estimate the breakpoint positions, analyse these highly haplotype resolved assemblies, and determine the variation in these breakpoint regions between populations and within the populations. I found that the breakpoint is conserved, with genetic variation building up as you move away from the breakpoint, within populations as well as between.

I will now discuss the protocol development and my findings in a broader context. First, I will discuss the population genomics of inversion breakpoints in Atlantic cod and the benefit of using PacBio sequencing for analyzing inversions. Secondly, I will discuss the challenges in the PCR protocol development and primer design regarding the repetitive and complex regions that spans these breakpoint areas. Lastly, I will discuss the implications of my findings and future perspectives.

4.1 Population genetics of inversion breakpoints

Population genetics, on the surface, is not a complicated idea; it is the study of how a population of the same species change genetically over time, leading to the species evolving. However, the population genetics of inversions are a different story. The genes and mutations associated with inversions are challenging to identify because of strong linkage disequilibrium within the inverted region (Huang & Rieseberg, 2020). Even so, with accurate long-read sequencing, it should be easier to pinpoint these genes within inversions for future analyses. Recombination plays an essential role in homogenizing nucleotide variability between homologous chromosomes. Therefore, chromosomal inversions can be an isolating mechanism between those who are homozygous for the inverted allele and those who are homozygous for the non-inverted (Figure 1) (Kirkpatrick, 2010). The inverted allele A, with the two breakpoints A/C and B/D, have proven beneficial as they have been selected for and prevailed in Atlantic cod. There is an expectation that regions near these inversion breakpoints have greater levels of differentiation because of the lack of genetic exchange between different gene arrangements. In contrast, the area within the breakpoint region is predicted to have lower levels of nucleotide differentiation due to greater levels of gene flux among different chromosomes. In my work, I found that the breakpoints of the derived inversion did not have significantly lower levels of nucleotide variability than breakpoints of ancestral inversion. The inverted alleles (the derived inversion) that were sequenced (A/C and B/D breakpoint) showed variability between individuals around the breakpoint region while having the same conserved region as all individuals had in the breakpoint area (Figure 26 and 30). I, therefore, want to challenge the hypothesis that the derived allele has less variation than the ancestral one. I found the breakpoints to be conserved and no recombination were detected.

Furthermore, I could detect two different haplotypes for the individuals who have the inverted allele on the A sequence through visual inspection on the multiple sequence alignment and phylogenetic tree (Figure 26 and 31). There is a need for more individuals to determine what the pattern is on how these haplotypes are divided, but from the individuals we have there seems to be a division between those who are heterozygote for the inverted allele and those who are homozygote. Also, it was primary coastal individuals, and since they are a smaller population (Dahle et al., 2018), it could be a sign of genetic drift. However, previous studies have done a genome-wide estimate of the temporal frequency change which suggested a large harmonic-mean effective population size (N_e), which suggest genetic drift within the Atlantic cod is weak (Pinsky et al., 2021).

There was a study done on inversion breakpoint regions on *Drosophila pseudoobscura (Wallace et al., 2013)* where they estimated nucleotide diversity at each breakpoint region to see whether ancestral inversion had higher levels of genetic variation at the breakpoint. Furthermore, they analyzed the levels of variation close to the breakpoint compared to distal segments. Their data indicates that variation was elevated near some breakpoints, but not in all, as I found in my data. Ancestral inversions failed to show greater levels of diversity than the derived inversions, even though breakpoint regions accumulate unique mutations. Breakpoints should, in theory, elevate levels of diversity and divergence, however I found that the breakpoints were conserved and similar. There seems to be linkage disequilibrium between the alleles of the breakpoint as they are being inherited together.

4.2 Advantages with HiFi-sequencing

Inversions, SVs in general, have substantial impact on evolution, but our understanding of inversions has been limited by technology. High-throughput short-read sequencing made it possible to sequence many genomes and has been an eminent tool to analyze genomic diversity (Altshuler et al., 2015). It has been used in multiple studies of transcription, gene regulation and epigenetics in many species (Henikoff et al., 2009). However, it does have its limitations, such as poor mapping to repetitive elements, limited ability to span indels or SVs and amplification artefacts during library construction (Sedlazeck et al., 2018). A lot of the SVs analysis we do today is based on short reads, though detecting SVs from short reads often suffers from low sensitivity (30-70%) and up to 85% false discovery (Sedlazeck et al., 2018). That is why I needed to do HiFi-sequencing to create accurate long reads of the breakpoint region. The inverted allele A on LG 2 was established, and the breakpoint area was narrowed down (Brieuc et. al in prep), even so, this region is highly repetitive and required high-throughput long-read sequencing to be able to span the whole breakpoint region. There are important advantages in using PacBio Hifi reads when creating *de novo* assemblies (Hon et al., 2020), compared to other long-read sequencing technologies such as Illumina . Which is that PacBio HiFi sequencing has a higher variant calling accuracy and variant calling confidence (Hon et al., 2020)

The primers were designed based on a few individuals from NoFima and two additional populations. However, these primers appear unsuitable for all cod populations, which probably explains why I was unable to successfully amplify the breakpoint regions from all the individuals available in the AquaGenome dataset, which would have included more Celtic individuals. However, I was able to sequence enough fragments for each breakpoint region to establish where the breakpoint is in both the NCC (coastal) and NEAC populations.

4.3 Methodological concerns

4.3.1 Developing the PCR-protocol

From developing the PCR protocol, I determined that to amplify multiple, but specific, fragments of the targeted region, one needed to do a two-step PCR. First a primary PCR round with DNA as template and Q5 as polymerase, followed by nested PCR with the primary PCR reaction as template and A2P as polymerase. This was concluded after multiple trial of amplifying fragments from one PCR session, but still low concentrations for fragments in the PCR reaction. What I established early on, was the difficulty of sequencing the breakpoint regions on LG1, 2 and 7. This could be because these are long stretches of repetitive DNA. However, it could also be because of primer choice. The DNA concentration was established to be good from using Qubit, therefore I continued to use the primary PCR reaction as a template and use primers within the amplified fragment for nested PCR. This yielded good and specific fragments. However, this resulted in long PCR runs that would lead to extra time in the lab. Because of this, I decided to continue focusing on one linkage group at a time to make sure that I would have good multiple fragments flanking the breakpoint region for the HiFi-sequencing. The idea when developing the PCR protocol was to establish a primer assay for rapid PCR sequencing of all the Atlantic cod individuals that were distributed from the Aqua Genome project. The primers that were designed for sequencing the breakpoint region was developed using only gadMor3 genome as reference, which does create a reference bias. Through the development of the PCR protocol, I discovered that even though the primer combination was

unique, it did not align well with all the individuals, which required an optimization of the PCR protocol. What I did not consider, when designing the primers, was population genetics.
The region around the breakpoint region varies enough between individuals, and especially between populations, that establishing only one primer combination assay would not be enough for sequencing multiple individuals. There is an expectation that regions near a breakpoint varies because of reduced genetic exchange between different gene arrangements (Wallace et al., 2013), and as mentioned earlier on, the primers were unique for these regions, but was only confirmed using the gadmor3 reference genome.

4.3.2 Repetitive regions

The benefit of analyzing the breakpoint region on a heterozygous individual is that we are comparing two alignments of the same individuals. So, the variation we see is the breakpoint variation, not variation because of ecotype. Using the multiple sequence alignment tool geneious, I could align the heterozygous individuals for both the inverted and non-inverted allele and compare the two alignments to see where the variation has occurred (Figure 21 and 27). In doing so, I could narrow down the breakpoint region and use that information when I align all the individuals for the A/B and A/C- and C/D and B/D breakpoint regions. While aligning the heterozygous individuals, I could confirm that the breakpoint region on these individuals is fixed, and there is minimal variation in these regions. However, there are still significant challenges in estimating breakpoint regions. The conserved regions surrounding the two breakpoints on LG 2 share sequence similarity, which in some cases, made it challenging to assign a sequence similarity and set a sequence to the A/C, A/B, C/D or B/D haplotype. Assembly algorithms have a hard time resolving repetitive regions, and have a much harder time making an assembly of sequencing data from short-reads technologies such as Illumina platform (Tørresen et al., 2017). This drawback was alleviated somewhat by my primers design and barcoding and the fact that we *a priori* knew which alleles (A or B) each individual carried (appendix figure 4A). To resolve this further, we would need to broaden the target region for sequencing, making both PCR and HiFI sequencing challenging. The targeted region that I was amplifying is a highly repetitive region which can be challenging to amplify. The Atlantic cod contains unusual high density of tandem repeats (TRs) compared to other vertebrates (Tørresen et al., 2017). From using the RepeatMasker, I discovered a transposable element in all the four breakpoint regions, retrotransposons LTR (Table 24). Transposable elements can induce a variety of chromosomal rearrangements, such as inversions (Sharma et al., 2021). Most likely there are more, but I only found the LTR that is annotated.

Deep oxford nanopore sequencing of LG 2 would possibly resolve even long and complex repetitive regions. Oxford Nanopore Technologies (ONT) PromethlON platform can produce even longer reads than HiFi-sequencing (up to 4 Mbp), which could enable us to flank an even larger breakpoint region and perhaps avoid the very repetitive areas. ONT does have a higher throughput at a lower cost, however, it produces less accurate reads than Sequel II system (De Coster et al., 2021)

Moreover, a different approach that I could have used for amplifying the breakpoint region, other than the traditional capture and PCR amplicons, is the CRIPR/Cas9 system and in silico sequencer-based selection (De Coster et al., 2021). These sequencing methods typically target 10-20 kbp regions, and the Cas9 system enriches a region without amplification. It can thus enable the assessment of methylation patterns and sequences that are hard to target, such as repeats (De Coster et al., 2021). These methods work well for PacBio sequencing. However, that was beyond the scope of this thesis.

5. Conclusion and future perspectives

What is the relationship between local adaptation and genome rearrangement within the Atlantic cod? I developed a PCR protocol through this study to amplify all four breakpoint regions of interest in LG 2 in Atlantic cod. In doing so, I conducted *de novo* haplotype assemblies of these regions and pinpointed where the breakpoints are and that they are conserved and fixed within and between populations. The population genetic data I have generated can potentially be used for further analyses, such as investigating the origin of the inversion and what regulatory regions are affected by the breakpoint. One can potentially better understand local adaptations and the mechanisms behind them. With the population data we now have, it is possible to understand better the genetic information regarding differences in behavioural ecology in different populations. Four genomic inversions in the Atlantic cod populations act as supergenes and have been associated with habitat differences in salinity, oxygen, and temperature (Barth et al., 2017; Berg et al., 2015; Berg et al., 2017; Berg et al., 2016; Kirubakaran et al., 2016; Matschiner et al., 2022). It shows that several regions of the Atlantic cod genomes are candidates for selection, and most of these regions are associated with local adaptations. Chromosomal inversions play a crucial role in maintaining diverging genomic regions, and I have through my thesis proved the possibility in amplifying complex regions and using accurate long-read sequencing to analyse inverted and non-inverted regions.

One main challenge in population-level studies is a scalable and streamlined analysis. What I did in my research is only one of many possible methods to analyse long-read sequencing. We are in a rapidly developing area of genomics, and new tools for population-level studies are constantly introduced. Nevertheless, we now have multiple *de novo* haplotype assemblies spanning the breakpoints and a better estimate of where the breakpoints are. Accurate long-read sequencing, such as HiFi-sequencing, is the key to identifying hidden SVs. With the advances in sequencing technology and bioinformatics, we are just getting started on achieving long-read sequencing on a population scale.

6. References

Altshuler, D. M., Albers, C. A., Abecasis, G. R., & et al. (2015). A global reference for human genetic variation. *Nature, 526*(7571), 68-74. https://doi.org/10.1038/nature15393

Atlantic cod : the bio-ecology of the fish. (2019). (First edition. ed.). Wiley Blackwell.

Barth, J. M. I., Berg, P. R., Jonsson, P. R., Bonanomi, S., Corell, H., Hemmer-Hansen, J., Jakobsen, K. S., Johannesson, K., Jorde, P. E., Knutsen, H., Moksnes, P.-O., Star, B., Stenseth, N. C., Svedäng, H., Jentoft, S., & André, C. (2017). Genome architecture enables local adaptation of Atlantic cod despite high connectivity. *Molecular Ecology, 26*(17), 4452-4466. https://doi.org/10.1111/mec.14207

Barth, J. M. I., Villegas-Ríos, D., Freitas, C., Moland, E., Star, B., André, C., Knutsen, H., Bradbury, I., Dierking, J., Petereit, C., Righton, D., Metcalfe, J., Jakobsen, K. S., Olsen, E. M., & Jentoft, S. (2019). Disentangling structural genomic and behavioral barriers in a sea of connectivity.

Berdan, E. L., Mérot, C., Pavia, H., Johannesson, K., Wellenreuther, M., & Butlin, R. K. (2021). A large chromosomal inversion shapes gene expression in seaweed flies (Coelopa frigida). *Evol Lett, 5*(6), 607-624. https://doi.org/10.1002/evl3.260

Berg, P. R., Jentoft, S., Star, B., Ring, K. H., Knutsen, H., Lien, S., Jakobsen, K. S., & André, C. (2015). Adaptation to Low Salinity Promotes Genomic Divergence in Atlantic Cod (Gadus morhua L.). https://doi.org/https://doi.org/10.1093/gbe/evv093

Berg, P. R., Star, B., Pampoulie, C., Bradbury, I. R., Bentzen, P., Hutchings, J., Jentoft, S., & Jakobsen, K. S. (2017). Trans-oceanic genomic divergence of Atlantic cod ecotypes is associated with large inversions. *Berg, Paul R. (2017) Genomic divergence in Atlantic cod populations. Doctoral thesis. http://urn.nb.no/URN:NBN:no-57964*.

Berg, P. R., Star, B., Pampoulie, C., Sodeland, M., Barth, J. M. I., Knutsen, H., Jakobsen, K. S., & Jentoft, S. (2016). Three chromosomal rearrangements promote genomic divergence between migratory and stationary ecotypes of Atlantic cod. *Sci Rep, 6*(1), 23246-23246. https://doi.org/10.1038/srep23246

BioLabsinc, N. E. *Q5® High-Fidelity DNA Polymerases*. New England BioLabs inc. https://international.neb.com/products/pcr-qpcr-and-amplification-technologies/q5-high-fidelity-dna-polymerases/q5-high-fidelity-dna-polymerases

Black, D., & Shuker, D. M. (2019). Supergenes. *Curr Biol, 29*(13), R615-R617. https://doi.org/10.1016/j.cub.2019.05.024

Butlin, R. K., Read, I. L., & Day, T. H. (1982). The effects of a chromosomal inversion on adult size and male mating success in the seaweed fly, coelopa frigida. *Heredity, 49*(1), 51-62. https://doi.org/10.1038/hdy.1982.64

Chouinard, G., & Fréchet, A. (1994, 01/01). Fluctuations in the cod stocks of the Gulf of St. Lawrence. *ICES Mar. Sci. Symp., 198*, 121-139.

Clark, D. P., & Pazdernik, N. J. (2013). *Molecular Biology* (Second edition ed.). AP cell press.

Dahle, G., Quintela, M., Johansen, T., Westgaard, J.-I., Besnier, F., Aglen, A., Jørstad, K. E., & Glover, K. A. (2018, 2018/07/09). Analysis of coastal cod (Gadus morhua L.) sampled on spawning sites reveals a genetic gradient throughout Norway's coastline. *BMC Genetics, 19*(1), 42. https://doi.org/10.1186/s12863-018-0625-8

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience, 10*(2). https://doi.org/10.1093/gigascience/giab008

De Coster, W., Weissensteiner, M. H., & Sedlazeck, F. J. (2021). Towards population-scale long-read sequencing. *Nat Rev Genet, 22*(9), 572-587. https://doi.org/10.1038/s41576-021-00367-3

Futuyma, D. J., & Kirkpatrick, M. (2017). *Evolution* (4th ed. ed.). Sinauer.

Henikoff, S., MacAlpine, D. M., Stein, L., Snyder, M., Lieb, J. D., Celniker, S. E., Dillon, L. A. L., White, K. P., Waterston, R. H., Micklem, G., Lai, E. C., Karpen, G. H., Gerstein, M. B., Gunsalus, K. C., Piano, F., & Kellis, M. (2009). Unlocking the secrets of the genome. *Nature, 459*(7249), 927-930. https://doi.org/10.1038/459927a

Hon, T., Mars, K., Young, G., Tsai, Y.-C., Karalius, J. W., Landolin, J. M., Maurer, N., Kudrna, D., Hardigan, M. A., Steiner, C. C., Knapp, S. J., Ware, D., Shapiro, B., Peluso, P., & Rank, D. R. (2020, 2020/11/17). Highly accurate long-read HiFi sequencing data for five complex genomes. *Scientific Data, 7*(1), 399. https://doi.org/10.1038/s41597-020-00743-4

Jay, P., Whibley, A., Frézal, L., Rodríguez de Cara, M. Á., Nowell, R. W., Mallet, J., Dasmahapatra, K. K., & Joron, M. (2018). Supergene Evolution Triggered by the Introgression of a Chromosomal Inversion. *Curr Biol, 28*(11), 1839-1845.e1833. https://doi.org/10.1016/j.cub.2018.04.072

Jonsson, P. R., Corell, H., André, C., Svedäng, H., & Moksnes, P.-O. (2016). Recent decline in cod stocks in the North Sea–Skagerrak–Kattegat shifts the sources of larval supply. *Fisheries Oceanography, 25*(3), 210-228. https://doi.org/https://doi.org/10.1111/fog.12146

Kirkpatrick, M., & Barrett, B. (2015). Chromosome inversions, adaptive cassettes and the evolution of species' ranges. *Mol Ecol, 24*(9), 2046-2055. https://doi.org/10.1111/mec.13074

Kirubakaran, T. G., Grove, H., Kent, M. P., Sandve, S. R., Baranski, M., Nome, T., De Rosa, M. C., Righino, B., Johansen, T., Otterå, H., Sonesson, A., Lien, S., & Andersen, Ø. (2016). Two adjacent inversions maintain genomic differentiation between migratory and stationary ecotypes of Atlantic cod. *Mol Ecol, 25*(10), 2130-2143. https://doi.org/10.1111/mec.13592

Kolmogorov, M., Bickhart, D. M., Behsaz, B., Gurevich, A., Rayko, M., Shin, S. B., Kuhn, K., Yuan, J., Polevikov, E., Smith, T. P. L., & Pevzner, P. A. (2020, 2020/11/01). metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nature Methods, 17*(11), 1103-1110. https://doi.org/10.1038/s41592-020-00971-x

Küpper, C., Stocks, M., Risse, J. E., Dos Remedios, N., Farrell, L. L., McRae, S. B., Morgan, T. C., Karlionova, N., Pinchuk, P., Verkuil, Y. I., Kitaysky, A. S., Wingfield, J. C., Piersma, T., Zeng, K., Slate, J., Blaxter, M., Lank, D. B., & Burke, T. (2016). A supergene determines highly divergent male reproductive morphs in the ruff. *Nat Genet, 48*(1), 79-+. https://doi.org/10.1038/ng.3443

Lank, D. B., Farrell, L. L., Burke, T., Piersma, T., & McRae, S. B. (2013). A dominant allele controls development into female mimic male and diminutive female ruffs. *Biol Lett, 9*(6), 20130653. https://doi.org/10.1098/rsbl.2013.0653

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

Martin, M., Patterson, M., Garg, S., O Fischer, S., Pisanti, N., Klau, G. W., Schöenhuth, A., & Marschall, T. (2016). WhatsHap: fast and accurate read-based phasing. *bioRxiv*, 085050. https://doi.org/10.1101/085050

Matschiner, M., Barth, J. M. I., Tørresen, O. K., Star, B., Baalsrud, H. T., Brieuc, M. S. O., Pampoulie, C., Bradbury, I., Jakobsen, K. S., & Jentoft, S. (2022). Supergene origin and maintenance in Atlantic cod. *Nat Ecol Evol*. https://doi.org/10.1038/s41559-022-01661-x

Michalsen, K., Johannesen, E., & Bogstad, B. (2008). Feeding of mature cod (Gadus morhua) on the spawning grounds in Lofoten. *ICES journal of marine science, 65*(4), 571-580. https://doi.org/10.1093/icesjms/fsn019

Mieszkowska, N., Genner, M. J., Hawkins, S. J., & Sims, D. W. (2009). Chapter 3 Effects of Climate Change and Commercial Fishing on Atlantic Cod Gadus morhua. In *Advances in Marine Biology* (Vol. 56, pp. 213-273). Academic Press. https://doi.org/https://doi.org/10.1016/S0065-2881(09)56003-8

Noor, M. A. F., Garfield, D. A., Schaeffer, S. W., & Machado, C. A. (2007). Divergence Between the Drosophila pseudoobscura and D. persimilis Genome Sequences in Relation to Chromosomal Inversions. *Genetics, 177*(3), 1417-1428. https://doi.org/10.1534/genetics.107.070672

Pinsky, M. L., Eikeset, A. M., Helmerson, C., Bradbury, I. R., Bentzen, P., Morris, C., Gondek-Wyrozemska, A. T., Baalsrud, H. T., Brieuc, M. S. O., Kjesbu, O. S., Godiksen, J. A., Barth, J. M. I., Matschiner, M., Stenseth, N. C., Jakobsen, K. S., Jentoft, S., & Star, B. (2021). Genomic stability through time despite decades of exploitation in cod on both sides of the Atlantic. *Proc Natl Acad Sci U S A, 118*(15), 1. https://doi.org/10.1073/pnas.2025453118

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011, 2011/01/01). Integrative genomics viewer. *Nature Biotechnology, 29*(1), 24-26. https://doi.org/10.1038/nbt.1754

Rodríguez-Ramilo, S. T., Baranski, M., Moghadam, H., Grove, H., Lien, S., Goddard, M. E., Meuwissen, T. H. E., & Sonesson, A. K. (2019). Strong selection pressures maintain divergence on genomic islands in Atlantic cod (Gadus morhua L.) populations. *Genet Sel Evol, 51*(1), 61-61. https://doi.org/10.1186/s12711-019-0503-5

Rose, G. A. (2019). *Atlantic cod : the bio-ecology of the fish* (First edition. ed.). Wiley Blackwell.

Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A., & Arnheim, N. (1985). Enzymatic Amplification of β-Globin Genomic Sequences and Restriction Site Analysis for Diagnosis of Sickle Cell Anemia. *Science, 230*(4732), 1350-1354. https://doi.org/10.1126/science.2999980

Scientific, T. *The Power of Two—Fusion DNA polymerases*. ThermoFisher SCIENTIFIC. https://www.thermofisher.com/no/en/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/spotlight-articles/the-power-of-two-fusion-dna-polymerases.html?ef_id=CjwKCAjwquWVBhBrEiwAt1Kmwq8RGSsrnEiZdPdyMZrlE4qy4MUJF5mjwGX4brogeBK1PCLgm1M_ExoCnhsQAvD_BwE:G:s&s_kwcid=AL!3652!3!394297685934!!!g!!&cid=bid_mol_pch_r01_co_cp1358_pjt0000_bid00000_0se_gaw_dy_pur_con&gclid=CjwKCAjwquWVBhBrEiwAt1Kmwq8RGSsrnEiZdPdyMZrlE4qy4MUJF5mjwGX4brogeBK1PCLgm1M_ExoCnhsQAvD_BwE

Sedlazeck, F. J., Lee, H., Darby, C. A., & Schatz, M. C. (2018). Piercing the dark matter: Bioinformatics of long-range sequencing and mapping. *Nat Rev Genet, 19*(6), 329-346. https://doi.org/10.1038/s41576-018-0003-4

Sharma, S. P., Zuo, T., & Peterson, T. (2021). Transposon-induced inversions activate gene expression in the maize pericarp. *Genetics, 218*(2). https://doi.org/10.1093/genetics/iyab062

Sætre, G. P., & Ravinet, M. (2019). *Evolutionary Genetics: Concepts, Analysis, and Practice*. Oxford University Press. https://books.google.no/books?id=XNqUDwAAQBAJ

Takara. *Advantage 2 Polymerase Mix*. Takara Bio Inc. https://www.takarabio.com/products/pcr/high-yield-pcr/advantage-2-products/advantage-2-polymerase-mix

Tamura, K., Stecher, G., & Kumar, S. (2021). MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution, 38*(7), 3022-3027. https://doi.org/10.1093/molbev/msab120

Tørresen, O. K., Star, B., Jentoft, S., Reinar, W. B., Grove, H., Miller, J. R., Walenz, B. P., Knight, J., Ekholm, J. M., Peluso, P., Edvardsen, R. B., Tooming-Klunderud, A., Skage, M., Lien, S., Jakobsen, K. S., & Nederbragt, A. J. (2017, 2017/01/18). An improved genome assembly uncovers prolific tandem repeats in Atlantic cod. *BMC Genomics, 18*(1), 95. https://doi.org/10.1186/s12864-016-3448-x

Villoutreix, R., Ayala, D., Joron, M., Gompert, Z., Feder, J. L., & Nosil, P. (2021). Inversion breakpoints and the evolution of supergenes. *Molecular Ecology, 30*(12), 2738-2755. https://doi.org/https://doi.org/10.1111/mec.15907

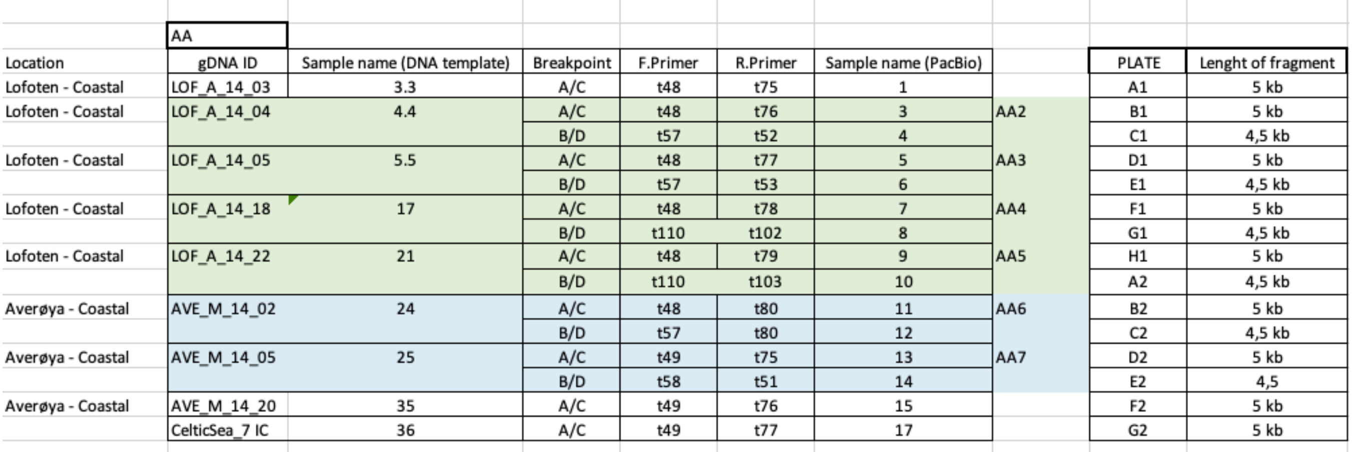
Wallace, A. G., Detweiler, D., & Schaeffer, S. W. (2013, Jul 8). Molecular population genetics of inversion breakpoint regions in Drosophila pseudoobscura. *G3 (Bethesda), 3*(7), 1151-1163. https://doi.org/10.1534/g3.113.006122

Wellenreuther, M., & Bernatchez, L. (2018, 2018/06/01/). Eco-Evolutionary Genomics of Chromosomal Inversions. *Trends in Ecology & Evolution, 33*(6), 427-440. https://doi.org/https://doi.org/10.1016/j.tree.2018.04.002

Woodruff, D. S. (2001). Populations, Species, and Conservation Genetics. In S. A. Levin (Ed.), *Encyclopedia of Biodiversity* (pp. 811-829). Elsevier. https://doi.org/https://doi.org/10.1016/B0-12-226865-2/00355-2

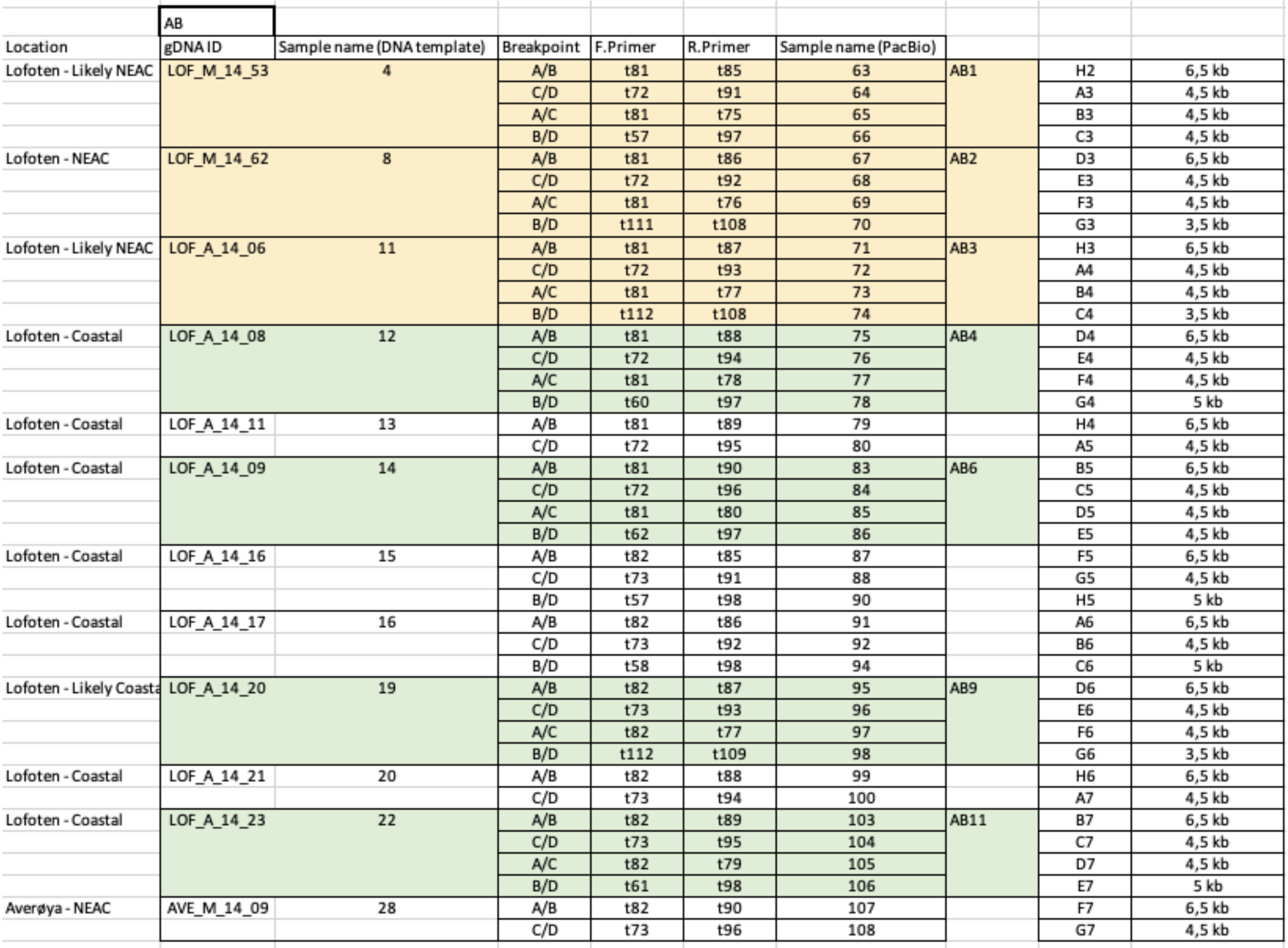
7. Appendix

Table 1A. *An overview over the individuals homozygote for the inverted allele A, with the forward and reverse barcoded primer for PacBio HiFi-sequecing.*



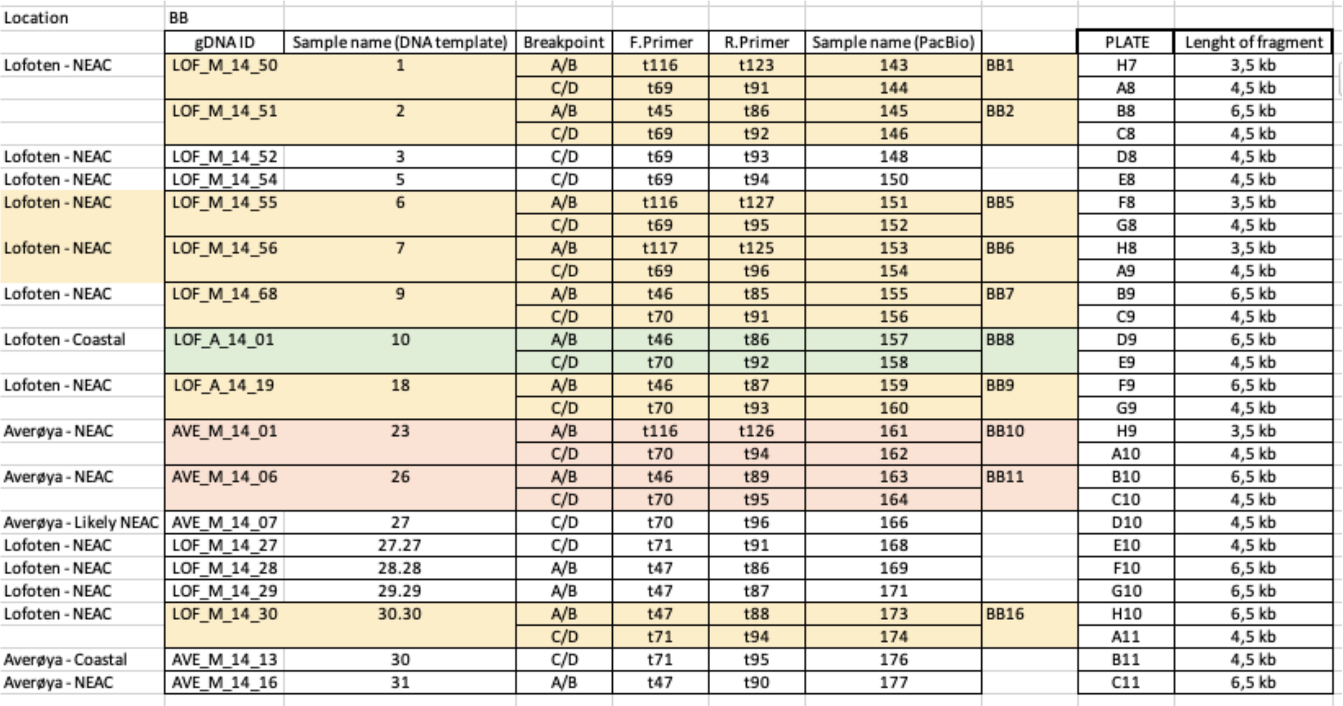
	AA								
Location	gDNA ID	Sample name (DNA template)	Breakpoint	F.Primer	R.Primer	Sample name (PacBio)		PLATE	Lenght of fragment
Lofoten - Coastal	LOF_A_14_03	3.3	A/C	t48	t75	1		A1	5 kb
Lofoten - Coastal	LOF_A_14_04	4.4	A/C	t48	t76	3	AA2	B1	5 kb
			B/D	t57	t52	4		C1	4,5 kb
Lofoten - Coastal	LOF_A_14_05	5.5	A/C	t48	t77	5	AA3	D1	5 kb
			B/D	t57	t53	6		E1	4,5 kb
Lofoten - Coastal	LOF_A_14_18	17	A/C	t48	t78	7	AA4	F1	5 kb
			B/D	t110	t102	8		G1	4,5 kb
Lofoten - Coastal	LOF_A_14_22	21	A/C	t48	t79	9	AA5	H1	5 kb
			B/D	t110	t103	10		A2	4,5 kb
Averøya - Coastal	AVE_M_14_02	24	A/C	t48	t80	11	AA6	B2	5 kb
			B/D	t57	t80	12		C2	4,5 kb
Averøya - Coastal	AVE_M_14_05	25	A/C	t49	t75	13	AA7	D2	5 kb
			B/D	t58	t51	14		E2	4,5
Averøya - Coastal	AVE_M_14_20	35	A/C	t49	t76	15		F2	5 kb
	CelticSea_7 IC	36	A/C	t49	t77	17		G2	5 kb

Table 2A. *An overview over the individual heterozygote for both the inverted and non-inverted allele with their respective forward and reverse barcoded primer for PacBio HiFi-sequencing.*



	AB								
Location	gDNA ID	Sample name (DNA template)	Breakpoint	F.Primer	R.Primer	Sample name (PacBio)			
Lofoten - Likely NEAC	LOF_M_14_53	4	A/B	t81	t85	63	AB1	H2	6,5 kb
			C/D	t72	t91	64		A3	4,5 kb
			A/C	t81	t75	65		B3	4,5 kb
			B/D	t57	t97	66		C3	4,5 kb
Lofoten - NEAC	LOF_M_14_62	8	A/B	t81	t86	67	AB2	D3	6,5 kb
			C/D	t72	t92	68		E3	4,5 kb
			A/C	t81	t76	69		F3	4,5 kb
			B/D	t111	t108	70		G3	3,5 kb
Lofoten - Likely NEAC	LOF_A_14_06	11	A/B	t81	t87	71	AB3	H3	6,5 kb
			C/D	t72	t93	72		A4	4,5 kb
			A/C	t81	t77	73		B4	4,5 kb
			B/D	t112	t108	74		C4	3,5 kb
Lofoten - Coastal	LOF_A_14_08	12	A/B	t81	t88	75	AB4	D4	6,5 kb
			C/D	t72	t94	76		E4	4,5 kb
			A/C	t81	t78	77		F4	4,5 kb
			B/D	t60	t97	78		G4	5 kb
Lofoten - Coastal	LOF_A_14_11	13	A/B	t81	t89	79		H4	6,5 kb
			C/D	t72	t95	80		A5	4,5 kb
Lofoten - Coastal	LOF_A_14_09	14	A/B	t81	t90	83	AB6	B5	6,5 kb
			C/D	t72	t96	84		C5	4,5 kb
			A/C	t81	t80	85		D5	4,5 kb
			B/D	t62	t97	86		E5	4,5 kb
Lofoten - Coastal	LOF_A_14_16	15	A/B	t82	t85	87		F5	6,5 kb
			C/D	t73	t91	88		G5	4,5 kb
			B/D	t57	t98	90		H5	5 kb
Lofoten - Coastal	LOF_A_14_17	16	A/B	t82	t86	91		A6	6,5 kb
			C/D	t73	t92	92		B6	4,5 kb
			B/D	t58	t98	94		C6	5 kb
Lofoten - Likely Coasta	LOF_A_14_20	19	A/B	t82	t87	95	AB9	D6	6,5 kb
			C/D	t73	t93	96		E6	4,5 kb
			A/C	t82	t77	97		F6	4,5 kb
			B/D	t112	t109	98		G6	3,5 kb
Lofoten - Coastal	LOF_A_14_21	20	A/B	t82	t88	99		H6	6,5 kb
			C/D	t73	t94	100		A7	4,5 kb
Lofoten - Coastal	LOF_A_14_23	22	A/B	t82	t89	103	AB11	B7	6,5 kb
			C/D	t73	t95	104		C7	4,5 kb
			A/C	t82	t79	105		D7	4,5 kb
			B/D	t61	t98	106		E7	5 kb
Averøya - NEAC	AVE_M_14_09	28	A/B	t82	t90	107		F7	6,5 kb
			C/D	t73	t96	108		G7	4,5 kb

Table 3A. *An overview over the individuals homozygote for the non-inverted allele B, with the forward and reverse barcoded primer for PacBio HiFi-sequencing.*



Location	BB								
	gDNA ID	Sample name (DNA template)	Breakpoint	F.Primer	R.Primer	Sample name (PacBio)		PLATE	Lenght of fragment
Lofoten - NEAC	LOF_M_14_50	1	A/B	t116	t123	143	BB1	H7	3,5 kb
			C/D	t69	t91	144		A8	4,5 kb
	LOF_M_14_51	2	A/B	t45	t86	145	BB2	B8	6,5 kb
			C/D	t69	t92	146		C8	4,5 kb
Lofoten - NEAC	LOF_M_14_52	3	C/D	t69	t93	148		D8	4,5 kb
Lofoten - NEAC	LOF_M_14_54	5	C/D	t69	t94	150		E8	4,5 kb
Lofoten - NEAC	LOF_M_14_55	6	A/B	t116	t127	151	BB5	F8	3,5 kb
			C/D	t69	t95	152		G8	4,5 kb
Lofoten - NEAC	LOF_M_14_56	7	A/B	t117	t125	153	BB6	H8	3,5 kb
			C/D	t69	t96	154		A9	4,5 kb
Lofoten - NEAC	LOF_M_14_68	9	A/B	t46	t85	155	BB7	B9	6,5 kb
			C/D	t70	t91	156		C9	4,5 kb
Lofoten - Coastal	LOF_A_14_01	10	A/B	t46	t86	157	BB8	D9	6,5 kb
			C/D	t70	t92	158		E9	4,5 kb
Lofoten - NEAC	LOF_A_14_19	18	A/B	t46	t87	159	BB9	F9	6,5 kb
			C/D	t70	t93	160		G9	4,5 kb
Averøya - NEAC	AVE_M_14_01	23	A/B	t116	t126	161	BB10	H9	3,5 kb
			C/D	t70	t94	162		A10	4,5 kb
Averøya - NEAC	AVE_M_14_06	26	A/B	t46	t89	163	BB11	B10	6,5 kb
			C/D	t70	t95	164		C10	4,5 kb
Averøya - Likely NEAC	AVE_M_14_07	27	C/D	t70	t96	166		D10	4,5 kb
Lofoten - NEAC	LOF_M_14_27	27.27	C/D	t71	t91	168		E10	4,5 kb
Lofoten - NEAC	LOF_M_14_28	28.28	A/B	t47	t86	169		F10	6,5 kb
Lofoten - NEAC	LOF_M_14_29	29.29	A/B	t47	t87	171		G10	6,5 kb
Lofoten - NEAC	LOF_M_14_30	30.30	A/B	t47	t88	173	BB16	H10	6,5 kb
			C/D	t71	t94	174		A11	4,5 kb
Averøya - Coastal	AVE_M_14_13	30	C/D	t71	t95	176		B11	4,5 kb
Averøya - NEAC	AVE_M_14_16	31	A/B	t47	t90	177		C11	6,5 kb

Table 4A. *The 73 Atltantic cod individuals used for DNA extraction. Lofoten (30), Averøya (28) and the celtic sea (15), all sampled for the Aqua Genome project.*

Species ID	Sample name	Population	Sampling date	Genotype	Species type
LOF_M_14_50	1	Lofoten	19.03.2014	BB	neac
LOF_M_14_51	2	Lofoten	19.03.2014	BB	neac
LOF_M_14_52	3	Lofoten	19.03.2014	BB	neac
LOF_M_14_53	4	Lofoten	19.03.2014	AB	likely_neac
LOF_M_14_54	5	Lofoten	19.03.2014	BB	neac
LOF_M_14_55	6	Lofoten	19.03.2014	BB	neac
LOF_M_14_56	7	Lofoten	19.03.2014	BB	neac
LOF_M_14_62	8	Lofoten	19.03.2014	AB	neac
LOF_M_14_68	9	Lofoten	19.03.2014	BB	neac
LOF_A_14_01	10	Lofoten	05.08.2014	BB	coastal
LOF_A_14_06	11	Lofoten	06.08.2014	AB	likely_neac
LOF_A_14_08	12	Lofoten	07.08.2014	AB	coastal
LOF_A_14_11	13	Lofoten	08.08.2014	AB	coastal
LOF_A_14_09	14	Lofoten	08.08.2014	AB	coastal
LOF_A_14_16	15	Lofoten	09.08.2014	AB	coastal
LOF_A_14_17	16	Lofoten	09.08.2014	AB	coastal
LOF_A_14_18	17	Lofoten	09.08.2014	AA	coastal
LOF_A_14_19	18	Lofoten	09.08.2014	BB	neac
LOF_A_14_20	19	Lofoten	09.08.2014	AB	likely_coastal
LOF_A_14_21	20	Lofoten	09.08.2014	AB	coastal
LOF_A_14_22	21	Lofoten	09.08.2014	AA	coastal
LOF_A_14_23	22	Lofoten	09.08.2014	AB	coastal
AVE_M_14_01	23	Averøya	24.03.2014	BB	neac
AVE_M_14_02	24	Averøya	24.03.2014	AA	coastal
AVE_M_14_05	25	Averøya	25.03.2014	AA	coastal
AVE_M_14_06	26	Averøya	25.03.2014	BB	neac
AVE_M_14_07	27	Averøya	25.03.2014	BB	likely_neac
AVE_M_14_09	28	Averøya	25.03.2014	AB	neac
AVE_M_14_10	29	Averøya	25.03.2014	AB	likely_coastal
AVE_M_14_13	30	Averøya	25.03.2014	BB	coastal
AVE_M_14_16	31	Averøya	25.03.2014	BB	neac
AVE_M_14_17	32	Averøya	25.03.2014	BB	neac
AVE_M_14_19	33	Averøya	25.03.2014	BB	neac
AVE_M_14_18	34	Averøya	25.03.2014	BB	neac
AVE_M_14_20	35	Averøya	25.03.2014	AA	coastal
CelticSea_7 IC	36			AA	
CelticSea_8 IC	37			AA	
CelticSea_9 IC	38			AA	
CelticSea_10 IC	39			AA	
CelticSea_11 IC	40			AA	
CelticSea_12 IC	41			AA	

CelticSea_16 IC	42			AA	
CelticSea_18 IC	43			AA	
CelticSea_20 IC	44			AA	
CelticSea_22 IC	45			AA	
CelticSea_30 IC	46			AA	
CelticSea_34 IC	47			AA	
CelticSea_35 IC	48			AA	
CelticSea_36 IC	49			AA	
CelticSea_38 IC	50			AA	
AVE_S_14_08	51	Averøya	15.09.2014	AA	coastal
AVE_S_14_11	52	Averøya	15.09.2014	AB	coastal
AVE_S_14_17	53	Averøya	16.09.2014	AA	coastal
AVE_S_14_19	54	Averøya	16.09.2014	AB	coastal
AVE_S_14_21	55	Averøya	16.09.2014	AA	coastal
AVE_S_14_22	56	Averøya	16.09.2014	AA	coastal
AVE_S_14_23	57	Averøya	17.09.2014	AA	coastal
AVE_S_14_24	58	Averøya	17.09.2014	AA	coastal
AVE_S_14_25	59	Averøya	17.09.2014	AB	coastal
AVE_S_14_27	60	Averøya	17.09.2014	AA	coastal
AVE_S_14_30	61	Averøya	17.09.2014	AB	coastal
AVE_S_14_33	62	Averøya	18.09.2014	AA	coastal
AVE_S_14_37	63	Averøya	18.09.2014	AB	coastal
AVE_S_14_38	64	Averøya	18.09.2014	AB	coastal
AVE_S_14_43	65	Averøya	18.09.2014	AB	coastal
LOF_A_14_03	3.3	Lofoten	06.08.2014	AA	coastal
LOF_A_14_04	4.4	Lofoten	06.08.2014	AA	coastal
LOF_A_14_05	5.5	Lofoten	06.08.2014	AA	coastal
LOF_M_14_27	27.27	Lofoten	18.03.2014	BB	neac
LOF_M_14_28	28.28	Lofoten	18.03.2014	BB	neac
LOF_M_14_29	29.29	Lofoten	18.03.2014	BB	neac
LOF_M_14_30	30.30	Lofoten	18.03.2014	BB	neac
LOF_M_14_31	31.31	Lofoten	18.03.2014	BB	neac

Table 5A. *The substitution type and the amount found when doing variant calling towards coastal reference genome*

Substitution types	*Count*
A > C	104
A > G	167
A > T	116
C > A	111
C > G	57
C > T	169
G > A	180
G < T	99
T > A	124
T > C	186
T > G	76

Table 6A. The substitution type and the amount found when doing variant calling towards *gadmor3 reference genome*

Substitution type	*Count*
A > C	59
A > G	136
A > T	86
C > A	84
C > G	44
C > T	121
G > A	118
G > C	36
G > T	77
T > A	85
T > C	130
T > G	49