Norwegian University of Life Sciences
Faculty of Biosciences
CIGENE

Philosophiae Doctor (PhD)
Thesis 2019:102

Genomics exploring population structure and sex determination in Atlantic cod *(Gadus morhua)*

Genomiske analyser av populasjonstruktur og kjønnbestemmelse hos Atlantisk torsk *(Gadus morhua)*

Graceline Tina Kirubakaran

Genomics exploring population structure and sex determination in Atlantic cod (*Gadus morhua*)

Genomiske analyser av populasjonstruktur og kjønnbestemmelse hos Atlantisk torsk (*Gadus morhua*)

Philosophiae Doctor (PhD) Thesis

Graceline Tina Kirubakaran

Norwegian University of Life Sciences
Faculty of Biosciences
CIGENE

Ås (2019)

Thesis number: 2019:102
ISSN: 1894-6402
ISBN: 978-82-575-1667-3

PhD supervisors

Prof. Sigbjørn Lien
Faculty of Biosciences,
Department of Animal and
Aquacultural Sciences, CIGENE,
Norwegian University of Life Sciences,
P.O. Box 5003 NMBU,
1432 Ås, Norway.
sigbjorn.lien@nmbu.no

Dr. Matthew Kent
Department of Animal and
Aquacultural Sciences, CIGENE,
Norwegian University of Life Sciences,
P.O. Box 5003 NMBU,
1432 Ås, Norway.
matthew.peter.kent@nmbu.no

Prof. Øivind Andersen
Senior scientist
Nofima,
Postboks 210, 1431 Ås, Norway.
oivind.andersen@nofima.no

PhD Evaluation Committee

Ass. Prof. Hendrik-Jan Megens
Wageningen University,
Wageningen, Gelderland,
The Netherlands.
hendrik-jan.megens@wur.nl

Dr. Erica Helen Leder
Senior scientist
Nofima,
Postboks 210, 1431 Ås, Norway.
erica.Leder@Nofima.no

Prof. Hans Magnus Gjøen
Norwegian University of Life Sciences,
P.O. Box 5003 NMBU,
1432 Ås, Norway.
hans.magnus.gjoen@nmbu.no

Acknowledgments

The thesis presented here is based on work performed at the Department of Biosciences (BIOVIT) at the Norwegian University of Life Sciences (NMBU). The project was funded by NMBU.

First and foremost, I would like to express my gratitude to my supervisors Prof. Sigbjørn Lien, Dr. Matthew Kent and Prof. Øivind Andersen, your scientific guidance, critical suggestions and valuable input during discussion sessions greatly improved my work. I appreciate all your contributions and ideas that made my PhD experience productive and stimulating. Sigbjørn, Thank you for taking me in. Matthew, I don't think I would have managed without your incredible support, thank you. My sincere thanks to all the co-authors, especially Torfinn for bioinformatics support, Terese Anderstuen, Kristina Hallan for the hours spent on PCR and Mariann Árnyasi for generating the long nanopore sequences.

Further, I would like to acknowledge all my colleagues, mainly the cheerful lab-ladies for creating a friendly environment. Thank you, Volha Paulouskaya for the good moments and friendship we shared this past few months. Thank you Mallikarjuna Rao and Peng Lei for the immense support and friendship. I also want to thank all the members of our church, especially Mats Jacobsson and Elias for the spiritual and friendly encouragement.

Finally, I want to thank God for blessing me with an amazing family. I fall a short of words to express my deep sense of gratitude to my dad and mom for their love, prayers and unending support. To my dear husband Jeevan, thank you for your encouragement, support and for always staying positive and cheerful in all circumstances. I dedicate this PhD thesis to my two lovely children, Nethan and Ivan, who are the pride and joy of my life. You have made me stronger, better, more fulfilled than I could have ever imagined. You are my greatest inspiration, my motivation and energy.

Ås, November 2019
Tina.

Table of Contents

List of papers

Paper I

Tina Graceline Kirubakaran, Harald Grove, Matthew P. Kent, Simen R. Sandve, Matthew Baranski, Torfinn Nome, Maria Cristina De Rosa, Benedetta Righino, Torild Johansen, Håkon Otterå, Anna Sonesson, Sigbjørn Lien, Øivind Andersen. **Two adjacent inversions maintain genomic differentiation between migratory and stationary ecotypes of Atlantic cod.** *Molecular Ecology.* 25, 2130–2143 (2016).

Paper II

Tina Graceline Kirubakaran, Øivind Andersen, Michel Moser, Mariann Arnyasi, Philip McGinnity, Sigbjørn Lien*, Matthew Kent*. **A nanopore based chromosome-level assembly representing Atlantic cod from the Celtic Sea.** *Manuscript*

Paper III

Kirubakaran TG, Andersen Ø, De Rosa MC, Andersstuen T, Hallan K, Kent MP, Lien S. **Characterization of a male specific region containing a candidate sex determining gene in Atlantic cod.** *Scientific Reports*. 2018; 9:116.

Summary

Atlantic cod (*Gadus morhua*) is a benthopelagic cold-water marine species found across the North Atlantic Ocean. Thanks to its historical and current significance as an important food resource it is one of the most well-known and well-studied teleost fish species. Genetic studies of Atlantic cod have been conducted since the 1930s, but with the advent of high-throughput DNA sequencing technologies investigations have been reinvigorated with high-resolution data allowing us to address previously unanswerable questions. This thesis explores the genome biology of Atlantic cod with the specific goals of; i) understanding the genomics underlying population divergence between cod populations in Norwegian waters, ii) using long-read sequencing technology to build a genome assembly representing the southern cod populations, and iii) characterizing genomic regions associated with sex determination. The knowledge generated in this thesis can enhance the understanding on how genomic architecture influence genome-wide variations contributing to population structures and sex determination.

Atlantic cod in the Norwegian Sea are classified into one of two different populations; migratory Northeast Arctic cod (NEAC) and stationary Norwegian coastal cod (NCC). In Paper I, we sought to understand how phenotypic and genetic differences are maintained despite interbreeding between NEAC and NCC. Utilizing genotype data from 192 parents of farmed families of NEAC, NCC or NEACxNCC crosses, we identified extended linkage disequilibrium (LD) in a 17.4Mb region on linkage group 1 (LG01). Furthermore, linkage analysis revealed two adjacent inversions within the region that repress meiotic recombination in NEACxNCC crosses. The haplotype block harbours 763 genes, including candidates regulating swim bladder pressure, heme synthesis and skeletal muscle organization conferring NEAC adaptation to long-distance migration and vertical movements to large depths. Our results document that inversion is the genetic mechanism that maintains the genetic differentiation despite interbreeding and we hypothesize the co-occurrence of possibly multiple adaptive genes forming a ‘supergene’ advantageous to NEAC.

The public reference genomes for Atlantic cod have all been derived from NEAC samples and therefore representing the northernmost cod population, adapted to near freezing temperatures. Several studies have demonstrated regions of genomic differentiation on linkage groups (LGs) 1, 2, 7 and 12 associated with adaptation to temperature along the

north-south gradient (Bradbury et al. 2010; 2013). In paper II, we generated a highly contiguous genome assembly representing the southern Celtic population of Atlantic cod using long-read nanopore sequencing data. By comparing this to the latest NEAC assembly gadMor3 we were able to characterize in detail the rearrangements creating the 'islands of genomic divergence' on LGs 1, 2, 7, and 12. The long contiguous genome assembly also facilitated the identification of a putative centromere-specific repeat.

In paper III, by comparing and contrasting whole genome short-read sequencing data from 49 male and 53 female cod, we detected a male specific region of 9,149 bp on LG11. A diagnostic PCR test was developed and confirmed the sex-specific nature of this male specific sequence in phenotypically sexed Atlantic cod, but also in closely related gadoids. A single gene named *zkY* (zinc knuckle on the Y chromosome), encoding a zinc knuckle protein, was detected within this male specific region and presented as a novel candidate sex-determining gene in Atlantic cod. We identified eight highly similar autosomal gene copies of *zkY* containing a zinc knuckle motif. The 3D modelling of the zinc knuckle domain of these suggests that amino acid changes might influence putative RNA binding specificity in six of the copies, whereas autosomal copies zk1 and zk2 possess identical zinc knuckle structure. Furthermore, gene expression data from early developmental stages provides additional evidence for function of *zkY* as a sex determining gene compared to autosomal copies. Collectively our results suggest that the *zkY* gene is involved in Atlantic cod sex determination.

Sammendrag

Atlantisk torsk (*Gadus morhua*) er en marin kaldtvannsart med bred utbredelse i Nordatlantiske farvann. Takket være den historiske og nåværende betydningen av denne viktige matressursen, er den blitt en av de mest kjente og undersøkte beinfisker. Genetiske studier av Atlantisk torsk har pågått siden 1930-tallet, men nyere storskala sekvenseringsteknologi har styrket undersøkelsene med høyoppløselige data som gjør det mulig å besvare tidligere ubesvarte spørsmål. Denne avhandlingen har undersøkt genombiologien hos Atlantisk torsk med spesifikk målsetting om å i) forståelse den genomiske basisen for utviklingen av torskepopulasjoner i norske farvann, ii) bruke ny sekvenseringsteknologi til å lage en referansegenomsekvens for som representerer sydlige torskepopulasjoner, iii) karakterisere genomiske områder i torskegenomet som er avgjørende for kjønnsbestemmelse. Kunnskapen fra denne avhandlingen vil forbedre forståelsen av hvordan oppbygningen av genomet påvirker den genomiske variasjonen og bidrar til populasjonsstruktur og kjønnsbestemmelse.

Atlantisk torsk i Norskehavet er klassifisert i to populasjoner; skrei (NEAC) og kysttorsk (NCC). I artikkel I ønsket vi å forstå hvordan fenotypisk og genetiske forskjeller opprettholdes til tross for krysning mellom kysttorsk og skrei. Med bruk av genotypedata fra 192 foreldre av ren kysttorsk, skrei og krysniner mellom disse, identifiserte vi utbredt koblingsulikevekt (LD) i en 17,4 Mb region på koblingsgruppe 1 (LG01). I det samme området avdekket vi to inversjoner som hindrer rekombinasjon hos krysninger mellom kysttorsk og skrei. Området inneholder 763 gener, blant annet kandidater som regulerer gasstrykket i svømmeblæren, hem-syntesen og organiseringen av skjelettmuskulaturn, noe som sannsynlig gjør skreien tilpasset vandringer over lange distanser og dykking ned til relativt store dyp. Våre resultater dokumenterer at inversjonen på LG01 er den genetiske mekanismen som opprettholder den genetiske forsskjeller til tross for kontakt mellom populasjoner over tid. Vi foreslår at organiseringen av mange adaptive gener danner et «supergen» som er fordelaktig for skrei.

Offentlig tilgjengelige referansegenomer hos Atlantisk torsk er basert på skrei som representerer en nordlig torskepopulasjon tilpasset temperaturer nær frysepunktet. Flere studier har vist at genomisk differensiering i regioner på koblingsgruppe 1, 2, 7 og 12, også kalt supergener,er assosiert med temperaturtilpasning langs en nord-syd-gradient (Berg et al. 2016; Bradbury et al. 2014; Bradbury et al. 2013; Sodeland et al. 2016). I artikkel II laget vi

en genomsekvens for en sydlig torskepopulasjon (Keltisk torsk) ved bruk av nanoporesekvensering. Ved å sammenlikne det nyeste skrei genomet «gadMor3» kunne vi avgrense supergenene på koblingsgruppe 1, 2, 7 og 12. Vi identifiserte også et karakteristisk repetert DNA-element som er knyttet til sentromerer i torskegenomet. Disse sekvensen ble brukt sammen med genetiske kart til å identifisere fire metasentriske kromosomer hos Atlantisk torsk.

I artikkel lII identifiserte vi en hann-spesifikk region på 9.149 bp på koblingsgruppe11 (LG11) ved å sammenlikne helgenom sekvensdata fra 49 hanner og 53 hunner av torsk. En diagnostisk PCR test ble utviklet og bekreftet kjønnsspesifisiteten av hann-sekvensen hos fenotypisk kjønnsbestemt Atlantisk torsk, men også hos nært relaterte torskefisk. Ett enkelt gen kalt zkY (zinc knuckle on the Y chromosome), som koder for et zinc knuckle protein, ble funnet i den hann-spesifikke regionen og representerer et nytt kjønnsestemmende gen hos Atlantisk torsk. Vi fant åtte svært like genkopier av zkY på andre torskekromosomer som alle inneholder zinc knuckle motivet. 3D modellering av zinc knuckle domenet hos disse proteinene indikerer at aminosyreforandringer i seks av kopiene kan påvirke spesifisiteten av den antatte RNA-bindingen. Kun de autosomale kopiene zk1 og zk2 har identiske zinc knuckle struktursom zkY, men genekspresjonsdata fra tidlige utviklingsstadier bidrar ytterligere til å underbygge funksjonen til zkY som det kjønnsbestemmende genet hos Atlantisk torsk. Samlet sett viser resultater at zkY-genet sannsynligvis er bestemmende for utvikling som hann eller hunn hos Atlantisk torsk.

1. Introduction

Atlantic cod is a cold-water marine fish widely distributed across the continental shelves of the Northwest and Northeast Atlantic Ocean. The abundance of cod in these waters has made it one of the most commercially important fish species in the North Atlantic. Norway has the world's largest cod stocks, and the abundance of this species has had a profound influence on the rise of settlements along its coastline, with the Lofoten archipelago especially influenced by the seasonal arrival of Atlantic cod from the Barents Sea for spawning between January and April. The arrival of the spawning fishes coincides with climatic condition that is ideal for freeze-drying, a technique where cod is left to dry by cold air and strong winds on large open-air timber racks. The freeze-dried cod also called stockfish can last for a long time without losing its flavour. During the middle Ages, the stockfish provided the Vikings with much-needed sustenance during their epic journeys and enabled them to establish international stockfish trade with communities around Baltic and Northern Europe (Barrett et al. 2011; Star et al. 2017). This iconic exchange since the Viking era shows that cod is popular on dinner tables today as it was centuries ago and is considered an important part of Norway's cultural heritage.

In the Northwest Atlantic, cod is found throughout the coastal inshore waters of Atlantic Canada. In fact, it was the abundance of cod in Grand Banks of Newfoundland that gave rise to the first settlements of modern European explorers in British North America in the late fifteenth century (Hubbard 2009). Cod fishery contributed significantly to the economy of the region and was nicknamed 'Newfoundland currency'. For centuries fishing was carried out as a subsistence activity with the fishing technology used at the time limiting the volume of the catch. As demand increased however, intensive fishing practices emerged. The introduction of modern fishing boats equipped with radar and sonar in late 1950s allowed fishermen to trawl larger areas and fish in deeper depths. This increased efficiency made cod stocks deplete at a rate faster than they could be replenished, which led to the decline of the entire cod biomass to critically low levels by 1995. The total catch of Atlantic cod in the 1970s was about 3.5 million tonnes in North Atlantic but now the total catch has declined to less than a million tonnes (Johansen et al. 2009), with overfishing being the major cause of this decline. The collapse of the Atlantic Canadian cod fishery, its economic, social, and cultural implications is one of the most commonly cited examples in the world of overfishing (Barrett et al. 2004). However currently, other anthropogenic effects like accelerated climate change, pollution, and habitat loss have also shown to trigger

declination of marine populations in general. With increasing human population and decline in capture fishery, a shift towards the production of Atlantic cod using aquaculture aims to bridge this gap. But one of the major problems for cod aquaculture is early sexual maturation, as the fish invest their energy in producing gonads and this slows or arrests growth and reduces flesh quality. To sustain future Atlantic cod fisheries, knowledge about the genes involved in sex determination (SD) may play an important role in production strategies. Acquiring this knowledge of candidate SD gene was the primary objective of paper III.

In general, marine species spread through a long geographic range and are generally assessed as large undifferentiated population with high gene flow. Atlantic cod is one such species with a large effective population size, occupying diverse habitats. With the advances in genomic technologies tools, studies within marine species are also revealing genetic differences that influence adaptation to local environment (Jones et al. 2012; Wang et al. 2013; Berg et al. 2015; Guo et al. 2015). In this thesis in paper I and paper II, we not only aim to identify significant genetic differences between cod populations, but also unravel the genetic mechanism that creates genetic barriers to evolve between populations despite high gene flow.

2. Atlantic cod population structures across the North Atlantic Ocean

The contemporary distribution of Atlantic cod (Figure 1) has been shaped by both current and historical patterns in ocean climate and habitat availability. One obvious prerequisite for genetically classifying cod populations is that there is sufficient genetic structure among populations. Many population genetics studies in Atlantic cod, some dating back to the 1930s, have reported population structure and generally disregarded the presumption of a single homogenous population (Rollefsen 1935; Moller 1966). The abundance of cod in the North Atlantic varies greatly among the different areas and is divided into several more or less separate stocks with different population sizes and harvest regimes.

In the Northeast Atlantic Ocean, cod habitats extend from the Barents Sea and White Sea in the North to the Celtic/Irish Sea in the south. In the Barents Sea, Atlantic cod is classified into two different populations: migratory Northeast Arctic cod (NEAC) or Barents Sea cod and stationary Norwegian coastal cod (NCC). In the Baltic Sea, cod inhabits brackish waters

tolerating varying levels of salinity from nearly freshwater in the northern Baltic to around 30% at the border to the North Sea (Berg et al. 2015).

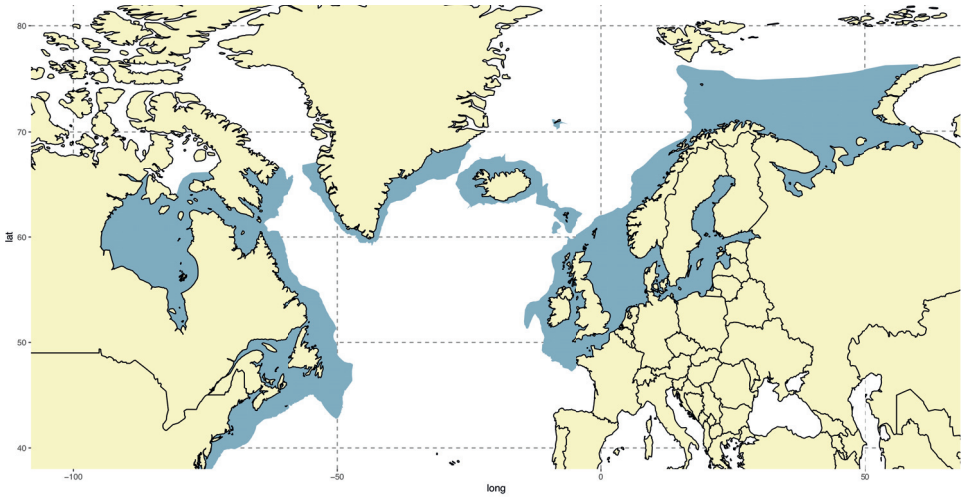

Figure 1: Atlantic cod distribution (blue areas). The distribution range was obtained from FAO Catalogues of Species (http://www.fao.org/geonetwork).

In Iceland, tagging studies have shown two distinct cod populations, resident coastal cod and migratory frontal cod (Pampoulie et al. 2006; Pampoulie et al. 2008b). Coastal cod reside mostly at depths less than 200 meters, whereas frontal cod are generally found at depths greater than 200 meters outside the spawning season and tend to conduct frequent vertical migrations. Atlantic cod in the North Sea is assessed as a single stock, although there is evidence of genetic structuring of coastal and locally adapted populations between the offshore and inshore northern and southern North Sea (Barth et al. 2017). Around the British Isles, cod is managed as the Celtic Sea, the Irish Sea and West of Scotland populations (Neat et al. 2014).

In the Northwest Atlantic, cod habitats range from Greenland to North Carolina in USA. Tagging and oceanographic studies on cod stocks in the Gulf of Maine indicate that Northern stocks undertake extensive annual migrations, spending winter at deep waters off the southern coast of Newfoundland and move towards the spawning grounds on the west coast of Newfoundland in late spring (Howell et al. 2008). In contrast, cod stocks in southern region display low levels of movement (Howell et al. 2008). A genome-wide

survey of genetic polymorphism has partitioned cod stocks as migratory northern (Can-N) and stationary southern (Can-S) population (Bradbury et al. 2013).

Sea water temperature is a key differentiator from the north to the south in the Atlantic Ocean. Given its wide distribution Atlantic cod is adapted to a wide range of temperatures from nearly freezing to 20°C and distributed from the shoreline down to the continental shelf with optimal temperature between 7°C and 15°C (Righton et al. 2010). The optimal temperature is generally accessed by growth, fecundity, etc. A study on four cod populations (Barents Sea, Icelandic waters, North Sea and Irish Sea) in the north eastern part of Atlantic showed extreme variability in size and age of sexual maturity, growth and fecundity with clear gradients from the north to the south (Thorsen et al. 2010). Cod in the cold waters have low growth rates and fecundity and mature later in life, whereas cod in warmer environment have much higher growth rates and fecundity mature earlier in life (Thorsen et al. 2010). Adult cod could occupy most temperatures during feeding season but the range of temperature at which spawning take place is narrower and peaks at 7°C (Righton et al. 2010). Studies have reported fluctuations in cod population linking to changes in Barents Sea temperature (Arthun et al. 2012; Dalpadado et al. 2012; Smedsrud et al. 2013; Hollowed et al. 2014; Fossheim et al. 2015).

3. Atlantic cod population structure in the Norwegian Sea

In Norwegian waters, Atlantic cod exhibits population structuring on both large and small spatial scales (Westgaard et al. 2007; Sundby et al. 2008). The first separation of these populations was based on differences in the shape of the otoliths and classified Atlantic cod into two groups, the NEAC and NCC (Rollefsen 1935). NEAC cod is the world's largest cod population, it spends most of the year in the Barents Sea but migrates seasonally to the Norwegian coast for spawning (Ottersen et al. 1998). In contrast, NCC populations are typically stationary living in coastal areas and fjords and experience different temperatures than that experienced by NEAC (Jakobsen 1987; Godo et al. 2000). Although the fundamental differences in migratory and non-migratory behaviour were recognized, controversy about whether the two populations are genetically different existed for several decades.

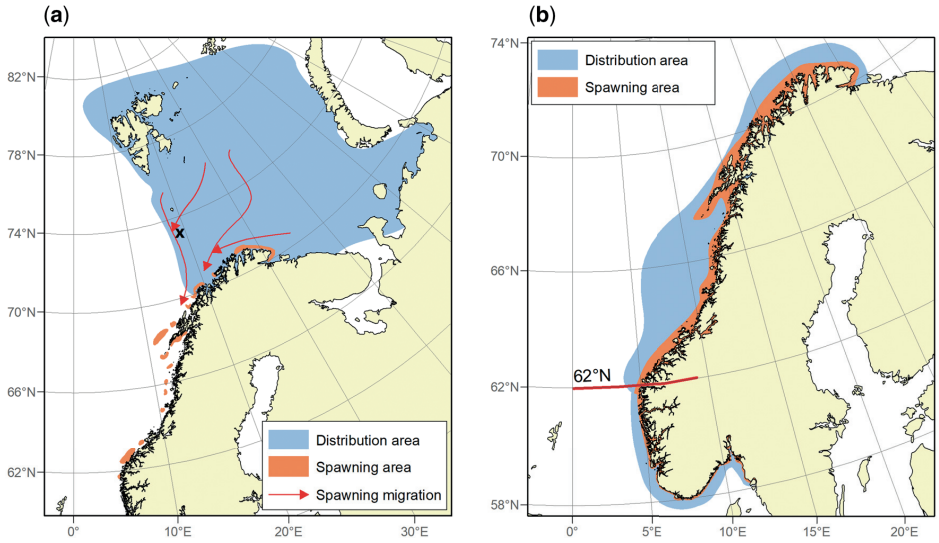

Figure 2. Distribution and spawning area of (a) NEAC and (b) NCC (Johansen et al. 2018).

Apart from migratory differences, the NEAC and NCC also display different life-history characteristics, for example when compared to NEAC, NCC grows faster as larvae, juveniles and adult fish (Vandermeeren et al. 1994; Otterlei et al. 1999). NCC matures at 5-6 years of age (length 40-50 cm; (Berg et al. 2001)), while NEAC generally mature at 6-8 years of age (length 75-90 cm (Bergstad et al. 1987)). They also differ in settling regimes, with NEAC juveniles most often settling in deep water compared to the shallow water preferred by NCC juveniles (Fevolden et al. 2012). With regards body shape, experienced fishermen can distinguish NEAC from NCC with the former generally being longer and thinner (Svasand et al. 1996). Once they mature, the eggs are spawned during March and April along the Norwegian coast, mainly around Lofoten archipelago (Bergstad et al. 1987) where spawning areas of the two cod populations overlap (Figure 2) (Jakobsen 1987; Berg et al. 2003; Johansen et al. 2018). Very little is known about the movements of young cod in their early years on the nursery grounds.

4. Early genetic studies of Atlantic cod populations

Microsatellites or short tandem repeats (STRs) have been the genetic markers of choice for more than two decades for studying population structures in Atlantic cod. For some studies, microsatellites can provide the necessary information at costs that are not prohibitive (Arthofer et al. 2018), for example for cod management and population assignment (Miller

et al. 2000; Delghandi et al. 2003; Jakobsdottir et al. 2006). Studies using microsatellites have also revealed genetic variation between populations and sub-populations that are spread across the Atlantic Ocean. Using this technology, genetic differences are found between cod sampled in different ecosystems such as, North Sea and Baltic Sea (Nielsen et al. 2001), and between individuals sampled at 22 spawning locations in Iceland coasts (Pampoulie et al. 2006). Even on small spatial scales, such as among neighbouring fjords in Norway, microsatellite markers have been resourceful in identifying genetic differences (Knutsen et al. 2003). A recent study using six microsatellite markers and pantophysin locus *(Pan 1)* on more than 4000 NCC fish sampled from 55 spawning sites displayed significant population genetic structure throughout its entire range, following a trend of isolation by distance (Dahle et al. 2018).

Over time, single nucleotide polymorphisms (SNPs) have challenged microsatellites as the most popular genetic marker type mainly due to their abundance in the genome and ease of genotyping and analysis, and several SNP based studies have been performed in Atlantic cod. Initial studies used a few thousand SNPs to study population divergence and revealed genetic differences between populations in the eastern and the western side of the Atlantic (O'Leary et al. 2007; Bradbury et al. 2010; 2013). Later, DNA from multiple large families was analysed using SNPs to construct genetic linkage maps describing the recombination between SNPs and assigning each marker a relative genetic position in LGs. The first linkage map of Atlantic cod was generated based on recombination frequency data from 12 full-sib families of Norwegian origin containing 174 SNPs and 33 microsatellites and defined 25 LGs (Moen et al. 2009). Shortly after, another higher density study used 924 SNPs detected within expressed sequence tags (EST) to define the 23 LGs (Hubert et al. 2010) reflecting the number of chromosomes identified by karyotyping and cytogenetic mapping (Ghigliotti et al. 2012). However, all these studies focus only on a limited number of genetic markers and relatively limited information about their association with genes, thus limiting our ability to study genome-wide patterns of divergence.

5. Genetic differences between the NEAC and NCC cod populations

Since the first studies on otolith patterns in 1935 (Rollefsen 1935) physical characteristics like body growth and sexual maturation, along with DNA polymorphisms, have been used in the cod populations in northern Norway to study its population structure. Some of the

first population studies used differences found in hemoglobin (*HbI*) to indicate heterogeneity among cod stocks (Frydenberg et al. 1965; Moller 1966, 1968). Later, it has been documented that the "cold" variant *HbI-2/2* is found in both NEAC and NCC in the north, while the "warm" variant *HbI-1/1* dominated the southern NCC populations (Andersen 2012). The *HbI-1/2* heterozygote seems to display intermediate properties of the two homozygotes (Andersen 2012). The microsatellite loci GMO34 and GMO132 also showed large genetic differences when comparing cod populations along the north to south gradient (Nielsen et al. 2006; Westgaard et al. 2007; Dahle et al. 2018), whereas analysis of mitochondrial genomes revealed no reproductive isolation between NEAC and NCC (Arnason et al. 1996). Variation within the *Pan I* locus (Fevolden & Pogson 1997) has been used to differentiate NEAC and NCC in multiple population genetic studies in Atlantic cod (Sarvas et al. 2005; Hemmer-Hansen et al. 2013; Andersen et al. 2015) but the physiological function of the protein remains unknown. The *Pan* I^A allele dominates in NCC population (*Pan* I^A allele frequency 0.81) and *Pan* I^B allele dominates in the NEAC population (*Pan* I^B allele frequency 0.90) (Fevolden et al. 1997; Pogson 2001). The Pan I^{AB} heterozygotes exhibit intermediate behaviors in Icelandic cod (Pampoulie et al. 2008a; Jakobsdottir et al. 2011).

6. Evolution of genomic resources and tools for Atlantic cod

The above-mentioned genetic markers provided researchers with evidence of genetic differentiation, but without a high-quality reference genome sequence it was not possible to accurately determine the location of these markers or to put them into context with other elements (e.g. genes) within the genome. Over the last 15 years, high-throughput sequencing has changed this picture dramatically where whole genome sequences are produced at an escalating scale (Zhang et al. 2011). The initial high-throughput sequencing technologies used different technical and biochemical strategies (e.g. 454 pyrosequencing vs Solid sequencing by ligation vs. Illumina's single base synthesis), but they were fundamentally alike in producing only relatively short reads. This fact has forced the bioinformatics community to develop novel approaches enabling researchers to assemble and interpret these tiny fragments of sequence into more continuous sequences to advances in genomic research. Utilizing the technology available at that time, the first *de novo* Atlantic cod genome sequence gadMor1 was generated based on 454-pyrosequencing technology data (Star et al. 2011). Although the estimated size of the assembly was 832 Mb,

the assembly was fragmented with 224 million base pairs in gaps with a contig N50 (the lengths at which half the assembly consists of sequences of those lengths or longer) of only 2.3kb. Further analyses revealed short tandem repeats located at the contig terminal indicated the repeats contributed to the observed level of fragmentation.

The availability of the first genome assembly for Atlantic cod (gadMor1), facilitated efficient detection of genetic variation in the genome, most commonly Single Nucleotide Polymorphisms (SNPs), and opened up for innovations in population genomics work in Atlantic cod. To generate a SNP set with wide utility, seven cod individuals from a wide geographic range across the Northeast Atlantic were paired-end sequenced using Illumina platform. The generated reads were aligned to gadMor1 and subsequent bioinformatics analysis detected approximately 2.8 million putative SNPs. One of the most cost-effective and reproducible ways to genotype a set of SNPs is by using oligonucleotide microarray technologies (SNP-arrays), which enable the genotyping of many thousands of selected SNPs. A set of 10,913 SNPs detected through resequencing were included on a custom Illumina cod SNP-array; '12K SNP-array'. A large family material of 2739 individuals was genotyped using this to construct linkage maps and it has been widely applied in population genomics to study divergence in Atlantic cod populations (Berg et al. 2015; Sodeland et al. 2016; Barth et al. 2017; Berg et al. 2017; Sinclair-Waters et al. 2017; Knutsen et al. 2018; Kess et al. 2019).

Recently, third-generation single molecule sequencing technologies have been developed, e.g. Pacific Biosciences (PacBio) and Oxford Nanopore Technology (ONT), which are able to deliver single contiguous reads that are tens of thousands of base pairs in length. Their technological principles upon which they are based are very different from any short-read technology and they directly sequence each unique target DNA or RNA molecule rather than an amplified population of target clones. By combining long-read sequencing reads from PacBio with short-read data from Illumina, mid-length read data from 454 and Sanger sequencing of bacterial artificial chromosome (BAC) ends, an improved genome assembly gadMor2 was generated. This resulted in a genome assembly with contig N50 of 116 Kb, genome size of 643 Mb and 1.69% gaps (Torresen et al. 2017). The inclusion of PacBio reads combined with the use of multiple assembly programs improved genome assembly statistics and confirmed the high content of tandem repeats in the Atlantic cod genome. The

gadMor2 scaffolds are ordered and oriented into LGs by matching the order of SNPs in the linkage map (Torresen et al. 2017).

While the ultimate goal would be to completely sequence a chromosome from end-to-end in a single read, this is beyond any technology currently available. However, any type of information that indicates the relative positions of genomic segments (e.g. linked reads; 10x Genomics, optical maps; BioNano genomics, and chromosome conformation capture techniques; HiC), can be used for genome scaffolding (Low et al. 2019; Pettersson et al. 2019). Very recently, a high-quality chromosome-scaled genome assembly gadMor3 was generated using PacBio and Illumina reads (NCBI accession ID: GCF_902167405.1). The initially assembly contained 1442 contigs (contig N50 = 1 Mb) that were put together into 227 scaffolds organised into 23 chromosomes amounting a total sequence length of 669 Mb with 3.5% gaps using 10x Genomics, Hi-C, BioNano genomics and a linkage map.

7. Chromosomal rearrangements and islands of genomic divergence

Larger chromosomal rearrangements like inversions may completely repress recombination between 'ancestral' and 'rearranged' homologous segments and preserve functional variation in 'island of genomic divergence' or 'supergenes' (Kirkpatrick et al. 2006; Nosil et al. 2009; Feder et al. 2012; Nosil et al. 2012). Evidence is accumulating regarding the importance of such rearrangements for phenotypic differences and among individuals and local adaptation (Hoffmann et al. 2008; Nosil et al. 2009; Kirkpatrick 2010). For example, inversions supressing recombination and creating supergenes that dictates colourful wing patterns have been demonstrated in *Heliconius* butterflies (Joron et al. 2006; Joron et al. 2011). In fruit fly (*Drosophila melanogaster)*, a large inversion (~8 Mb) on the right arm of the third chromosome (3R) is associated with climatic adaptation (Rane et al. 2015). In monkey flower (*Mimulus ringens*) a large inversion has been shown to differentiate between annual and perennial forms (Twyford et al. 2015). In three-spined stickleback (*Gasterosteus aculeatus*), inversions on three chromosomes are associated with differentiating marine and freshwater stickleback (Jones et al. 2012). In rainbow trout (*Oncorhynchus mykiss)* a double inversion on chromosome 5 mediates sex-specific migratory tendency differing anadromous (steelhead) and resident (rainbow trout) populations (Pearse et al. 2018). In Atlantic herring (*Clupea harengus)* a large inversion on chromosome 12 is associated with ecological

adaptation *(Pettersson et al. 2019).* The above-mentioned studies suggest that chromosomal inversions may be important mechanisms maintaining locally adapted genomic regions.

In Atlantic cod, large islands of genomic divergence, have been identified on LGs 1, 2, 7, 12 and their role in promoting differences between populations has been investigated in a number of studies (Bradbury et al. 2010; Bradbury et al. 2013; Hemmer-Hansen et al. 2013; Karlsen et al. 2013; Bradbury et al. 2014; Berg et al. 2015; Berg et al. 2016; Kirubakaran et al. 2016; Sodeland et al. 2016; Barney et al. 2017; Barth et al. 2017; Berg et al. 2017; Barth et al. 2019; Clucas et al. 2019a; Clucas et al. 2019b; Kess et al. 2019; Puncher et al. 2019). The region on LG01 coincides with a double inversion and has been associated with strong genetic differentiation between migratory and stationary ecotypes on both sides of the Atlantic Ocean (Kirubakaran et al. 2016; Sinclair-Waters et al. 2017). The regions on LGs 2, 7 and 12 have been associated with ocean temperatures on both sides of the Atlantic Ocean (Bradbury et al. 2010), as well as salinity and oxygen concentrations in Baltic Sea (Berg et al. 2015). The supergenes on LGs 2 and 12 have also been linked to differences in coastal (Skagerrak) and oceanic environments (North Sea) in the south of Norway (Sodeland et al. 2016). Collectively these results suggest that on a genome-wide scale a few potentially large regions or genomic islands are under selection in Atlantic cod populations.

8. Sex determination in fish species

The major role of sex determination (SD) gene is to initially determine the sex by triggering testicular or ovarian differentiation of a sexually bi-potential gonad. The genes and pathways that control SD in teleosts are not conserved and not consistently determined by any single genetic cascade. In teleosts, SD can be either genetically determined or influenced by environment or both. Among the almost 26,000 highly diverse species of teleost fish, the mechanism(s) responsible for SD has only been investigated in a handful of fish species revealing diverse SD mechanisms (Volff et al. 2001; Devlin et al. 2002; Yoshida et al. 2011). Some examples of sex determining genes are *dmY, dmrt1Yb, gsdfY* and *sox3Y* in the medaka genus (Matsuda et al. 2002; Nanda et al. 2002; Myosho et al. 2012; Takehana et al. 2014), *amhr2* in fugu (Kamiya et al. 2012), *amhy* in Patagonian pejerry (Hattori et al. 2012) and Nile Tilapia (Li et al. 2015), and *sdY* in salmonid species (Yano et al. 2012). Some of the above-mentioned genes involved in SD mechanism show striking diversity and suggest the rapid evolution of SD mechanism in teleosts. For instance, in the

genus medaka (*Oryzias*) closely related sister species have recruited different master SD genes (Tanaka et al. 2007). The sister species *O. latipes*, *O. curvinotus*, and *O. luzonensis* possess an XX–XY sex determination system. In *O. latipes*, *O. curvinotus* the SD genes *dmY, dmrt1Yb* respectively are found in the male Y chromosome (Matsuda et al. 2002; Nanda et al. 2002). But no *dmY/ dmrt1Yb* gene is found in *O. luzonensis*, instead *gsdfY* (gonadal soma derived growth factor on the Y chromosome) acts as the SD gene (Myosho et al. 2012). Later in another sister species *O. dancena*, *sox3Y* (which is orthologous to mammalian SD gene *srY*) acts as SD gene (Takehana et al. 2014).

9. Atlantic cod sex determination

Several papers provide evidence for genetic sex-determination in Atlantic cod rather than environmental (Volff et al. 2007; Avise et al. 2009; Heule et al. 2014), with the expression of several genes like *amh, cyp19a1, dmrt1, dmrt2, sox9* shown to differ between males and females (Johnsen et al. 2013). The possibility to create all-female population by breeding with masculinized females is not only evidence for genetic sex determination, but also provides evidence for male-heterogametic sex-determination, or XX-XY system (Haugen et al. 2011). Morphologically, Atlantic cod exhibit variable levels of sexual dimorphism where females become longer and heavier than males (Keyl et al. 2015). Sex is also known to influence the time of puberty, where males mature earlier than females in aquaculture (Haugen et al. 2011).

Recently, a whole-genome sequence of 227 wild-caught specimens was used to identify 166 sex-associated variants in six distinct regions on five LGs in Atlantic cod (Star et al. 2016). The largest region was an approximately 55Kb region on LG11 that contained the majority of genotypes that segregate closely to XX-XY system (Star et al. 2016), however, the study did not report candidate sex-determining gene(s). Identifying the SD gene in Atlantic cod can help address the issue with early sexual maturation in farmed cod males and can be of significance to other codfishes as well.

Aim of the thesis

To explore the genomic architecture of Atlantic cod genome with emphasis on genome-wide variations contributing to sex determination and population structures. Specific topics are to:

i) Resolve the mechanism causing the underlying differences between stationary NCC and migratory NEAC populations (Paper I)

ii) Develop a high-quality reference genome sequence from an Atlantic cod captured in south-eastern extreme of the species distribution and representing an alternative to current genome assemblies produced from the NEAC population feeding in the cold Barents Sea. (Paper II)

iii) Provide insight into the sex determining region and candidate sex determining gene in Atlantic cod (Paper III)

Brief summary of papers

One of the long-standing controversies in Atlantic cod research is how genetic differences are maintained between migratory NEAC and stationary NCC despite gene flow. Earlier studies had associated LG01 with genetic differentiation between NEAC and NCC (Bradbury et al. 2010; Hemmer-Hansen et al. 2013; Karlsen et al. 2013) characterized by extended blocks of elevated linkage disequilibrium (LD) spanning several Mb, but the genomic architecture underlying these islands of genomic divergence has so far been unclear. To investigate if large adjacent inversions may explain these phenomenon, we used a combinatory approach consisting of whole genome sequencing and linkage mapping in a family material consisting of pure NEAC, pure NCC and NEACxNCC crosses together with samples of wild cod collected at 14 different locations distributed across the Northeast Atlantic Ocean from the Barents Sea in the north to the Celtic/Irish Sea in the south. Our results document that the island of genomic divergence at LG01 are instigated by two large adjacent inversions that differ between NEAC and NCC and suppress recombination in heterozygotes. We looked closely into the inverted region and identified a few candidate genes that may play an important role for NEAC fish to make long-distance migrations and vertical movements down to large depths. By comparing the synteny blocks with Northern pike and stickleback we revealed which haplotype represent the ancestral state of the inverted structure.

Extended LD spanning several Mb have also been detected on LGs 2, 7 and 12 in Atlantic cod and association studies with SNP-markers have linked these regions with temperature along the North-South gradient (Bradbury et al. 2013; Sodeland et al. 2016). However, the relatively low resolution of the SNP-arrays used in these studies prevent the identification of precise inversion breakpoints and the exact complement of genes they contain. Current genome assemblies for Atlantic cod, including the latest gadMor3, have been generated from NEAC, which is a migratory population feeding in the cold Barents Sea. To generate a genomic resource serving as a contrast to these assemblies we captured and sequences a male Atlantic cod from the Celtic sea, representing one of the most southernmost extreme populations of the species. A chromosome-level genome assembly (gadMor_Celtic), with a contig N50 of 10Mb was developed using nanopore long-read data and error corrected to reference quality using Illumina reads. By comparing the gadMor_Celtic and gadMor3 assemblies, that harbour alternative configurations of the four chromosomal rearrangements,

inversion breakpoints on LGs 1, 2, 7 and 12 were identified. In fact, in most cases these breakpoints are located within single contig in the gadMor_Celtic assembly.

Earlier studies have shown that sex in Atlantic cod is genetically determined (Haugen et al. 2011) but the sex determining locus and its exact genomic location has not been identified. By comparing the read depth of whole genome sequence data from multiple phenotypically sexed male and females, we identified a Y-specific sequence of 9,149 bp and a X-sequence of 425 bp on LG11. Long reads from PacBio and nanopore sequencing aligning specifically to either the X- or Y- sequence validated the integrity of the newly identified sequence. A diagnostic PCR test assessed on phenotypically sexed males and females confirmed the sex-specific nature of the X- and Y-sequences. Annotation using RNA sequence data from early larval stages revealed a novel single exon gene containing a zinc knuckle motif on the Y-sequence (named *zkY)*, which we propose to be the master masculinization gene in Atlantic cod. PCR amplification of Y-specific sequences in Arctic cod (*Arctogadus glacialis*) and Greenland cod (*Gadus macrocephalus ogac*) suggests that the Y-specific region emerged in codfishes more than 7.5 million years ago.

Discussion

The collapse of Atlantic cod populations across the North Atlantic is highly affected by overexploitation and rebuilding the population biomass is challenging (Sguotti et al. 2019). Apart from overfishing, other factors like pollution, acidification, and increased sea surface temperature may also affect Atlantic cod populations. For example, the on-going rise in sea surface temperature has been associated with negatively affecting cod productivity (Freitas et al. 2015; Clucas et al. 2019a; Sguotti et al. 2019) and growth (Brander 1995). Rapid changes in sea temperature especially at spawning time may also lead to subsequent effects on the distribution of adult cod population (Righton et al. 2010). During the last years, the NEAC population has increased substantially, whereas the NCC populations suffer by reduction (ICES 2015a, b). Due to the diminishing levels of NCC populations, fishing restrictions are now being enforced in coastal regions around Norway (Fiskeridirektoratet.no 2019). Therefore, understanding the population structure, its connectivity and adaptation to different environments is much needed for the management of Atlantic cod populations in a sustainable way.

In this thesis we have developed genomic resources utilizing high-throughput sequencing technologies, which enabled us to investigate not just the genetic differences between populations but specifically reveal mechanism that enables genetic divergence to occur between Atlantic cod populations. Additionally, we also explored the genome to identify genes involved in SD that can be beneficial to cod aquaculture initiatives.

Atlantic cod chromosomal inversions and adaptation to different environments

For decades, Atlantic cod in the Norwegian Sea are classified into two main groups, the migratory Northeast Arctic cod (NEAC) and stationary Norwegian coastal cod (NCC). In paper I, we characterized a double inversion in LG01, documented that it represses recombination in NEAC/NCC heterozygotes, and over time creates genomic divergence between the two populations. The island of genomic divergence on LG01 was found in earlier studies (Bradbury et al. 2010; 2013; Hemmer-Hansen et al. 2013; Karlsen et al. 2013), but paper 1 was the first study to document that large inversions is the underlying genetic mechanism responsible for the divergence between NEAC and NCC. It is suggested that chromosomal polymorphisms involving more than one inversion can suppress recombination effectively for prolonged periods of time and lead to genetically distinct

haplotypes (Huynh et al. 2011). Studies proceeding paper I confirmed that the inversions in LG01 is associated with divergence between migratory and stationary populations along the Icelandic and Canadian coasts as well (Berg et al. 2017; Sinclair-Waters et al. 2017). One noticeable difference between NEAC and NCC is the ability of NEAC to migrate and adjust the buoyancy of the body to high hydrostatic pressure at large sea depths. Studies have shown that the foraging and spawning migrations of NEAC involve vertical movements at depths of 200–400 m along stable thermal paths while stationary NCC are found in much shallower habitats (Stensholt 2001). This is supported by behavioral differences between juvenile NEAC and NCC settling at different depths (Fevolden et al. 2012). Data is suggesting that the large region on LG01 is under selection as a supergene. To identify possible underlying functional variation within the supergene we analysed short-read sequencing data from almost 100 individuals representing pure NEAC, pure NCC and NEACxNCC crosses using Illumina-sequencing technology. We revealed 849 plausible functional variants (e.g. non-synonymous SNP) in 321 genes that are fixed for alternative alleles in NEAC, NCC and heterozygous in NEACxNCC crosses. No fixed sequence differences were found outside the inverted region. This raises an interesting question, are there candidate genes within the inverted region (with plausible functional variants) that may facilitate NEAC to migrate and adapt to high hydrostatic pressure at large sea depths? In paper I, we focused on a few candidate genes within the inversion, mainly the enzyme *ca6* gene. *Ca6* is involved in the regulation of blood pH associated with changes in depth pressure in swim bladder of fishes (Fange 1953; Skinazi 1953; Pelster 2014). Using predictive modeling of the CA6 protein we show that NEAC may contain a more efficient variant of this protein than NCC, which may help NEAC in deep-water feeding. However, our approach to identify *ca6* as the best candidate for this function was largely observational and needs to be followed up in other studies to prove functional differences between the CA6 variants. While we focused mostly into the *ca6* gene, future studies should also consider other genes within the inverted region that may cause advantages for the migratory behavior of NEAC.

Previous studies have proposed that the large islands of genomic divergence on LGs 2, 7 and 12 are associated with environmental factors like water temperature (Bradbury et al. 2010; Berg et al. 2015), oxygen levels and salinity (Berg et al. 2015). Similar to LG01, these regions show similar patterns of divergence on both sides of the Atlantic (Barney et al. 2017; Berg et al. 2017). One strategy to identify the islands of divergence on the different

LGs was to utilize genotypes from multiple samples and populations to identify blocks of extended LD (Berg et al. 2015; Sodeland et al. 2016; Barth et al. 2017; Berg et al. 2017; Sinclair-Waters et al. 2017; Knutsen et al. 2018; Kess et al. 2019). Another approach was to use whole genome sequence, or at least partly sequence, multiple samples and compare them to the gadMor2 assembly (Barney et al. 2017; Clucas et al. 2019a; Clucas et al. 2019b; Puncher et al. 2019). The majority of papers exploring this have used Atlantic cod genome sequences generated from a NEAC individual (Star et al. 2011; Torresen et al. 2017). Very recently, the chromosome-scale gadMor3 assembly also generated from a NEAC fish was built from a broad range of data types including PacBio, optical maps and Hi-C. In paper II, we used nanopore sequencing from Oxford Nanopore Technologies (ONT) to generate a high-quality genome assembly from an Atlantic cod fished in the Celtic Sea, representing a southern population and displaying alternate configuration of all four rearrangements. This allowed us to characterize the inversion boundaries between the two chromosome-scale genome assemblies. Pairwise comparison of the two genomes revealed additional putative rearrangements on LGs 6,11 and 21, not been reported before. Whether these are assembly artefacts or real rearrangements need to be confirmed.

Traditionally, coupling de novo assembly with linkage mapping is a powerful way to produce a high-quality reference genome (Fierst 2015). In paper II, we use linkage mapping to order and orient the large contigs. An alternative approach would be to build Hi-C contact maps or optical maps, which are both methods, used to build the gadMor3 assembly. On top of these physical maps, gadMor3 also used linkage map (a different version than ours) to orient their scaffolds. Although the contig N50 of the gadMor_Celtic was higher than for gadMor3 (10 Mb and 1 Mb respectively), pairwise comparison of the two assemblies show that both approaches have been successful to produce chromosome-level assemblies for Atlantic cod.

In the future, advances in sequencing technology will likely lead to comprehensive sequencing of multiple Atlantic cod populations. This will enable researchers to explore the genomic architecture with increased resolution and understand the genetic basis of adaptive traits. Enabled by new sequencing technologies researchers are no longer constrained to ‘one reference genome for a species’ dogma. Thoughts are shifting towards a graph genome structures or pan-genomes that will allow reference genomes to evolve beyond a single, linear representation and capture the full diversity of a population. Hopefully, this will allow

us to spend less time assembling genomes and spend more time exploring the genome and further understand cod biology.

Characterisation of the sex determination region in Atlantic cod

The characterization of the male specific region in paper III brings us one step closer to elucidating the molecular mechanism of sex-determination (SD) in Atlantic cod and facilitates development of diagnostic tests to determine gender prior to physical maturation. However, to determine whether the *zkY* located within the region is a functional sex determiner demands further investigations. One way to study this may be to perform gene-editing experiments using CRISPR studies and to characterize the resulting biological effects.

References

Andersen O. 2012. Hemoglobin polymorphisms in Atlantic cod - a review of 50 years of study. *Marine Genomics* 8:59-65.

Andersen O, Johnsen H, De Rosa MC, et al. 2015. Evolutionary history and adaptive significance of the polymorphic Pan I in migratory and stationary populations of Atlantic cod (*Gadus morhua*). *Marine Genomics* 22:45-54.

Arnason E, Palsson S. 1996. Mitochondrial cytochrome b DNA sequence variation of Atlantic cod *Gadus morhua*, from Norway. *Molecular Ecology* 5:715-724.

Arthofer W, Heussler C, Krapf P, et al. 2018. Identifying the minimum number of microsatellite loci needed to assess population genetic structure: A case study in fly culturing. *Fly* 12:13-22.

Arthun M, Eldevik T, Smedsrud LH, et al. 2012. Quantifying the influence of Atlantic heat on Barents Sea ice variability and retreat. *Journal of Climate* 25:4736-4743.

Avise JC, Mank JE. 2009. Evolutionary perspectives on hermaphroditism in fishes. *Sexual Development* 3:152-163.

Barney BT, Munkholm C, Walt DR, et al. 2017. Highly localized divergence within supergenes in Atlantic cod (*Gadus morhua*) within the Gulf of Maine. *BMC Genomics* 18:271.

Barrett JH, Locker AM, Roberts CM. 2004. 'Dark Age Economics' revisited: the english fish bone evidence AD 600-1600. *Antiquity* 78:618-636.

Barrett JH, Orton D, Johnstone C, et al. 2011. Interpreting the expansion of sea fishing in medieval Europe using stable isotope analysis of archaeological cod bones. *Journal of Archaeological Science* 38:1516-1524.

Barth JMI, Berg PR, Jonsson PR, et al. 2017. Genome architecture enables local adaptation of Atlantic cod despite high connectivity. *Molecular Ecology* 26:4452-4466.

Barth JMI, Villegas-Rios D, Freitas C, et al. 2019. Disentangling structural genomic and behavioural barriers in a sea of connectivity. *Molecular Ecology* 28:1394-1411.

Berg E, Albert OT. 2003. Cod in fjords and coastal waters of North Norway: distribution and variation in length and maturity at age. *ICES Journal of Marine Science* 60:787-797.

Berg E, Pedersen T. 2001. Variability in recruitment, growth and sexual maturity of coastal cod (*Gadus morhua L.*) in a fjord system in northern Norway. *Fisheries Research* 52:179-189.

Berg PR, Jentoft S, Star B, et al. 2015. Adaptation to low salinity promotes genomic divergence in Atlantic cod (*Gadus morhua L.*). *Genome Biology and Evolution* 7:1644-1663.

Berg PR, Star B, Pampoulie C, et al. 2017. Trans-oceanic genomic divergence of Atlantic cod ecotypes is associated with large inversions. *Heredity* 119:418-428.

Berg PR, Star B, Pampoulie C, et al. 2016. Three chromosomal rearrangements promote genomic divergence between migratory and stationary ecotypes of Atlantic cod. *Scientific Reports* 6:23246. DOI: 10.1038/srep23246.

Bergstad OA, Jorgensen T, Dragesund O. 1987. Life-history and ecology of the gadoid resources of the Barents Sea. *Fisheries Research* 5:119-161.

Bradbury IR, Bowman S, Borza T, et al. 2014. Long distance linkage disequilibrium and limited hybridization suggest cryptic speciation in Atlantic cod. *PLoS One* 9(9): e106380.

Bradbury IR, Hubert S, Higgins B, et al. 2010. Parallel adaptive evolution of Atlantic cod on both sides of the Atlantic Ocean in response to temperature. *Proceedings of the Royal Society B-Biological Sciences* 277:3725-3734.

Bradbury IR, Hubert S, Higgins B, et al. 2013. Genomic islands of divergence and their consequences for the resolution of spatial structure in an exploited marine fish. *Evolutionary Applications* 6:450-461.

Brander KM. 1995. The effect of temperature on growth of Atlantic Cod (*Gadus morhua L*). *ICES Journal of Marine Science* 52:1-10.

Clucas GV, Kerr LA, Cadrin SX, et al. 2019a. Adaptive genetic variation underlies biocomplexity of Atlantic Cod in the Gulf of Maine and on Georges Bank. *PLoS One* 14.

Clucas GV, Lou RN, Therkildsen NO, et al. 2019b. Novel signals of adaptive genetic variation in northwestern Atlantic cod revealed by whole-genome sequencing. *Evolutionary Applications*. DOI: 10.1111/eva.12861

Dahle G, Quintela M, Johansen T, et al. 2018. Analysis of coastal cod (*Gadus morhua L.*) sampled on spawning sites reveals a genetic gradient throughout Norway's coastline. *BMC Genetics* 19:42.

Dalpadado P, Ingvaldsen RB, Stige LC, et al. 2012. Climate effects on Barents Sea ecosystem dynamics. *ICES Journal of Marine Science* 69:1303-1316.

Delghandi M, Mortensen A, Westgaard JI. 2003. Simultaneous analysis of six microsatellite markers in Atlantic cod (*Gadus morhua*): A novel multiplex assay system for use in selective breeding studies. *Marine Biotechnology* 5:141-148.

Devlin RH, Nagahama Y. 2002. Sex determination and sex differentiation in fish: an overview of genetic, physiological, and environmental influences. *Aquaculture* 208:191-364.

Fange R. 1953. The mechanisms of gas transport in the euphysoclist swimbladder. *Acta Physiol Scand Suppl* 30:1-133.

Feder JL, Egan SP, Nosil P. 2012. The genomics of speciation-with-gene-flow. *Trends in Genetics* 28:342-350.

Fevolden SE, Pogson GH. 1997. Genetic divergence at the synaptophysin (Syp I) locus among Norwegian coastal and north-east Arctic populations of Atlantic cod. *Journal of Fish Biology* 51:895-908.

Fevolden SE, Westgaard JI, Pedersen T, et al. 2012. Settling-depth vs. genotype and size vs. genotype correlations at the Pan I locus in 0-group Atlantic cod Gadus morhua. *Marine Ecology Progress Series* 468:267-278.

Fierst JL. 2015. Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools. *Frontiers in Genetics* 6:220.

Fossheim M, Primicerio R, Johannesen E, et al. 2015. Recent warming leads to a rapid borealization of fish communities in the Arctic. *Nature Climate Change* 5:673.

Freitas C, Olsen EM, Moland E, et al. 2015. Behavioral responses of Atlantic cod to sea temperature changes. *Ecology and Evolution* 5:2070-2083.

Frydenberg O, Moller D, Naevdal G, et al. 1965. Haemoglobin polymorphism in Norwegian cod populations. *Hereditas-Genetiskt Arkiv* 53:257.

Ghigliotti L, Fevolden SE, Cheng CH, et al. 2012. Karyotyping and cytogenetic mapping of Atlantic cod (*Gadus morhua* Linnaeus, 1758). *Animal Genetics* 43:746-752.

Godo OR, Michalsen K. 2000. Migratory behaviour of north-east Arctic cod, studied by use of data storage tags. *Fisheries Research* 48:127-140.

Guo BC, DeFaveri J, Sotelo G, et al. 2015. Population genomic evidence for adaptive differentiation in Baltic Sea three-spined sticklebacks. *BMC Biology* 13. doi: 10.1186/s12915-015-0130-8

Hattori RS, Murai Y, Oura M, et al. 2012. A Y-linked anti-mullerian hormone duplication takes over a critical role in sex determination. *PNAS* 109:2955-2959.

Haugen T, Andersson E, Norberg B, et al. 2011. The production of hermaphrodites of Atlantic cod (*Gadus morhua*) by masculinization with orally administered 17-alpha-methyltestosterone, and subsequent production of all-female cod populations. *Aquaculture* 311:248-254.

Hemmer-Hansen J, Nielsen EE, Therkildsen NO, et al. 2013. A genomic island linked to ecotype divergence in Atlantic cod. *Molecular Ecology* 22:2653-2667.

Heule C, Salzburger W, Bohne A. 2014. Genetics of sexual development: An evolutionary playground for fish. *Genetics* 196:579-591.

Hoffmann AA, Rieseberg LH. 2008. Revisiting the impact of inversions in evolution: From population genetic markers to drivers of adaptive shifts and speciation? *Annual Review of Ecology Evolution and Systematics* 39:21-42.

Hollowed AB, Sundby S. 2014. Ecology. Change is coming to the northern oceans. *Science* 344:1084-1085.

Howell WH, Morin M, Rennels N, et al. 2008. Residency of adult Atlantic cod (*Gadus morhua*) in the western Gulf of Maine. *Fisheries Research* 91:123-132.

Hubbard J. 2009. Cod: The ecological history of the North atlantic fisheries. *Technology and Culture* 50:249-250.

Hubert S, Higgins B, Borza T, et al. 2010. Development of a SNP resource and a genetic linkage map for Atlantic cod (*Gadus morhua*). *BMC Genomics* 11, 191.

Huynh LY, Maney DL, Thomas JW. 2011. Chromosome-wide linkage disequilibrium caused by an inversion polymorphism in the white-throated sparrow (*Zonotrichia albicollis*). *Heredity* 106:537-546.

ICES. 2015a. Cod (*Gadus morhua*) in subareas I and II (Northeast Arctic). ICES

ICES. 2015b. Cod (*Gadus morhua*) in Subareas I and II (Norwegian coastal waters cod). ICES.

Jakobsdottir KB, Jorundsdottir OD, Skirnisdottir S, et al. 2006. Nine new polymorphic microsatellite loci for the amplification of archived otolith DNA of Atlantic cod, *Gadus morhua L. Molecular Ecology Notes* 6:337-339.

Jakobsdottir KB, Pardoe H, Magnusson A, et al. 2011. Historical changes in genotypic frequencies at the Pantophysin locus in Atlantic cod (*Gadus morhua*) in Icelandic waters: evidence of fisheries-induced selection? *Evolutionary Applications* 4:562-573.

Jakobsen T. 1987. Coastal Cod in Northern Norway. *Fisheries Research* 5:223-234.

Johansen SD, Coucheron DH, Andreassen M, et al. 2009. Large-scale sequence analyses of Atlantic cod. *New Biotechnology* 25:263-271.

Johansen T, Westgaard JI, Seliussen BB, et al. 2018. "Real-time" genetic monitoring of a commercial fishery on the doorstep of an MPA reveals unique insights into the interaction between coastal and migratory forms of the Atlantic cod. *ICES Journal of Marine Science* 75:1093-1104.

Johnsen H, Tveiten H, Torgersen JS, et al. 2013. Divergent and sex-dimorphic expression of the paralogs of the Sox9-Amh-Cyp19a1 regulatory cascade in developing and adult atlantic cod (*Gadus morhua L.*). *Molecular Reproduction and Development* 80:358-370.

Jones FC, Grabherr MG, Chan YF, et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* 484:55-61.

Joron M, Frezal L, Jones RT, et al. 2011. Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature* 477:203-206.

Joron M, Papa R, Beltran M, et al. 2006. A conserved supergene locus controls colour pattern diversity in Heliconius butterflies. *PLoS Biology* 4:1831-1840.

Kamiya T, Kai W, Tasumi S, et al. 2012. A trans-species missense SNP in Amhr2 is associated with sex determination in the tiger pufferfish, *Takifugu rubripes* (fugu). *PLoS Genetics* 8:e1002798.

Karlsen BO, Klingan K, Emblem A, et al. 2013. Genomic divergence between the migratory and stationary ecotypes of Atlantic cod. *Molecular Ecology* 22:5098-5111.

Kess T, Bentzen P, Lehnert SJ, et al. 2019. A migration-associated supergene reveals loss of biocomplexity in Atlantic cod. *Science Advances* 5:eaav2461.

Keyl F, Kempf AJ, Sell AF. 2015. Sexual size dimorphism in three North Sea gadoids. *Journal of Fish Biology* 86:261-275.

Kirkpatrick M. 2010. How and why chromosome inversions evolve. *PLoS Biology* 8, e1000501.

Kirkpatrick M, Barton N. 2006. Chromosome inversions, local adaptation and speciation. *Genetics* 173:419-434.

Kirubakaran TG, Grove H, Kent MP, et al. 2016. Two adjacent inversions maintain genomic differentiation between migratory and stationary ecotypes of Atlantic cod. *Molecular Ecology* 25:2130-2143.

Knutsen H, Jorde PE, Andre C, et al. 2003. Fine-scaled geographical population structuring in a highly mobile marine species: the Atlantic cod. *Molecular Ecology* 12:385-394.

Knutsen H, Jorde PE, Hutchings JA, et al. 2018. Stable coexistence of genetically divergent Atlantic cod ecotypes at multiple spatial scales. *Evolutionary Applications* 11:1527-1539.

Li M, Sun Y, Zhao J, et al. 2015. A tandem duplicate of anti-mullerian hormone with a missense SNP on the Y chromosome is essential for male sex determination in nile tilapia, *Oreochromis niloticus*. *PLoS Genet* 11:e1005678.

Low WY, Tearle R, Bickhart DM, et al. 2019. Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity. *Nature Communications* 10:260.

Matsuda M, Nagahama Y, Shinomiya A, et al. 2002. *DMY* is a Y-specific DM-domain gene required for male development in the medaka fish. *Nature* 417:559-563.

Miller KM, Le KD, Beacham TD. 2000. Development of tri- and tetranucleotide repeat microsatellite loci in Atlantic cod (*Gadus morhua*). *Molecular Ecology* 9:238-239.

Moen T, Delghandi M, Wesmajervi MS, et al. 2009. A SNP/microsatellite genetic linkage map of the Atlantic cod (*Gadus morhua*). *Animal Genetics* 40:993-996.

Moller D. 1966. Genetic differences between cod groups in Lofoten area. *Nature* 212:824.

Moller D. 1968. Genetic diversity in spawning cod along Norwegian coast. *Hereditas-Genetiskt Arkiv* 60:1.

Myosho T, Otake H, Masuyama H, et al. 2012. Tracing the emergence of a novel sex-determining gene in medaka, *Oryzias luzonensis*. *Genetics* 191:163-170.

Nanda I, Kondo M, Hornung U, et al. 2002. A duplicated copy of DMRT1 in the sex-determining region of the Y chromosome of the medaka, *Oryzias latipes*. *PNAS* 99:11778-11783.

Neat FC, Bendall V, Berx B, et al. 2014. Movement of Atlantic cod around the British Isles: implications for finer scale stock management. *Journal of Applied Ecology* 51:1564-1574.

Nielsen EE, Hansen MM, Meldrup D. 2006. Evidence of microsatellite hitch-hiking selection in Atlantic cod (*Gadus morhua L.*): implications for inferring population structure in nonmodel organisms. *Molecular Ecology* 15:3219-3229.

Nielsen EE, Hansen MM, Schmidt C, et al. 2001. Fisheries - Population of origin of Atlantic cod. *Nature* 413:272-272.

Nosil P, Feder JL. 2012. Genomic divergence during speciation: causes and consequences. *Philos Trans R Soc Lond B Biol Sci* 367:332-342.

Nosil P, Harmon LJ, Seehausen O. 2009. Ecological explanations for (incomplete) speciation. *Trends in Ecology & Evolution* 24:145-156.

O'Leary DB, Coughlan J, Dillane E, et al. 2007. Microsatellite variation in cod *Gadus morhua* throughout its geographic range. *Journal of Fish Biology* 70:310-335.

Otterlei E, Nyhammer G, Folkvord A, et al. 1999. Temperature- and size-dependent growth of larval and early juvenile Atlantic cod (*Gadus morhua*): a comparative study of Norwegian coastal cod and northeast Arctic cod. *Canadian Journal of Fisheries and Aquatic Sciences* 56:2099-2111.

Ottersen G, Michalsen K, Nakken O. 1998. Ambient temperature and distribution of north-east Arctic cod. *ICES Journal of Marine Science* 55:67-85.

Pampoulie C, Jakobsdottir KB, Marteinsdottir G, et al. 2008a. Are vertical behaviour patterns related to the pantophysin locus in the Atlantic cod (*Gadus morhua L.*)? *Behavior Genetics* 38:76-81.

Pampoulie C, Ruzzante DE, Chosson V, et al. 2006. The genetic structure of Atlantic cod (*Gadus morhua*) around Iceland: insight from microsatellites, the Pan I locus, and tagging experiments. *Canadian Journal of Fisheries and Aquatic Sciences* 63:2660-2674.

Pampoulie C, Stefansson MO, Jorundsdottir TD, et al. 2008b. Recolonization history and large-scale dispersal in the open sea: the case study of the North Atlantic cod, *Gadus morhua L. Biological Journal of the Linnean Society* 94:315-329.

Pearse DE, Nicola J. Barson, Torfinn Nome, et al. 2018. Sex-dependent dominance maintains migration supergene in rainbow trout. *BioRxiv*. DOI: https://doi.org/10.1101/504621.

Pelster B. 2014. Swimbladder function and the spawning migration of the European eel *Anguilla anguilla*. *Frontiers in Physiology* 5:486.

Pettersson ME, Rochus CM, Han F, et al. 2019. A chromosome-level assembly of the Atlantic herring genome-detection of a supergene and other signals of selection. *Genome Research* 29:1919–1928.

Pogson GH. 2001. Nucleotide polymorphism and natural selection at the pantophysin (Pan I) locus in the Atlantic cod, *Gadus morhua* (L.). *Genetics* 157:317-330.

Puncher GN, Rowe S, Rose GA, et al. 2019. Chromosomal inversions in the Atlantic cod genome: Implications for management of Canada's Northern cod stock. *Fisheries Research* 216:29-40.

Rane RV, Rako L, Kapun M, et al. 2015. Genomic evidence for role of inversion 3RP of Drosophila melanogaster in facilitating climate change adaptation. *Molecular Ecology* 24:2423-2432.

Righton DA, Andersen KH, Neat F, et al. 2010. Thermal niche of Atlantic cod *Gadus morhua*: limits, tolerance and optima. *Marine Ecology Progress Series* 420:1-U344.

Rollefsen G. 1935. The spawning zone in cod otoliths and prognosis of stock. Bergen: A.s J. Griegs boktrykkeri.

Sarvas TH, Fevolden SE. 2005. Pantophysin (Pan I) locus divergence between inshore v. offshore and northern v. southern populations of Atlantic cod in the north-east Atlantic. *Journal of Fish Biology* 67:444-469.

Sguotti C, Otto SA, Frelat R, et al. 2019. Catastrophic dynamics limit Atlantic cod recovery. *Proceedings of the Royal Society B-Biological Sciences* 286.

Sinclair-Waters M, Bradbury IR, Morris CJ, et al. 2017. Ancient chromosomal rearrangement associated with local adaptation of a postglacially colonized population of Atlantic Cod in the northwest Atlantic. *Molecular Ecology* 27:339-351.

Skinazi L. 1953. Carbonic anhydrase in two closely related teleosts; inhibition of the secretion of gas from the air bladder of perch by sulfonamides. *C R Seances Soc Biol Fil* 147:295-299.

Smedsrud LH, Esau I, Ingvaldsen RB, et al. 2013. The Role of the Barents Sea in the Arctic Climate System. *Reviews of Geophysics* 51:415-449.

Sodeland M, Jorde PE, Lien S, et al. 2016. "Islands of Divergence" in the Atlantic cod genome represent polymorphic chromosomal rearrangements. *Genome Biology and Evolution* 8:1012-1022.

Star B, Boessenkool S, Gondek AT, et al. 2017. Ancient DNA reveals the Arctic origin of Viking Age cod from Haithabu, Germany. *PNAS* 114:9152-9157.

Star B, Nederbragt AJ, Jentoft S, et al. 2011. The genome sequence of Atlantic cod reveals a unique immune system. *Nature* 477:207-210.

Star B, Torresen OK, Nederbragt AJ, et al. 2016. Genomic characterization of the Atlantic cod sex-locus. *Scientific Reports* 6:31235.

Stensholt BK. 2001. Cod migration patterns in relation to temperature: analysis of storage tag data. *ICES Journal of Marine Science* 58:770-793.

Sundby S, Nakken O. 2008. Spatial shifts in spawning habitats of Arcto-Norwegian cod related to multidecadal climate oscillations and climate change. *ICES Journal of Marine Science* 65:953-962.

Svasand T, Jorstad KE, Ottera H, et al. 1996. Differences in growth performance between Arcto-Norwegian and Norwegian coastal cod reared under identical conditions. Jo*urnal of Fish Biology* 49:108-119.

Takehana Y, Matsuda M, Myosho T, et al. 2014. Co-option of Sox3 as the male-determining factor on the Y chromosome in the fish Oryzias dancena. *Nature Communications* 5.

Tanaka K, Takehana Y, Naruse K, et al. 2007. Evidence for different origins of sex chromosomes in closely related Oryzias fishes: Substitution of the master sex-determining gene. *Genetics* 177:2075-2081.

Thorsen A, Witthames PR, Marteinsdottir G, et al. 2010. Fecundity and growth of Atlantic cod (*Gadus morhua L.*) along a latitudinal gradient. *Fisheries Research* 104:45-55.

Torresen OK, Star B, Jentoft S, et al. 2017. An improved genome assembly uncovers prolific tandem repeats in Atlantic cod. *BMC Genomics* 18:95.

Twyford AD, Friedman J. 2015. Adaptive divergence in the monkey flower Mimulus guttatus is maintained by a chromosomal inversion. *Evolution* 69:1476-1486.

Vandermeeren T, Jorstad KE, Solemdal P, et al. 1994. Growth and survival of cod larvae (*Gadus-Morhua L*) - comparative enclosure etudies of Northeast Arctic cod and coastal cod from western Norway. Cod and climate change - Proceedings of a Symposium 198:633-645.

Volff JN, Nanda I, Schmid M, et al. 2007. Governing sex determination in fish: Regulatory putsches and ephemeral dictators. *Sexual Development* 1:85-99.

Volff JN, Schartl M. 2001. Variability of genetic sex determination in poeciliid fishes. *Genetica* 111:101-110.

Wang L, Liu S, Zhuang Z, et al. 2013. Population genetic studies revealed local adaptation in a high gene-flow marine fish, the small yellow croaker (*Larimichthys polyactis*). *PLoS One* 8:e83493.

Westgaard JI, Fevolden SE. 2007. Atlantic cod (*Gadus morhua L.*) in inner and outer coastal zones of northern Norway display divergent genetic signature at non-neutral loci. *Fisheries Research* 85:306-315.

Yano A, Guyomard R, Nicol B, et al. 2012. An immune-related gene evolved into the master sex-determining gene in rainbow trout, *Oncorhynchus mykiss*. *Current Biology* 22:1423-1428.

Yoshida K, Terai Y, Mizoiri S, et al. 2011. B chromosomes have a functional effect on female sex determination in lake victoria cichlid fishes. *PLoS Genetics* 7: e1002203.

Zhang J, Chiodini R, Badr A, et al. 2011. The impact of next-generation sequencing on genomics. *Journal of Genetics and Genomics* 38:95-109.

Papers I-III

Paper I

ISSN 0962-1083

VOLUME 25
NUMBER 10
MAY
2016

MOLECULAR ECOLOGY

FROM THE COVER: Two adjacent inversions maintain genomic differentiation between migratory and stationary ecotypes of Atlantic cod. See pp. 2130–2143.

Published by
WILEY Blackwell

MOLECULAR ECOLOGY

Molecular Ecology (2016) **25**, 2130–2143 doi: 10.1111/mec.13592

FROM THE COVER

Two adjacent inversions maintain genomic differentiation between migratory and stationary ecotypes of Atlantic cod

TINA GRACELINE KIRUBAKARAN,* HARALD GROVE,* MATTHEW P. KENT,* SIMEN R. SANDVE,* MATTHEW BARANSKI,† TORFINN NOME,* MARIA CRISTINA DE ROSA,‡ BENEDETTA RIGHINO,‡ TORILD JOHANSEN,§ HÅKON OTTERÅ,§ ANNA SONESSON,† SIGBJØRN LIEN* and ØIVIND ANDERSEN*†

**Centre for Integrative Genetics (CIGENE), Department of Animal and Aquaculture Sciences (IHA), Norwegian University of Life Sciences (NMBU), PO Box 5003, Ås N-1430, Norway, †Nofima, PO Box 5010, Ås N-1430, Norway, ‡Institute of Chemistry of Molecular Recognition – CNR and Institute of Biochemistry and Clinical Biochemistry, Catholic University of Rome, Rome 00168, Italy, §Institute of Marine Research, PO Box 6404, Tromsø N-9294, Norway*

Abstract

Atlantic cod is composed of multiple migratory and stationary populations widely distributed in the North Atlantic Ocean. The Northeast Arctic cod (NEAC) population in the Barents Sea undertakes annual spawning migrations to the northern Norwegian coast. Although spawning occurs sympatrically with the stationary Norwegian coastal cod (NCC), phenotypic and genetic differences between NEAC and NCC are maintained. In this study, we resolve the enigma by revealing the mechanisms underlying these differences. Extended linkage disequilibrium (LD) and population divergence were demonstrated in a 17.4-Mb region on linkage group 1 (LG1) based on genotypes of 494 SNPs from 192 parents of farmed families of NEAC, NCC or NEACxNCC crosses. Linkage analyses revealed two adjacent inversions within this region that repress meiotic recombination in NEACxNCC crosses. We identified a NEAC-specific haplotype consisting of 186 SNPs that was fixed in NEAC sampled from the Barents Sea, but segregating under Hardy–Weinberg equilibrium in eight NCC stocks. Comparative genomic analyses determine the NEAC configuration of the inversions to be the derived state and date it to ~1.6–2.0 Mya. The haplotype block harbours 763 genes, including candidates regulating swim bladder pressure, haem synthesis and skeletal muscle organization conferring adaptation to long-distance migrations and vertical movements down to large depths. Our results suggest that the migratory ecotype experiences strong directional selection for the two adjacent inversions on LG1. Despite interbreeding between NEAC and NCC, the inversions are maintaining genetic differentiation, and we hypothesize the co-occurrence of multiple adaptive alleles forming a 'supergene' in the NEAC population.

Keywords: chromosomal inversion, gene flow, local adaptation, recombination, supergene, swim bladder

Received 30 November 2015; revision received 8 February 2016; accepted 17 February 2016

Introduction

Adaptive divergence among populations and ultimately speciation is one of the most central subjects in evolutionary biology. Marine fishes are typically associated with shallow population structure and are highly dispersed in various ecological habitats, thus providing excellent models for studying interactions between the homogenizing effect of gene flow and the diversifying effect of selection (Hauser & Carvalho 2008; Nielsen *et al*. 2009). While reduced genetic exchange between

Correspondence: Øivind Andersen, Fax: +47 77629100; E-mail: oivind.andersen@nofima.no

sympatric cryptic species has mainly been explained by various reproductive barriers (Palumbi 1994; Monteiro *et al*. 2012; Shen *et al*. 2015), the adaptive significance of chromosome rearrangements by suppressing recombination between heterozygotes has been evidenced in many plant and animal studies, including coral reef fish species and in marine and freshwater ecotypes of threespined stickleback (*Gasterosteus aculeatus*) (Jones *et al*. 2012; Martinez *et al*. 2015).

Atlantic cod are widely distributed on the continental shelves and banks on both sides of the North Atlantic Ocean and represent the main demersal fish resource in these regions. The success of this highly exploited fish seems to be related to the different life history strategies of the multiple migratory and stationary populations, but careful management is required as several stocks have been dramatically reduced as a result of overfishing, climate change and pollution (Myers *et al*. 1997; Christensen *et al*. 2003; Mackenzie *et al*. 2004; Robichaud & Rose 2004). Cod fishery dates back to the tenth century A.D. when Vikings used dried Skrei (Old Norse *skriða* means wandering) as a source of nutrition and currency along the European trade routes. Today, Skrei are synonymous with the large Northeast Arctic cod (NEAC) population, which feeds in the Barents Sea and near Svalbard, but the adults undertake annual long-distance migrations to and from the spawning banks along the coast of North Norway, mainly offshore the Lofoten Archipelago (Bergstad & Dragesun 1987; Sundby & Nakken 2008; Ottersen *et al*. 2014). During foraging and spawning migrations, NEAC perform vertical movements down to depths of about 500 m with frequent descending and ascending swimming spanning up to 250 m (Godø & Michalsen 2000; Stensholt 2001). In contrast, the stationary Norwegian coastal cod (NCC) live in shallow coastal waters and fjords throughout the year and generally migrate only short distances at depths down to about 100 m (Hobson *et al*. 2007; Michalsen *et al*. 2014). The vertical divergence between NEAC and NCC is apparent at the 0-group stage when juveniles settle in deep and shallow water, respectively, in northern Norwegian fjords (Løken *et al*. 1994; Fevolden *et al*. 2012). In Iceland, similar ecotypes are represented by the frontal (migratory) and coastal (nonmigratory) populations, which exploit different habitats at depths of 200–600 m and <200 m, respectively (Pálsson & Thorsteinsson 2003; Pampoulie *et al*. 2008; Grabowski *et al*. 2011).

The population structure of NEAC and NCC and the interactions between gene flow and natural selection have been a controversial subject since the 1930s. Although NEAC and NCC occur sympatrically on local spawning grounds with potentially high levels of gene flow, phenotypic and genetic differences are maintained between the two populations (Rollefsen 1933, 1954; Jakobsen 1987; Løken & Pedersen 1996; Nordeide & Pettersen 1998; Nordeide *et al*. 2011). Whereas phenotypic traits such as otolith morphology and vertebrae numbers are likely influences by temperature and hydrostatic pressure at the nursery grounds, strongly divergent allele frequencies have been found in several candidate genes for adaptation to different ecosystems, including pantophysin (*Pan* I), haemoglobin and rhodopsin (Møller 1966, 1968; Fevolden & Pogson 1997; Pogson 2001; Andersen *et al*. 2015; Pampoulie *et al*. 2015). Studying the genetic diversity in Atlantic cod along the Norwegian coast, Møller (1966, 1968, 1969) concluded that NEAC and NCC form two genetically separated populations or sibling species. The extreme divergence between NEAC and NCC at the *Pan* I locus has later been explained by differences in breeding structure, as selection alone would be insufficient to cause the observed levels of genetic differentiation (Fevolden & Sarvas 2001; Sarvas & Fevolden 2005; Westgaard & Fevolden 2007). The Icelandic frontal and coastal ecotypes were consistently proposed to be cryptic species within the Atlantic cod complex that have adapted to environmental factors in shallow and deep waters (Halldórsdóttir & Árnason 2016). Differences in courtship, spawning behaviour or spawning depths could hinder interbreeding between the populations or siblings (Hutchings *et al*. 1999; Nordeide & Folstad 2000; Grabowski *et al*. 2011). However, analysis of the mitochondrial genome revealed no reproductive isolation between NEAC and NCC (Árnason & Pálsson 1996; Karlsen *et al*. 2014), supporting the hypothesis that there is genetic communication between all cod stocks in the North Atlantic (Mork & Sundnes 1985). Accordingly, the genetic differences between NEAC and NCC were suggested to be produced by contemporary selection acting on settling cohorts (Williams 1975; Mork & Sundnes 1985).

The genetic differentiation of NEAC and NCC was recently shown to be uniquely associated with a large genomic region potentially harbouring hundreds of genes on linkage group 1 (LG1) (Hemmer-Hansen *et al*. 2013; Karlsen *et al*. 2013). The underlying mechanism was not unravelled in these studies, but a combination of divergence hitchhiking and chromosomal rearrangement was proposed to be responsible for the genomic island of divergence. We have finally resolved the enigma behind the strong genetic divergence between migratory NEAC and stationary NCC by the identification of a double inversion on LG1 that repress recombination within heterozygotes preventing introgression between cosegregating haplotypes.

Materials and methods

Fish material and DNA extraction

Wild cod were collected from 14 locations ranging from the Irish Sea in the south to the Barents Sea in the north (see Fig. 2). On average, 48 samples were collected from each location. To obtain a representational sampling of NEAC, we collected cod from two locations in the Barents Sea. Farmed cod were sampled from 88 families of the National cod breeding programme maintained by Nofima in Tromsø, Norway, and from eight families of the CODBIOBANK at the Institute of Marine Research in Bergen, Norway.

Totally, 104 cod from the National cod breeding programme were selected for sequencing comprising 50 fish of NEAC origin, 11 fish of NCC origin and 43 fish offspring of NEAC × NCC crosses. The sequenced fish belonged to year classes 2005 (P) and 2006 (F1) and represented the second generation of cod produced in captivity. The original broodstock in the base population were sampled from different geographical areas along the Norwegian coast and were assigned to the NCC and NEAC populations based on sampling locations and the *Pan* I^A and I^B alleles (Fevolden & Pogson 1997; Bangera *et al*. 2013). The Greenland cod (*Gadus macrocephalus ogac*) used to date the inversion was sampled at the Uummannaq Island, Northwest Greenland.

DNA was extracted using either a DNeasy kit from Qiagen (Hilden, Germany) according to the manufacturer's instructions or a high salt precipitation method (http://www.liv.ac.uk/~kempsj/IsolationofDNA.pdf). DNA quality was assessed by electrophoresis on 1% agarose gel to estimate the proportion of high molecular weight (HMW) DNA, and low-quality samples with negligible levels of HMW DNA were excluded from analysis. DNA concentration was assessed fluorometrically using Qubit technologies (Thermo Fisher Scientific, Carlsbad, USA).

Genotyping

Farmed (n = 2951) and wild fish (n = 959) were genotyped for 10 913 SNPs using an Illumina custom Infinium II SNP array (M. Kent, T. G. Kirubakaran, P. R. Berg, M. Baranski, G, Dahle, K. S. Jakobsen, S. Jentoft, T. Johansen, L. Nederbragdt, T, Nome, B. Star & S. Lien, in prep) according to the manufacturer's instructions (Illumina, San Diego, USA). Poorly performing samples displaying call rates below 0.9 were excluded from the analysis. Genotype data was preprocessed by removing low (<0.05) MAF (minor allele frequency) SNPs, and Mendelian errors were set to missing and imputed along with any other failed genotypes using BEAGLE v4 (Browning & Browning 2007). Wild populations were phased using SHAPEIT v2 (https://mathgen.stats.ox.ac.uk/shapeit), and the family material was accurately phased using linkage information. Phased data for 192 parents were used to estimate linkage disequilibrium (LD) between SNPs using HAPLOVIEW 4.2 (Barrett *et al*. 2005). All NEAC samples from the Barents Sea were homozygous for a haplotype consisting of 186 SNPs from the SNP array (see Table S1, Supporting information), and the wild fish were assigned to NEAC, NCC or a cross using this NEAC haplotype.

Linkage mapping and inversion detection

The construction of linkage maps for cod using 12K SNP array is described in detail elsewhere (H. Grove, T. G. Kirubakaran, M. Kent, M. Baranski, T. Nome, S. Sandve, P. R. Berg, M. Sodeland, G. Dahle, A. Sonesson, T. Johansen, Ø. Andersen & S. Lien, in prep). The analyses are mainly based on the CRIMAP linkage program (Green *et al*. 1990) and begin with performing two-point linkage in CRIMAP to sort SNPs into linkage groups. In the present study, we used 494 SNPs mapping to LG1 for the construction of separate multipoint linkage maps for pure NEAC, pure NCC and NEAC × NCC crosses. The sorting of family material for these analyses was determined by haplotyping parents using the 186 SNP set described above.

SNPs on the cod 12K SNP array were carefully chosen to tag as many contigs as possible (M. Kent, T. G. Kirubakaran, P. R. Berg, M. Baranski, G. Dahle, K. S. Jakobsen, S. Jentoft, T. Johansen, L. Nederbragdt, T, Nome, B. Star & S. Lien, in prep) and are well distributed along the linkage groups, thereby also forming a good foundation for building a chromosome sequence for LG1. Scaffolds from two draft assemblies, containing at least one SNP from the linkage map, were selected and used for the construction of the chromosome files. Erroneous scaffolds containing SNPs from more than one LG were broken between conflicting SNP positions. Overlapping scaffolds were identified by comparing SNPs mapping to both assemblies and were merged using coordinates from alignment with LASTZ (Harris 2007), resulting in a total of 40 scaffolds that were used to build the final chromosome sequence. Subsequently, linkage maps were then updated to take into account the more precise SNP order given by individual scaffolds. Finally, all scaffolds were oriented, ordered and concatenated into a new chromosome sequence based on information from the linkage map. The size of the final chromosome sequence for LG1 was 29 521 491 bp.

Sequencing and variant detection

Genomic DNA from the 104 breeding programme fish was prepared for sequencing using the Truseq Library

prep kit from Illumina (Illumina, San Diego, USA). Paired-end sequencing (2 × 100 nts) was carried out using an Illumina HiSeq 2000 instrument to generate approximately 10× coverage of the genome for each sample. Reads were processed using default parameters in TRIMMOMATIC version 0.32 (Bolger *et al.* 2014) before being aligned to the unmasked reference genome based on the NCC map described above using BOWTIE2 version 2.2.3 (Langmead & Salzberg 2012). Within-sample variant detection was performed using GATK HAPLOTYPECALLER version 2.8-1-g932cd3a (McKenna *et al.* 2010). SNPEFF version 4.0e (Cingolani *et al.* 2012) was used to annotate allelic variants. Individual variant calls with a quality score of <20 were excluded from further analysis, as were INDELs and genotypes with read depths below 6 or above 27. Also, variants not detected in >70% of the samples were removed across all samples.

Genomic DNA from a single Greenland cod was prepared for sequencing using a Nextera XT library preparation kit generating a library with an average size of 650 bp. Sequencing was performed using a MiSeq platform with V3 kit chemistry to generate 2 × 301-nt paired-end reads. A total of 18.7 M reads generated 11.2-Gb sequence data. Reads were mapped and variants detected as described above.

Pairwise LD (measured as r^2) for the whole linkage group was calculated based on 48 sequenced NEAC samples using PLINK v1.9 (https://www.cog-genomics.org/plink2) with MAF > 0.1 and HWE > 0.001.

Gene annotation

Gene models were built from multiple data sources including (i) approximately 3 million transcriptome reads (http://www.ncbi.nlm.nih.gov/sra?term=SRP013269) obtained from liver, egg, brain, head, kidney, hindgut, gonad and spleen, generated using GS-FLX 454 Titanium platform (Roche, Switzerland), (ii) ESTs from NCBI (n = 257 218), (iii) predicted RNAs (n = 1541, http://www.codgenome.no/data/ATLCOD1_ANN/) and (iv) approximately 35 million short read mRNA sequences from whole NEAC larvae at 12 and 35 days posthatching (Johnsen & Andersen 2012). To enable model building, short reads were mapped to the reference genome sequence using STAR (version 2.3.1z12) (Dobin *et al.* 2013), while long 454 transcriptome reads were mapped using GMAP (version 2014-07-28) (Thomas & Watanabe 2005) with '–no-chimeras' parameter in addition to default parameters. CUFFLINKS (version 2.2.1) (Trapnell *et al.* 2010) with '–multi-read-correct' parameter in addition to the default parameter assembled the aligned RNA-Seq reads and transcriptome reads into transcripts. Transcript models from RNA-Seq and 454 transcriptome were merged using cuffmerge.

Open reading frame (ORF) prediction was carried out using TransDecoder (http://transdecoder.github.io/) (Haas *et al.* 2013) using the pfamA and pfamB databases for homology searches (–search_pfam) and a minimum length of 30 amino acids for ORFs without pfam support (-m 30). In addition to the pfam homology evidence, we also performed BLASTP (evalue<1e-10) for all predicted proteins against zebrafish (*Danio rerio*) (v9.75) and three-spined stickleback (BROADS1.75) annotations downloaded from Ensembl. Only gene models with support from at least one type of homology search (pfam or BLASTP) were kept.

In total, we mapped 35 million RNA-Seq reads and 3.3 million 454 transcriptome sequences to the whole genome and used this to annotate LG1. A total of 2323 transcripts were left after merging transcript models using cuffmerge. Functional annotations of the transcripts were carried out using BLASTX against the SWISSPROT database. Results from TransDecoder and homology support filtering of putative protein-coding loci are shown in Table S2 (Supporting information).

Origin and dating of inversions

To determine whether NEAC or NCC represents the ancestral state of the inversions, we aligned LG1 sequences representing possible arrangements of the inversions with Northern pike (*Esox lucius*) and stickleback using LASTZ in gap-free mode requiring ≥75% identity and match-count filtering of 100 (Harris 2007).

Hierarchical clustering of the wild stocks was estimated based on genotypes from the SNP array using the R package SNPrelate (Zheng *et al.* 2012). Four linkage groups (LG1, LG2, LG7 and LG12) were excluded from this analysis due to the presence of extended LD (own data; Bradbury *et al.* 2010; Hemmer-Hansen *et al.* 2013). Reads generated from whole-genome sequencing of a single Greenland cod were compared with NEAC and NCC variant calls to identify a set of fixed sequence differences (FSD; single nucleotides fixed within populations) along LG1. FSD counts were then used to calculate pairwise differences among Greenland cod, NEAC and NCC. Under the assumption of a constant clock, we then estimated the NEAC-NCC divergence age relative to their divergence from Greenland cod by calculating the ratio between NEAC-NCC FSD distance and the mean FSD distance between Greenland cod-NEAC and Greenland cod-NCC (i.e. $FSD_{NEAC\text{-}NCC}/FSD_{mean\ (Greenland\ cod\text{-}NEAC,\ Greenland\ cod\text{-}NCC)}$) as shown in Fig. S1 (Supporting information).

Protein modelling

The carbonic anhydrase isoform 6 (*ca6*) was identified as a candidate gene within the double inversion because of

its key role in reducing pH levels in the gas gland of the swim bladder. Homology modelling was performed with the MODELLER software (Sali & Blundell 1993) to build the three-dimensional structure of the NEAC and NCC variants of Ca6 based on the crystal structure of human Ca6 as template (PDB code 3FE4, Pilka *et al.* 2012). The sequences were aligned using CLUSTALW, and identities between targets and template of 58% (NEAC) and 56% (NCC) allowed using the standard MODELLER protocol implemented in DISCOVERYSTUDIO v4.5 (Biovia). We ascertained that no other protein with a known related structure displayed a greater sequence similarity. The best of 50 models according to the PDF (Probability Density Function) score included in MODELLER was selected. The structures were inspected with PROCHECK (Laskowski *et al.* 1993) for inappropriate stereochemistry. Ramachandran maps of NEAC and NCC models revealed that they contained 91.7% of non-Gly-non-Pro residues in most favoured, 7.8% in additional allowed, 0.5% in generously allowed and 0.0% in disallowed regions. These models were further validated for their structure quality by Verify 3D available at http://services.mbi.ucla.edu/, and 95% of the residues of the modelled proteins showed a satisfactory 3D-1D score (>0.2). DISCOVERYSTUDIO v4.5 (Biovia) software was used to visualize the generated models.

Results

Linkage map and LD calculations

A genetic map describing 23 linkage groups in Atlantic cod (H. Grove, T. G. Kirubakaran M. Kent, M. Baranski, T. Nome, S. Sandve, P. R. Berg, M. Sodeland, G. Dahle, A. Sonesson, T. Johansen, Ø. Andersen & S. Lien, in prep.) was constructed by genotyping a large family material of 2739 individuals using a 12K SNP array (Kent *et al.* in prep). The map constructed for LG1 contained 494 SNPs (Fig. 1a; Table S1, Supporting information) and was used to integrate, order and orientate scaffolds from two draft cod assemblies into a cohesive chromosome sequence comprising 29.52 Mb. Accurately phased genotypes from 192 parents were used to estimate LD between SNPs and revealed a distinct block of extended LD from 10 to 27 Mb (Fig. 1c), embracing the *Pan* I locus located at 17.4 Mb. The parents were of known origin and classed as pure NEAC, pure NCC or NEAC × NCC crosses based on Pan I genotyping. Analyses of pure NEAC cod, captured from two locations in the Barents Sea, identified a single haplotype of 186 nonconsecutive SNPs that were homozygous in all individuals (Table S1, Supporting information). All

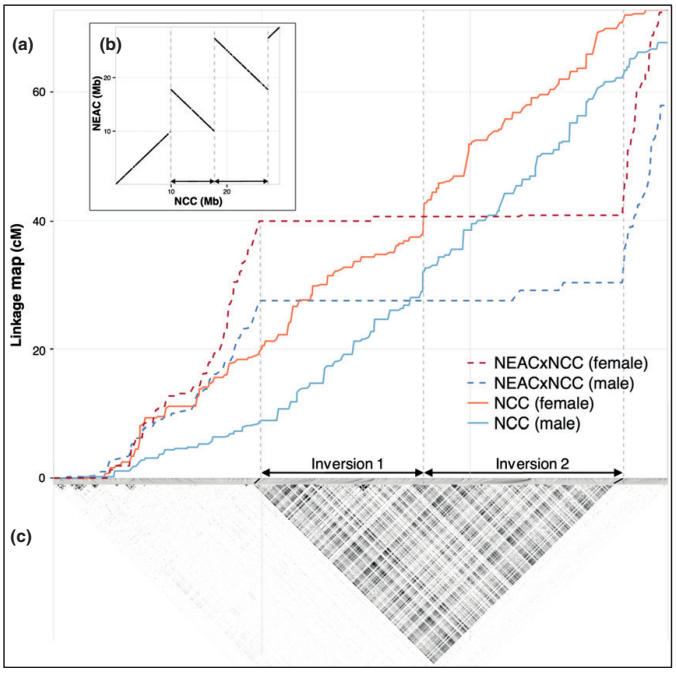

Fig. 1 (a) Linkage map for LG1 created separately for pure NCC and NEAC × NCC crosses. (b) Whole chromosome alignment between the NCC and NEAC sequence. (c) Pairwise LD calculated in 192 parents from the linkage families. Two large inversions inhibit recombination in NEACxNCC crosses corresponding to a region of extended LD on LG1.

NEAC × NCC crosses had one copy of this haplotype, while the NEAC haplotype was completely absent in pure NCC samples.

Because this distinct haplotype in NEAC indicated a substantial differentiation between NEAC and NCC cod, we constructed linkage maps separately for pure NEAC, pure NCC and NEAC × NCC crosses (Table S1, Supporting information). Pure NEAC and NCC showed typical recombination rates between SNPs along the length of LG1, but comparing the linkage maps disclosed a different SNP order within the block with extended LD. NEAC × NCC crosses displayed almost complete repression of recombination within this block, but showed elevated recombination outside the block (Fig. 1b). NEAC and NCC linkage maps were used to order and orient scaffolds to create specific assemblies of LG1 for these two ecotypes of Atlantic cod. Alignment of these sequences revealed the presence of two adjacent inversions of 9.55 and 7.82 Mb (Fig. 1a). Additional evidence for two inversions, in contrast to a single inversion, was found in the LD pattern of 48 whole-genome sequenced NEAC samples (Fig. S2, Supporting information). High LD was found between polymorphisms at 18 and 28 Mb in the NCC version of the assembly. In contrast, the proposed NEAC orientation of the inversions rearranges these two regions to be located close together.

Geographical distribution of NEAC haplotype

To validate whether the NEAC haplotype precisely identified the differing genotypes of migratory and stationary cod ecotypes, we analysed 48 adult cod captured in the Barents Sea and representing pure NEAC based on *Pan* I genotyping. All samples were homozygous for the 186 SNPs within the haplotype block, which endorses its utility as a tool to classify cod as NEAC, NCC or crosses. To explore the distribution of the NEAC haplotype, we tested adult individuals from 14 different localities across the Northeast Atlantic Ocean. In sharp contrast to the fixation in the two locations in the Barents Sea, frequencies of the NEAC haplotype were low or nonexistent in the southern stocks and in the White Sea, while intermediate frequencies were found among samples collected along the Norwegian coast north from Bergen (Porsanger, Senja, Verrabotn, Borgund) (Fig. 2a). The NEAC haplotype was in HWE in all the stocks examined. These results contrasts with the hierarchical clustering analysis performed on SNPs in genomic locations outside regions with suspected inversions causing extended LD on LG1, LG2, LG7 and LG12 (Fig. 2b). The extremely long terminal branches and the absence of a clusters formed by the NEAC fish (Fig. 2b) suggest a complete lack of genetic structure between Barents Sea NEAC and the NCC stocks outside these chromosomal regions.

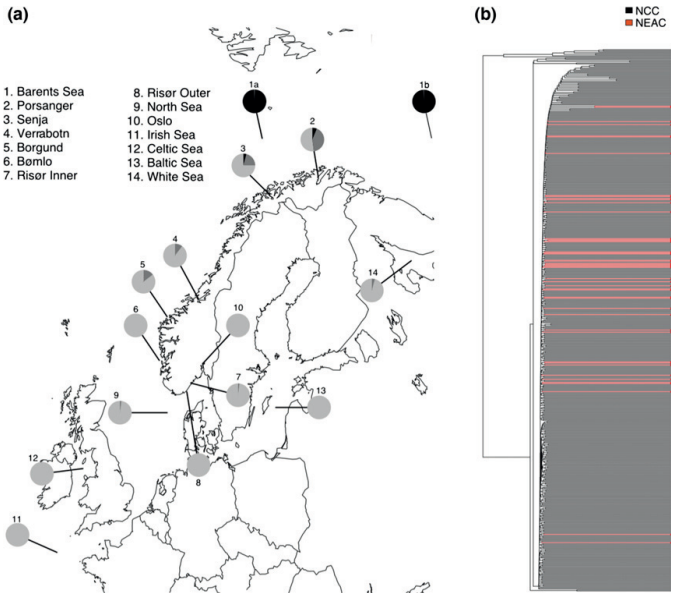

Fig. 2 Genomic divergence between NEAC and NCC. (a) Proportion of fish containing two (black), one (mid-grey) or no (light grey) copies of the NEAC haplotype in different Northeast Atlantic stocks. (b) Hierarchical clustering of SNP variation excluding genomic regions with suspected inversions due to large LD blocks (LG1, LG2, LG7, LG12). NEAC and NCC were represented by red and black tips, respectively. The genetic distance was calculated as identity by state across 7238 SNP loci.

Origin and age of inversions

To determine whether NEAC or NCC represents the ancestral state, we aligned LG1 sequences representing the arrangements of the inversions with Northern pike and stickleback to possibly identify conserved synteny blocks spanning breakage points defining the inversions. This analysis revealed a large block in pike spanning the break points flanking inversion 1, and a smaller block in stickleback bridging the two inversions (Fig. 3, Fig. S3, Supporting information). Taken together, these results suggest that NCC represents the ancestral state of the inverted structure.

The relative SNP density between NEAC and NCC across LG1 was calculated using whole-genome resequencing data from samples classified on the basis of the NEAC haplotype. Analysis of homozygous NEAC (n = 50), homozygous NCC (n = 11) or NEAC × NCC crosses (n = 43) revealed 540 685 SNPs with an average sequencing coverage of 17x. Relative heterozygosity expressed as the number of SNPs per 100 Kb in NEAC divided by the number in NCC revealed a dramatically reduced SNP density in NEAC samples within the LD block (Fig. 3). In contrast, the diversity outside the block was comparable for NEAC and NCC samples and to the rest of the genome. This result suggests a bottleneck event specific to the NEAC haplotype region on LG1 and supports the conclusion that NCC represents the ancestral state of the inverted structure.

The NEAC-NCC divergence relative to their divergence from Greenland cod were estimated to 0.57 and 0.13 within and outside the LG1 inversions, respectively. Assuming a divergence age of 3.5 million years between Greenland cod and Atlantic cod (Carr *et al.* 1999; Coulson *et al.* 2006), the inversion is estimated to be ~2 million years old (3.5*0.57 = 1.99). Although SNP data revealed no apparent genetic population structure between NEAC and NCC outside the inversion (Fig. 2a), we find 1553 FSD counts on LG1 outside the inversion. These divergent FSD sites are likely caused by a sample bias within NEAC and NCC fish because they represent a narrow genetic pool of interrelated individuals from a breeding programme rather than being a true random sample from both populations. Taking this background bias in FSD into account, the normalized Greenland cod–Atlantic cod divergence within the inversion would be ~1.6 million years (3.5*(0.57-0.13) = 1.57).

Candidate genes for adaptation to migratory behaviour

We annotated the LG1 sequence to search for genes involved in the adaptive divergence between the migratory and stationary ecotypes of Atlantic cod. The annotation resulted in the prediction of 1262 gene models for the whole chromosome, whereof 763 genes were located within the 17.4-Mb region containing the two inversions (357 and 406 genes, respectively). Variant detection within the same region revealed 19 206 SNPs that were fixed or very close to fixation for alternative alleles in NEAC and NCC and heterozygous in NEAC × NCC crosses, and included 849 plausible functional variants in 321 genes presenting good hits in the SWISSPROT database (Table S4, Supporting information). The corresponding protein variants displaying several amino acid substitutions included key enzymes in swim bladder function and haem synthesis together with important factors involved in muscle organization and behaviour (Fig. 3). Carbonic anhydrase catalyses the reversible conversion of carbon dioxide and water to bicarbonate and protons of importance for blood acidification and gas secretion into the swim bladder. The predicted NEAC and NCC variants of the secretory carbonic anhydrase isoform 6 (Ca6) differ at five positions, and the replacement of the highly conserved Gln196 with the novel His residue was shown by 3D modelling to reduce the interactions at the dimeric surface (Fig. 4, Table S5, Supporting information). We therefore predict reduced enzyme activity of the NCC variant as dimeric assembly of the enzyme confers an advantage for efficient CO_2 hydration in a variable extracellular milieu, such as the strong pH fluctuations in the gas gland (Pelster 2004; Pilka *et al.* 2012). We identified four additional genes within the double inversion involved in gas secretion by regulating glucose uptake and the production of acid metabolites. Glut1a facilitates glucose transport across cell membrane and is highly expressed in the gas gland cells of Atlantic cod (Hall *et al.* 2014). Two amino acid changes in the Glut1a protein and SNPs in the upstream region might have functional and regulatory effects. We also noted many SNPs in the genes coding for the three enzymes enolase 1 (Eno1), muscle-type phosphofructokinase (Pfkm) and glucose-6-phosphate dehydrogenase (G6pd) catalysing the anaerobic conversion of glucose to the acidic metabolites lactate and CO_2.

The inversions also contained candidate genes associated with the strenuous migratory behaviour, including two *alas* genes, and we found two amino acids changed in the predicted erythroid-specific Alas2 of crucial importance for haemoglobin production. Whereas the polymorphic *hb-1* gene is not located on LG1 (Borza *et al.* 2010), two amino acid changes in the rhesus type B glycoprotein (Rhbg) are probably responsible for the different blood type allele frequencies in NEAC and NCC (Møller 1966). The two populations also differ at four positions in the muscle protein leiomodin (Lmod3) essential for the organization of the sarcomeric thin filaments in skeletal muscle, and precise regulation of the

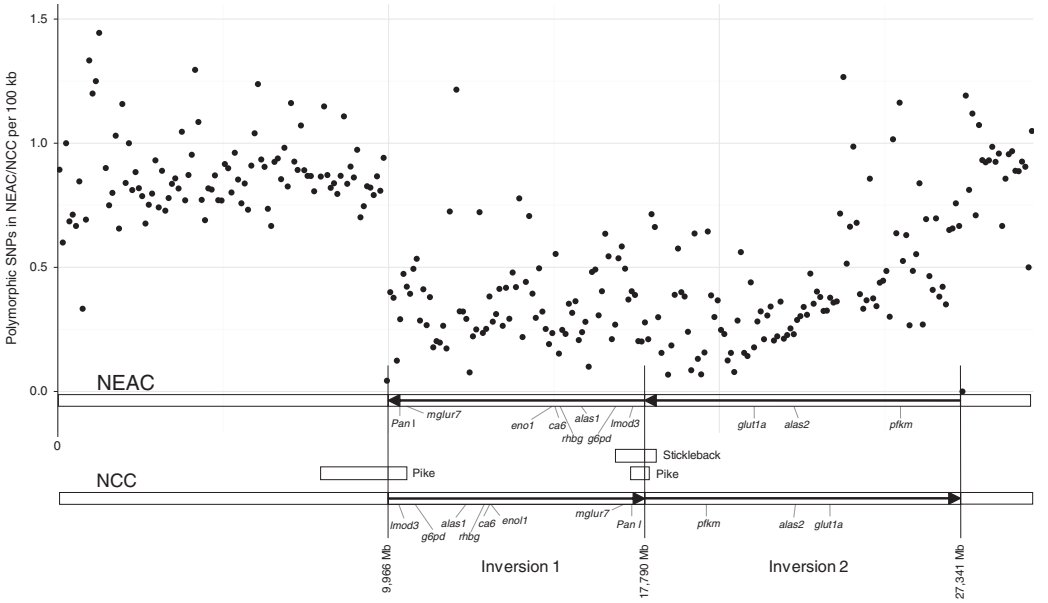

Fig. 3 Graphical representation of two adjacent inversions on LG1 present in NEAC and NCC. The upper part shows the relative difference in heterozygosity, measured as the number of polymorphisms per 100 kb in NEAC divided by the corresponding values in NCC. Conserved synteny blocks bridging inversion breakage points 1 and 2 suggest that NCC is holding the ancestral state of the inversions. Putative adaptive genes within the inversions are indicated.

filament length is crucial for optimal force generation during muscle contraction (Nworu *et al.* 2014; Yuen *et al.* 2014). Intriguingly, the double inversion contains the metabotropic glutamate receptor genes *mglur4* and *mglur7*, which are broadly expressed in the zebrafish brain, including olfactory bulb and retina (Haug *et al.* 2013). mGlur7 plays an important role in hippocampus-dependent spatial learning in mice (Goddyn *et al.* 2015), while the NEAC and NCC variants of the cod receptor were found to differ at three positions in a highly flexible region (data not shown).

Discussion

Population differentiation and adaptive evolution of Atlantic cod have been associated with four discrete islands of genomic divergence located on different chromosomes (Bradbury *et al.* 2010; Hemmer-Hansen *et al.* 2013; Karlsen *et al.* 2013). However, the fragmented nature of the current cod genome assembly (Star *et al.* 2011) has largely restricted our ability to identify genes associated with selection as well as our ability to reveal the mechanisms responsible for the observed patterns. To overcome these constraints, we constructed a dense linkage map and integrated it with draft genome assemblies to produce a cohesive chromosome sequence for Atlantic cod LG1. Separate linkage maps were constructed for pure NEAC, pure NCC and NEAC × NCC crosses in order to study differences in recombination patterns and potentially highlight rearrangements distinguishing the two ecotypes. These analyses revealed two adjacent inversions of 9.55 and 7.82 Mb (Fig. 1b), which clearly differentiate NEAC from NCC, as well as revealing a mechanism resulting in almost complete suppression of homologous recombination in individuals heterozygous for the inversions (Fig. 1a). Unlike the theoretic model of gene flow for simple inversions in *Drosophila* (Navarro *et al.* 1997), chromosomal polymorphisms involving more than one inversion prevent double crossovers and suppress recombination across the entire length of the rearrangements (Munte′ *et al.* 2005; Dyer *et al.* 2007; Huynh *et al.* 2011).

Chromosomal inversions have been associated with adaptive phenotypes in various plants and animals, including migratory species displaying high gene flow between the diverging populations (Rieseberg 2001; Hoffmann *et al.* 2004; Hoffmann & Rieseberg 2008). Polymorphic wing colour mimicry in butterflies is maintained by chromosomal rearrangements in the *Papilio* genus and in *Heliconius numata* (Joron *et al.* 2011;

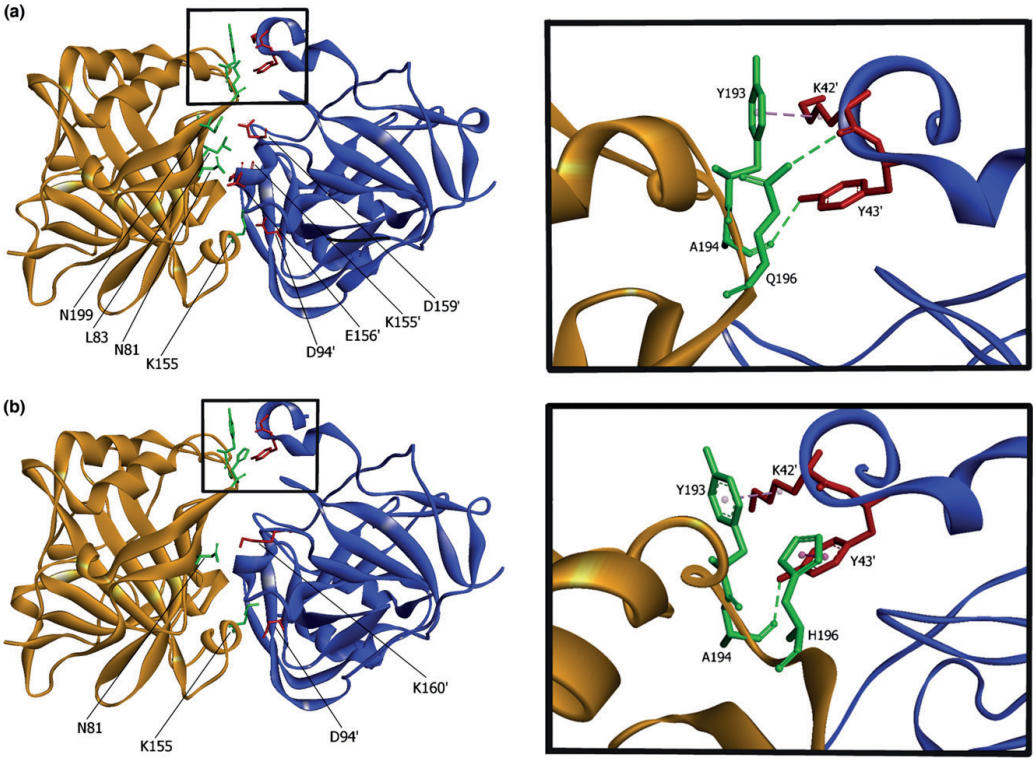

Fig. 4 Ribbon plots of the modelled carbonic anhydrase (Ca6) dimer interface in a) NEAC and b) NCC. The monomer subunits and key interacting residues (Table S5, Supporting information) are given in different colours. The enlarged sections show the dimeric interactions of Gln (Q) and His (H) at position 196.

Nishikawa *et al.* 2015), and at least two inversions spanning about 100 Mb in white-throated sparrow (*Zonotrichia albicollis*) were shown to harbour genes associated with territorial song and plumage (Thomas *et al.* 2008; Huynh *et al.* 2011; Zinzow-Kramer *et al.* 2015). Similarly, alternative reproductive strategies in the ruff (*Philomachus pugnax*) were recently found to be controlled by polymorphic genes affecting sex hormone levels and plumage within an inversion that occurred about 3.8 million years ago (Küpper *et al.* 2015; Lamichhaney *et al.* 2015). The repeated evolution of distinct marine and freshwater ecotypes of three-spined stickleback involves three chromosomal inversions, and alternative orientations of the voltage-gated potassium channel gene *kcnh4* might generate marine- and freshwater-specific isoforms (Jones *et al.* 2012). In rainbow trout (*Oncorhynchus mykiss*), different life history strategies of anadromous (steelhead) and resident ecotypes were recently shown to be associated with multiple loci with strong LD, suggesting the presence of an inversion suppressing recombination (Pearse *et al.* 2014).

The absence of genetic differentiation between NCC and NEAC populations outside the inversions on LG1 (Fig. 2b) supports previous conclusions of high levels of gene flow between migratory and stationary cod ecotypes in the North Atlantic Ocean (Hemmer-Hansen *et al.* 2013; Karlsen *et al.* 2013). However, the fact that the inversion is homozygous in NEAC, but polymorphic and under HWE in NCC populations, suggests that it is under strong directional selection in the migratory ecotype, while confers little or no fitness effects in the stationary ecotype. The absence of the NEAC haplotype in southern populations of NCC is interesting and raises the question whether this chromosomal rearrangement confers negative fitness effects in more southern marine ecosystems, for example due to suboptimal performance at higher water temperatures (Sarvas & Fevolden 2005). However, long-distance LD has been

reported between LG1 and LG2, LG7 and LG12 (Bradbury *et al.* 2013; Halldórsdóttir & Árnason 2016), making it possible that genes conferring the lack of fitness in warmer waters could be located elsewhere.

One of the obvious divergent adaptive challenges between NEAC and NCC populations is the adaptation to high hydrostatic pressure at large depths. The foraging and spawning migrations of NEAC involve vertical movements at depths of 200–400 m along stable thermal paths (Stensholt 2001), while stationary NCC fish exploit much shallower habitats (Hobson *et al.* 2007; Michalsen *et al.* 2014). This is supported by behavioural differences between juvenile NEAC and NCC settling at different depths, whereas the pelagic eggs have similar buoyancy (Løken *et al.* 1994; Fevolden *et al.* 2012; Jung *et al.* 2012). The swim bladder is a crucial organ by maintaining neutral buoyancy that allows fish to stay at their current depth without expending much energy swimming (Fänge 1953; Pelster 2004). Hence, impairment of the swim bladder function was assumed to significantly threaten the success of the spawning migration in the European eel (*Anguilla anguilla*) (Pelster 2014). Frequent descents and ascents lead to negative buoyancy, because gas secretion from the gas gland lags behind gas resorption in the swim bladder (Harden Jones & Scholes 1985; Godø & Michalsen 2000). This effect is amplified at greater depths, and the migratory NEAC should therefore benefit from enhanced gas secretion by increased blood acidification in the gas gland. The important role played by carbonic anhydrase in swim bladder function was demonstrated by inhibiting the enzyme activity in the gas gland that resulted in significantly reduced proton production and gas secretion (Fänge 1953; Skinazi 1953; Pelster 1995; Wurtz *et al.* 1999). While reduced carbonic anhydrase activity predicted for the NCC variant might not be critical for fish inhabiting shallow coastal water, the ability to maintain buoyancy is probably crucial for NEAC during frequent vertical movements to large depths. The energetic costs associated with the strenuous migrations may be further reduced by increased oxygen delivery and enhanced muscular capacity involving a suite of adaptive alleles identified within the inversions.

Gene flow between populations with divergent adaptive challenges can result in large fitness costs when recombination disrupts coinheritance of advantageous genetic variants. A genetic architecture that enforces strong LD between coselected gene variants would therefore be highly favourable under extensive gene flow from divergent populations. Such 'supergenes' have been shown to maintain population-specific adaptations in various organisms and are often caused by larger chromosome rearrangements (Joron *et al.* 2011; Thompson & Jiggins 2014; Twyford & Friedman 2015). We therefore hypothesize that the inversions on LG01 act as a supergene to efficiently maintain coinheritance of several highly favourable genetic variants, which over time have generated the island of genomic divergence observed between migratory and stationary ecotypes of cod.

Similar to NEAC and NCC, Icelandic migratory and stationary cod populations inhabiting different depths show genetic differentiation at the same genomic region as found in the Norwegian cod populations (Pampoulie *et al.* 2008, 2015; Grabowski *et al.* 2011). Additionally, offshore Atlantic cod from Davis Strait clustered with presumably NEAC sampled in the Barents Sea by SNP genotyping trans-Atlantic populations lacking NCC (Bradbury *et al.* 2013). Together, this suggests a common ancestry for the migratory cod populations and supports an old origin of the inversion polymorphism on LG1 associated with divergent behavioural adaptations in Atlantic cod. We estimated that the inversion arose about 1.6–2 million years ago during Pleistocene when glacial barriers and lowered sea level greatly influenced the abundance and distribution of marine species. This epoch probably represented the most important vicariance event in the evolution of Arctic fishes (Mecklenburg *et al.* 2011; Owens 2015). Atlantic cod survived in glacial refugia, but also moved southwards to ice-free regions during the glacial periods (Bigg *et al.* 2008; Kettle *et al.* 2011). We propose that beneficial alleles were captured within the two inversions that occurred in an isolated refugial population and later became fixed. During interglacial periods, local adapted individuals may have dispersed in the Arctic region and are today represented by the large migratory cod populations exploiting the high seasonal productivity in the most northerly environments on both sides of North Atlantic (Robichaud & Rose 2004). Adaptation to the polar environment might have been refined in the hybrid species walleye pollock (*Gadus chalcogrammus*) between Atlantic cod and polar cod (*Boreogadus saida*) that trespasses the ecology of its parents (Halldórsdóttir & Árnason 2016).

In conclusion, we reveal a major difference in the genomic architecture of the migratory NEAC and stationary NCC ecotypes by documenting two adjacent inversions spanning 17.4 Mb on LG1 that effectively block recombination in individuals heterozygous for the inversions. Despite clear signs of interbreeding, this lack of recombination has caused a supergene comprising adaptive alleles related to the migratory ecotype to be preserved without dilution from the stationary ecotype. The status of the two ecotypes differs substantially today that will probably increase in future. The NEAC population has increased greatly in recent years and the spawning stock in Barents Sea is historic high. In contrast, dramatic low levels in the NCC stocks and recruitment have resulted in restrictions on commercial cod fishing in several

coastal regions. As a consequence of increased water temperatures, Atlantic cod is expected to spread northwards and occupy larger areas of Barents Sea, while southern stocks will probably decline or disappear within year 2100 (Drinkwater 2005). Knowledge about the breeding structure of NEAC and NCC and the functional roles played by the differentiating genes might be important for the fisheries management and for predicting the response of the ecotypes to future climate change.

Acknowledgements

We are grateful to Kim Præbel for providing Greenland cod sample. Tom Cross, Phil McGinnity, Richard Fitzgerald, Halvor Knutsen and Roman Wenne are acknowledged for providing Atlantic cod population samples from the Irish Sea, Risør, North Sea and Baltic Sea. Genotypes for LG01 were generated as a part of the Cod SNP Consortium (CSC), a collaboration between CIGENE, CEES, IMR and Nofima. Mariann Arnyasi is acknowledged for genotyping and filtering of SNP data. Ole Kristian Tørresen and Alexander J. Nederbragt are acknowledged for providing draft genome assemblies for Atlantic cod. Funding was provided by the Norwegian University of Life Sciences to cover the salary of T.G.K. and from the Research Council of Norway (Project Nos.: 207680/E40 and 22734/O30).

References

Andersen O, Johnsen H, De Rosa MC *et al.* (2015) Evolutionary history and adaptive significance of the polymorphic *Pan* I in migratory and stationary populations of Atlantic cod (*Gadus morhua*). *Marine Genomics*, **22**, 45–54.

Árnason E, Pálsson S (1996) Mitochondrial cytochrome b DNA sequence variation of Atlantic cod *Gadus morhua*, from Norway. *Molecular Ecology*, **5**, 715–724.

Bangera R, Odegard J, Nielsen HM, Gjoen HM, Mortensen A (2013) Genetic analysis of vibriosis and viral nervous necrosis resistance in Atlantic cod (*Gadus morhua* L.) using a cure model. *Journal of Animal Science*, **91**, 3574–3582.

Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.

Bergstad OAJT, Dragesun O (1987) Life history and ecology of the gadoid resources of the Barents Sea. *Fisheries Research*, **5**, 119–161.

Bigg GR, Cunningham CW, Ottersen G *et al.* (2008) Ice-age survival of Atlantic cod: agreement between palaeoecology models and genetics. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **275**, 163–172.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Borza T, Higgins B, Simpson G, Bowman S (2010) Integrating the markers *Pan* I and Haemoglobin with the genetic linkage map of Atlantic cod (*Gadus morhua*). *Genetics*, **157**, 317–330.

Bradbury IR, Hubert S, Higgins B *et al.* (2010) Parallel adaptive evolution of Atlantic cod on both sides of the Atlantic Ocean in response to temperature. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **277**, 3725–3734.

Bradbury IR, Hubert S, Higgins B *et al.* (2013) Genomic islands of divergence and their consequences for the resolution of spatial structure in a exploited marine fish. *Evolutionary Applications*, **6**, 450–461.

Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, **81**, 1084–1097.

Carr SMKDS, Pepin P, Crutcher DC (1999) Molecular systematics of gadid fishes: implications for the biogeographic origins of Pacific species. *Canadian Journal of Zoology*, **77**, 19–26.

Christensen V, Guenette S, Heymans JJ *et al.* (2003) Hundred-year decline of North Atlantic predatory fishes. *Fish and Fisheries*, **4**, 1–24.

Cingolani P, Platts A, le Wang L *et al.* (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*, **6**, 80–92.

Coulson MW, Marshall HD, Pepin P, Carr SM (2006) Mitochondrial genomics of gadine fishes: implications for taxonomy and biogeographic origins from whole-genome data sets. *Genome*, **49**, 1115–1130.

Dobin A, Davis CA, Schlesinger F *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.

Drinkwater KF (2005) The response of Atlantic cod (*Gadus morhua*) to future climate change. *ICES Journal of Marine Science*, **62**, 1327–1337.

Dyer KA, Charlesworth B, Jaenike J (2007) Chromosome-wide linkage disequilibrium as a consequence of meiotic drive. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 1587–1592.

Fänge R (1953) The mechanisms of gas transport in the euphysoclist swimbladder. *Acta Physiologica Scandinavica. Supplementum*, **30**, 1–133.

Fevolden SE, Pogson GH (1997) Genetic divergence at the synaptophysin (Syp I) locus among Norwegian coastal and north-east Arctic populations of Atlantic cod. *Journal of Fish Biology*, **51**, 895–908.

Fevolden SE, Sarvas T (2001) *Distinct genetic divergence between cod (Gadus morhua) in fjords and cod in offshore waters in northern Norway*. ICES CM 2001/L:04.

Fevolden SE, Westgaard JI, Pedersen T, Praebel K (2012) Settling-depth vs. genotype and size vs. genotype correlations at the Pan I locus in 0-group Atlantic cod *Gadus morhua*. *Marine Ecology Progress Series*, **468**, 267–278.

Goddyn H, Callaerts-Vegh Z, D'Hooge R (2015) Functional dissociation of group III metabotropic glutamate receptors revealed by direct comparison between the behavioral profiles of knockout mouse lines. *International Journal of Neuropsychopharmacology*, **18**, pyv053, doi:10.1093/ijnp/pyv053.

Godø OR, Michalsen K (2000) Migratory behaviour of north-east Arctic cod, studied by use of data storage tags. *Fisheries Research*, **48**, 127–140.

Grabowski TB, Thorsteinsson V, McAdam BJ, Marteinsdottir G (2011) Evidence of segregated spawning in a single marine fish stock: sympatric divergence of ecotypes in Icelandic cod? *PLoS ONE*, **6**, e17528.

Green P, Falls K (1990) *Documentation for CRI-MAP version 2.4*. Washington University School of Medicine, St. Louis, Missouri.

Haas BJ, Papanicolaou A, Yassour M *et al.* (2013) *De novo* transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. *Nature Protocols*, **8**, 1494–1512.

Hall JR, Clow KA, Short CE, Driedzic WR (2014) Transcript levels of class I GLUTs within individual tissues and the direct relationship between GLUT1 expression and glucose metabolism in Atlantic cod (*Gadus morhua*). *Journal of Comparative Physiology B*, **184**, 483–496.

Halldórsdóttir K, Árnason E (2016) *Whole-genome sequencing uncovers cryptic and hybrid species among Atlantic and Pacific cod-fish. bioRxiv*. preprint first posted online Dec. 20, 2015; doi: http://dx.doi.org/10.1101/034926.

Harden Jones FR, Scholes P (1985) Gas secretion and resorption in the swimbladder of the cod *Gadus morhua*. *Journal of Comparative Physiology B*, **155**, 319–331.

Harris RS (2007) *Improved pairwise alignment of genomic DNA*. Ph.D. Thesis, The Pennsylvania State University.

Haug MF, Gesemann M, Mueller T, Neuhauss SC (2013) Phylogeny and expression divergence of metabotropic glutamate receptor genes in the brain of zebrafish (*Danio rerio*). *The Journal of Comparative Neurology*, **521**, 1533–1560.

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

Hemmer-Hansen J, Nielsen EE, Therkildsen NO *et al.* (2013) A genomic island linked to ecotype divergence in Atlantic cod. *Molecular Ecology*, **22**, 2653–2667.

Hobson VJ, Righton D, Metcalfe JD, Hays GC (2007) Vertical movements of North Sea cod. *Marine Ecology Progress Series*, **347**, 101–110.

Hoffmann AA, Rieseberg LH (2008) Revisiting the impact of inversions in evolution: from population genetic markers to drivers of adaptive shifts and speciation? *Annual Review of Ecology, Evolution, and Systematics*, **39**, 21–42.

Hoffmann AA, Sgrò CM, Weeks AR (2004) Chromosomal inversion polymorphism and adaptation. *Trends in Ecology and Evolution*, **19**, 482–488.

Hutchings JA, Bishop TD, McGregor-Shaw CR (1999) Spawning behaviour of Atlantic cod, *Gadus morhua*: evidence of mate competition and mate choice in a broadcast spawner. *Canadian Journal of Fisheries and Aquatic Sciences*, **56**, 97–107.

Huynh LY, Maney DL, Thomas JW (2011) Chromosome-wide linkage disequilibrium caused by an inversion polymorphism in the white-throated sparrow (*Zonotrichia albicollis*). *Heredity*, **106**, 537–546.

Jakobsen T (1987) Coastal cod in northern Norway. *Fisheries Research*, **5**, 223–234.

Johnsen H, Andersen Ø (2012) Sex dimorphic expression of five dmrt genes identified in the Atlantic cod genome. The fish-specific *dmrt2b* diverged from *dmrt2a* before the fish whole-genome duplication. *Gene*, **505**, 221–232.

Jones FC, Grabherr MG, Chan YF *et al.* (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, **484**, 55–61.

Joron M, Frezal L, Jones RT *et al.* (2011) Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature*, **477**, 203–206.

Jung KM, Folkvord A, Kjesbu OS *et al.* (2012) Egg buoyancy variability in local populations of Atlantic cod (*Gadus morhua*). *Marine Biology*, **159**, 1969–1980.

Karlsen BO, Klingan K, Emblem A *et al.* (2013) Genomic divergence between the migratory and stationary ecotypes of Atlantic cod. *Molecular Ecology*, **22**, 5098–5111.

Karlsen BO, Emblem A, Jorgensen TE *et al.* (2014) Mitogenome sequence variation in migratory and stationary ecotypes of North-east Atlantic cod. *Marine Genomics*, **15**, 103–108.

Kettle AJ, Morales Muñiz A, Roselló-Izquierdo E, Heinrich D, Vollestad A (2011) Refugia of marine fish in the northeast Atlantic during the last glacial maximum: concordant assessment from archaeozoology and paleotemperature reconstructions. *Climate of the Past*, **7**, 181–201.

Küpper C, Stocks M, Risse JE (2015) A supergene determines highly divergent male reproductive morphs in the ruff. *Nature Genetics*, **48**, 79–83.

Lamichhaney C, Fan G, Widemo F *et al.* (2015) Structural genomic changes underlie alternative reproductive strategies in the ruff (*Philomachus pugnax*). *Nature Genetics*, **48**, 84–88.

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

Laskowski RA, Moss DS, Thornton JM (1993) Main-chain bond lengths and bond angles in protein structures. *Journal of Molecular Biology*, **231**, 1049–1067.

Løken S, Pedersen T (1996) Effect of parent type and temperature on vertebrae number in juvenile cod, *Gadus morhua* (L), in Northern Norway. *Sarsia*, **80**, 293–298.

Løken S, Pedersen T, Berg E (1994) Vertebrae numbers as an indicator for the recruitment mechanism of coastal cod of northern Norway. *Cod and Climate Change – Proceedings of a Symposium*, **198**, 510–519.

Mackenzie BR, Almesjo L, Hansson S (2004) Fish, fishing, and pollutant reduction in the Baltic Sea. *Environmental Science & Technology*, **38**, 1970–1976.

Martinez PA, Zuranoa JP, Amadoa TF *et al.* (2015) Chromosomal diversity in tropical reef fishes is related to body size and depth range. *Molecular Phylogenetics and Evolution*, **93**, 1–4.

McKenna A, Hanna M, Banks E *et al.* (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.

Mecklenburg N, Garcia-Lopez R, Puelles E, Sotelo C, Martinez S (2011) Cerebellar oligodendroglial cells have a mesencephalic origin. *Glia*, **59**, 1946–1957.

Michalsen K, Johansen T, Subbey S, Beck A (2014) Linking tagging technology and molecular genetics to gain insight in the spatial dynamics of two stocks of cod in Northeast Atlantic waters. *ICES Journal of Marine Science*, **71**, 1417–1432.

Møller D (1966) Genetic differences between cod groups in the Lofoten area. *Nature*, **212**, 824

Møller D (1968) Genetic diversity in spawning cod along the Norwegian coast. *Hereditas*, **60**, 1–32.

Møller D (1969) The relationship between arctic and coastal cod in their immature stages illustrated by frequencies of genetic characters. *Fiskeridirektoratets Skrifter. Serie Havundersøkelser*, **15**, 220–223.

Monteiro CA, Serrão EA, Pearson GA (2012) Prezygotic barriers to hybridization in marine broadcast spawners: reproductive timing and mating system variation. *PLoS ONE*, **7**, e35978.

Mork J, Sundnes G (1985) 0-group cod (*Gadus morhua*) in captivity: different survival of certain genotypes. *Helgoländer Meeresuntersuchungen*, **39**, 63–70.

Munte′ A, Rozas J, Aguade M, Segarra C (2005) Chromosomal inversion polymorphism leads to extensive genetic structure: a multilocus survey in *Drosophila subobscura*. *Genetics*, **169**, 1573–1581.

Myers RA, Hutchings JA, Barrowman NJ (1997) Why do fish stocks collapse? The example of cod in Atlantic Canada. *Ecological Applications*, **7**, 91–106.

Navarro A, Betran E, Barbadilla A, Ruiz A (1997) Recombination and gene flux caused by gene conversion and crossing over in inversion heterokaryotypes. *Genetics*, **146**, 695–709.

Nielsen EE, Hemmer-Hansen J, Poulsen NA *et al*. (2009) Genomic signatures of local directional selection in a high gene flow marine organism; the Atlantic cod (*Gadus morhua*). *BMC Evolutionary Biology*, **9**, 276.

Nishikawa H, Iijima T, Kajitani R *et al*. (2015) A genetic mechanism for female-limited Batesian mimicry in *Papilio* butterfly. *Nature Genetics*, **47**, 405–409.

Nordeide JT, Folstad I (2000) Is cod lekking or a promiscuous group spawner? *Fish and Fisheries*, **1**, 90–93.

Nordeide JT, Pettersen IH (1998) Haemoglobin frequencies and vertebral numbers of cod (*Gadus morhua* L.) off northern Norway – test of a population structure hypothesis. *ICES Journal of Marine Science*, **55**, 134–140.

Nordeide J, Johansen S, Jørgensen T, Karlsen BMoum T (2011) Population connectivity among migratory and stationary cod *Gadus morhua* in the Northeast Atlantic – a review of 80 years of study. *Marine Ecology Progress Series*, **435**, 269–283.

Nworu C, Kraft R, Schnurr D, Gregorio CKrieg P (2014) Leiomodin 3 and tropomodulin 4 have overlapping functions during skeletal myofibrillogenesis. *Journal of Cell Science*, **128**, 239–250.

Ottersen G, Bogstad B, Yaragina N *et al*. (2014) A review of early life history dynamics of Barents Sea cod (*Gadus morhua*). *ICES Journal of Marine Science*, **71**, 2064–2087.

Owens HL (2015) Evolution of codfishes (Teleostei: *Gadinae*) in geographical and ecological space: evidence that physiological limits drove diversification of subarctic fishes. *Journal of Biogeography*, **42**, 1091–1102.

Pálsson OK, Thorsteinsson V (2003) Migration patterns, ambient temperature, and growth of Icelandic cod (*Gadus morhua*): evidence from storage tag data. *Canadian Journal of Fisheries and Aquatic Sciences*, **60**, 1409–1423.

Palumbi SR (1994) Genetic divergence, reproductive isolation, and marine speciation. *Annual Review of Ecology and Systematics*, **25**, 547–572.

Pampoulie C, Jakobsdottir KB, Marteinsdottir G, Thorsteinsson V (2008) Are vertical behaviour patterns related to the pantophysin locus in the Atlantic cod (*Gadus morhua* L.)? *Behavior Genetics*, **38**, 76–81.

Pampoulie C, Skirnisdottir S, Star B *et al*. (2015) Rhodopsin gene polymorphism associated with divergent light environments in Atlantic cod. *Behavior Genetics*, **45**, 236–244.

Pearse DE, Miller MR, Abadia-Cardoso A, Garza JC (2014) Rapid parallel evolution of standing variation in a single, complex, genomic region is associated with life history in steelhead/rainbow trout. *Proceedings of the Royal Society B*, **281**, 20140012.

Pelster B (1995) Mechanisms of acid release in isolated gas gland cells of the European eel Anguilla anguilla. *American Journal of Physiology*, **269**, R793–R799.

Pelster B (2004) pH regulation and swimbladder function in fish. *Respiratory Physiology & Neurobiology*, **144**, 179–190.

Pelster B (2014) Swimbladder function and the spawning migration of the European eel *Anguilla anguilla*. *Frontiers in Physiology*, **5**, 486.

Pilka ES, Kochan G, Oppermann U, Yue WW (2012) Crystal structure of the secretory isozyme of mammalian carbonic anhydrases CA VI: implications for biological assembly and inhibitor development. *Biochemical and Biophysical Research Communications*, **419**, 485–489.

Pogson GH (2001) Nucleotide polymorphism and natural selection at the pantophysin (*Pan* I) locus in the Atlantic cod, *Gadus morhua* (L.). *Genetics*, **157**, 317–330.

Rieseberg L (2001) Chromosomal rearrangements and speciation. *Trends in Ecology and Evolution*, **16**, 351–358.

Robichaud D, Rose G (2004) Migratory behaviour and range in Atlantic cod: inference from a century of tagging. *Fish and Fisheries*, **5**, 185–214.

Rollefsen G (1933) The otoliths of the cod. *Fiskeridirektoratets Skrifter. Serie Havundersøkelser*, **4**, 1–4.

Rollefsen G (1954) Observations on cod and cod fisheries of Lofoten. *Rapp. P.-v Rapports et procès-verbaux des réunions/ Conseil permanent international pour l'exploration de la mer*, **136**, 40–47.

Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, **234**, 779–815.

Sarvas TH, Fevolden SE (2005) Pantophysin (*Pan* I) locus divergence between inshore v. offshore and northern v. southern populations of Atlantic cod in the north-east Atlantic. *Journal of Fish Biology*, **67**, 444–469.

Shen KN, Chang CW, Durand JD (2015) Spawning segregation and philopatry are major prezygotic barriers in sympatric cryptic *Mugil cephalus* species. *Comptes Rendus Biologies*, **338**, 803–811.

Skinazi L (1953) Carbonic anhydrase in two closely related teleosts; inhibition of the secretion of gas from the air bladder of perch by sulfonamides. *Comptes Rendus des Seances de la Societe de Biologie et de ses Filiales*, **147**, 295–299.

Star B, Nederbragt AJ, Jentoft S *et al*. (2011) The genome sequence of Atlantic cod reveals a unique immune system. *Nature*, **477**, 207–210.

Stensholt BK (2001) Cod migration patterns in relation to temperature: analysis of storage tag data. *ICES Journal of Marine Science*, **58**, 770–793.

Sundby S, Nakken O (2008) Spatial shifts in spawning habitats of Arcto-Norwegian cod related to multidecadal climate oscillations and climate change. *ICES Journal of Marine Science*, **65**, 953–962.

Thomas DW, Watanabe CK (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–1875.

Thomas JW, Caceres M, Lowman JJ *et al*. (2008) The chromosomal polymorphism linked to variation in social behavior in the white-throated sparrow (*Zonotrichia albicollis*) is a complex rearrangement and suppressor of recombination. *Genetics*, **179**, 1455–1468.

Thompson MJ, Jiggins CD (2014) Supergenes and their role in evolution. *Heredity*, **113**, 1–8.

Trapnell C, Williams BA, Pertea G *et al.* (2010) Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. *Nature Biotechnology*, **28**, 511–515.

Twyford A, Friedman J (2015) Adaptive divergence in the monkey flower Mimulus guttatus is maintained by a chromosomal inversion. *Evolution*, **69**, 1476–1486.

Westgaard JI, Fevolden SE (2007) Atlantic cod (*Gadus morhua* L.) in inner and outer coastal zones of northern Norway display divergent genetic signature at non-neutral loci. *Fisheries Research*, **85**, 306–315.

Williams GC (1975) *Sex and Evolution*. Princeton University Press, Princeton, New Jersey.

Wurtz J, Salvenmoser W, Pelster B (1999) Localization of carbonic anhydrase in swimbladder of European eel (*Anguilla anguilla*) and perch (*Perca fluviatilis*). *Acta Physiologica Scandinavica*, **165**, 219–224.

Yuen M, Sandaradura SA, Dowling JJ *et al.* (2014) Leiomodin-3 dysfunction results in thin filament disorganization and nemaline myopathy. *Journal of Clinical Investigations*, **124**, 4693–4708.

Zheng X, Levine D, Shen J *et al.* (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, **28**, 3326–3328.

Zinzow-Kramer WM, Horton BM, McKee CD *et al.* (2015) Genes located in a chromosomal inversion are correlated with territorial song in white-throated sparrows. *Genes, Brain and Behavior*, **14**, 641–654.

S.L. and Ø.A. designed the study with input from T.G.K. H.G and M.P.K. T.G.K, S.L., H.G., S.R.S and T.N. analysed the data. Ø.A., T.G.K. and S.L. examined candidate genes. T.J., H.O. and M.B. provided samples from family material and wild populations. M.B. and A.S. provided sequence data from the National cod breeding programme. M.C.D.R and B.R. modelled the protein variants. Ø.A., T.G.K., H.G, S.R.S., M.P.K. and S.L. wrote the manuscript with contributions from all authors.

Data accessibility

Resequencing data from 104 farmed cod were submitted to the European Nucleotide Archive and can be retrieved under Accession no. PRJEB12803. Genotype data from SNP array are available in Dryad (http://datadryad.org/). All SNPs are referred to by their ss# or rs# available in dbSNP (http://www.ncbi.nlm.nih.gov/SNP/).

Supporting information

Additional supporting information may be found in the online version of this article.

Table S1 Linkage maps generated from 96 families of farmed cod. Maps and map distances were calculated separately for NEAC, NCC and NEAC × NCC crosses, and also split between males and females.

Table S2 Predicted function of open reading frames were found with TransDecoder and homology search using BLASTP against zebrafish and stickleback protein databases.

Table S3 Reads from a Greenland cod and a NCC where aligned to the NEAC reference. Fixed sequence differences were counted for LG1, both inside the inversions (top right) and outside (bottom left).

Table S4 Genes with nonsynonymous SNPs within the inversions. Number of individuals with reference or alternative alleles from NEAC, NCC and NEAC × NCC cross are indicated together with GI number and accession IDs.

Table S5 Interdimeric contacts in carbonic anhydrase (Ca6) of NEAC (Gln196) and NCC (His196). Protein contacts (within 4.5 Å) in the interfaces between A- and B-monomers of the homology models are reported.

Fig. S1 The double inversion was dated by comparing estimates of sequence divergence between NEAC-NCC within the inversion with sequence divergence between Greenland cod and Atlantic Cod under the assumption that this latter species divergence occurred 3.5 million years ago (Carr *et al.* 1999; Coulson *et al.* 2006). The level of sequence divergence was measured in units of fixed sequence differences (FSD). The ratio between NEAC-NCC FSD distance (red line) and the mean FSD distance between Greenland cod-NEAC (blue line) and Greenland cod-NCC (green line) was calculated and then multiplied with the Greenland cod- Atlantic Cod divergence age (3.5 Mya) to get the absolute age of NEAC-NCC divergence.

Fig. S2 Pairwise LD for 48 NEAC, measured as r^2, between all SNPs (MAF>0.1) detected by resequencing on LG1. Left figure is the NCC map, while right figure is the NEAC map. Only values above $r^2 = 0.7$ are shown. Circle indicates a region within the second inversion being in high LD with a region at the end of the chromosome. The NEAC map minimizes the distance between these two regions.

Fig. S3 Comparative map using whole chromosome alignment between the NCC version of LG1 and stickleback LGXIII (a) and pike LG12 (b).

Paper II

A nanopore based chromosome-level assembly representing Atlantic cod from the Celtic Sea.

Tina Graceline Kirubakaran[1], Øivind Andersen[1,2], Michel Moser[1], Mariann Arnyasi[1], Philip McGinnity[3], Sigbjørn Lien[1], Matthew Kent[1].

[1]Centre for Integrative Genetics and Department of Animal and Aquacultural Sciences, Faculty of Biosciences, Norwegian University of Life Sciences, Ås, Norway.
[2]Nofima, Ås, Norway.
[3]School of Biological, Earth & Environmental Sciences University College Cork, Cork, Ireland

ABSTRACT

Background

Currently available genome assemblies for Atlantic cod (*Gadus morhua*) have been constructed using DNA extracted from fish belonging to the Northeast Arctic Cod (NEAC) population; a migratory population feeding in the cold Barents Sea. These genome assemblies have been crucial for the development of genetic markers that have been used together with whole genome resequencing data to study population differentiation and adaptive evolution in Atlantic cod, pinpointing four discrete islands of genomic divergence located on linkage groups (LGs) 1, 2, 7 and 12.

Results

To generate a resource representing a contrast to the current assemblies from the northern NEAC population, we have produced a high-quality reference genome from a male Atlantic cod representing the southernmost population inhabiting the Celtic sea. Structurally, the genome assembly (gadMor_Celtic) was produced exclusively from long-read nanopore data and has a combined contig size of 686 Mb with a N50 of 10 Mb. Integration of the contigs with genetic linkage mapping information enabled the construction of 23 chromosome sequences adding up to 643.4 Mb. The gadMor_Celtic chromosome sequences map with high confidence to the latest assembly generated from the NEAC population (gadMor3) and allowed us to characterize in detail the large (4.5 – 17.4 Mb) chromosomal inversions underlying the supergenes on LGs 1, 2, 7 and 12. In most cases, inversion breakpoints could be located within single nanopore contigs. Our results also suggest the presence of inversions on LGs 6, 11 and 21, although these remain to be confirmed. Further, we identified a repetitive

element highly enriched at predicted centromeric regions with low recombination that was to reliably categorize four metacentric Atlantic cod chromosomes.

Conclusions
Our gadMor_Celtic assembly provides a chromosome-level resource representing a southern cod population which is complementary to the existing northern population based genome assemblies and represents the first step towards developing pan-genomic resources for Atlantic cod.

Keywords
Atlantic cod, genome assembly, nanopore, chromosomal rearrangements, linkage map, centromere repeats.

Background
Atlantic cod (*Gadus morhua*) is a commercially exploited high-fecundity fish with a wide geographical distribution extending across the North Atlantic Ocean and including nearly freezing temperatures in the Arctic to variable high temperatures in the southern extremities of its Eastern Atlantic distribution [1-3]. It has been proposed that increases in water temperatures will see Atlantic cod spread northwards and occupy larger areas of Barents Sea, while southern populations will decline and possibly disappear [3, 4]. Characterizing the genomic diversity among fish populations, and understanding its relationship to phenotypic variation has therefore become increasingly important in fisheries management and for predicting the response of various ecotypes to environmental fluctuations, such as climatic changes [2, 5]. Earlier studies in Atlantic cod have provided evidence for elevated genomic divergence between populations mainly in four discrete regions, also referred as supergenes, located on linkage groups (LGs) 1, 2, 7 and 12 [6-21]. Relationships between these regions and environmental conditions has documented that the region on LG01 is associated with strong genetic differentiation between migratory and stationary ecotypes on both sides of the Atlantic Ocean [8-10, 19, 21, 22]. The supergene coincides with a double inversion that suppresses homologous recombination in heterozygotes and effectively prevents admixing between co-segregating haplotypes [8]. The genomic islands of divergence on LGs 2, 7 and 12 are also found on both sides of the Atlantic Ocean associated with mean ocean temperatures along the north-south gradient [11-13, 16]. Genomic divergence in these regions have also been associated with other environmental factors in studies concerning Baltic and North Sea populations (Berg *et al.* 2015), as well as oceanic and coastal populations in the North Sea [14]. Elevated linkage disequilibrium (LD) detected across the regions on LGs 2, 7, and 12 suggests that they have arisen as a result of chromosomal inversions, but high-resolution

sequence data showing this and describing the precise locations, sizes and genomic structure underlying these regions has so far been lacking.

Most present fish genome sequences are built from short-read Illumina data, which is a computationally challenging and error prone process especially when the genomes contain extensive repetitive regions. Long-read sequencing technologies provide the means to directly read through repetitive elements and thereby potentially produce much more complete *de novo* assemblies. The recently released gadMor3 assembly (NCBI accession ID: GCF_902167405.1) was developed based on long-read sequence data produced from a NEAC fish and represent a significant improvement over the former gadMor1 and gadMor2 assemblies generated from the same northern population [23, 24]. In this paper, we used long-read nanopore data to construct a reference genome assembly for a male Atlantic cod from the southern population of the Celtic Sea and integrated the assembly with linkage data to build high-quality chromosomes sequences. The genome sequence was utilized to detect a potential centromeric repeat sequence differentiating chromosomal morphology and to characterise with high precision the chromosomal rearrangements underlying the notable supergenes on LGs 1, 2, 7 and 12.

Results and Discussion

Genome assembly

Current paradigm in population genomics is to build genomic tools and interpret results based on the information embodied by one arbitrary reference genome, constructed from a single individual and used to represent the whole species. In accord with this, genome assemblies for Atlantic cod have been generated from NEAC fish, which is a northern migratory population feeding in the cold Barents Sea. However, with the advent of new, cheaper sequencing platforms and long-read technology it is timely to develop multiple reference genome sequences representing a species. To produce a resource representing a contrast to NEAC, we decided to generate a high-quality reference genome from a male Atlantic cod captured in the Celtic sea, a region representing the southernmost extreme of the Eastern Atlantic distribution [3, 25] and where cod experience suboptimal summer temperatures [5]. Our gadMor_Celtic assembly was built in a stepwise process involving; (i) testing multiple combinations of assembly parameters to generate initial assemblies using wtdgb2 [26], (ii) merging of contigs from selected initial assemblies into a primary assembly using quickmerge [27], (iii) multiple rounds of base error correction using Racon [28] and Pilon [29], and finally (iv) anchoring and orientation of polished contigs into linkage groups.

The two 'best' initial assemblies (see Materials and Methods for details), were similar in regards to their total size (bp), number of contigs, and contig N50 (see Table 1), and their Benchmarking Universal Single-Copy Orthologs (BUSCO)

scores of 20-40% indicating a poor content of identifiable reference genes. This last observation likely reflects the fact that they are constructed from nanopore reads alone (which suffer from relatively high rates of substitution and deletion errors; e.g. 13% and 5% respectively [30]) and that the assemblies have not been corrected with higher quality reads such as those from Illumina sequencing [31]. To improve assembly contiguity, contigs showing a sequence overlap of more than 5kb with >95% similarities were combined using quickmerge. This increased the contig N50 from 6 to 10.4Mb and concurrently reduced number of contigs. Thereafter, two rounds of error correction were performed. First round used Racon to generate consensus sequences using the 70X nanopore data alone and resulted in a BUSCO score of 66.5%. Second round used Pilon and 50X coverage high-quality Illumina data (16.5Mb paired-end 250 bp reads) and saw the BUSCO genome completeness score increase to 94.2% which is comparable to other high quality fish genomes (e.g. [32, 33]. The resulting gadMor_Celtic assembly is composed of 1253 contigs (contig N50=10.5 Mb, average contig length 0.55 Mb) and includes 686 Mb of sequence.

	Total size (bp)	**Total number of contigs**	**Contig N50 (bp)**	**BUSCO score (%)**
Wtdgb2 assembly 1	668,357,526	1600	6,012,173	23.2
Wtdgb2 assembly 2	670,278,278	1666	6,004,590	42.4
Quickmerge contigs	677,547,349	1253	10,448,158	Not done
Racon polishing	683,672,734	1253	10,518,163	66.5
Pilon polishing	**685,982,295**	**1253**	**10,559,872**	**94.2**

Table 1. Assembly statistics*. Metrics describing genome statistics of the initial assemblies, the quickmerge assembly, and the final gadMor_Celtic assembly after polishing with nanopore (Racon) and Illumina (Pilon) data.*

High-quality linkage maps of densely spaced markers provide the means to reliably anchor genomic fragments (contigs and scaffolds) to chromosomes. If constructed in a large pedigree, and with an adequate number of markers, it may also serve as the backbone for ordering, orienting and concatenating the fragments into chromosome sequences. However, the ability to order and orientate fragments is constrained by the frequency and location of recombination events and thus is limited by the resolution of the map. In this study we used a genetic map consisting of 9,178 SNPs (Additional file 1), constructed in a large pedigree of 2,951 individuals to order and orientate 149 contigs (totalling 643.4 Mb; 93% of assembly) into 23 chromosome sequences. The average number of SNPs per contig was 56.1, with only 12 contigs containing fewer than five SNPs. The high contiguity of the gadMor_Celtic assembly is evidenced by the fact that for one linkage group (LG14), the entire genetic map was correctly captured by a single contig of more than 30 Mb. The total length of the female linkage map (1,662.7 cM) was approximately 1.3 times larger than the male map (1,262.3 cM). Notably, the linkage maps were

constructed in a family material segregating the large inversions on LGs 1, 2, 7 and 12, which introduce pronounced gaps in the linkage maps at the boarders of these inversions (see Additional file 2).

Chromosomal inversions

The detection of extended blocks of LD between SNPs has been used in several studies to define the regions of genetic differentiation on Atlantic cod LGs 1 2, 7 and 12 [6, 11, 13, 14]. Large chromosomal inversions have been hypothesized for all four regions but only documented for LG01 (Kirubakaran *et al.* 2016). While regions of extended LD are symptomatic of large polymorphic inversions, no studies have directly compared reference genomes from different cod ecotypes to define and confirm the underlying mechanism, locate the genomic regions containing the inversion breakpoints and define the exact complement of genes they contain. We aligned the recently released gadMor3 assembly (NCBI accession ID: GCF_902167405.1) constructed from a NEAC individual to our gadMor_Celtic assembly using LASTZ [34]. The gadMor3 assembly was generated following a comprehensive sequencing effort combining long-read sequence data from Pacific BioSciences with various datasets for scaffolding and polishing, and resulted in 1442 contigs (contig N50=1.015 Mb). Despite being an order of magnitude smaller than our gadMor_Celtic contigs, the gadMor3 scaffolds nevertheless mapped with a high confidence to the assembly and documented that the two assemblies display alternative configurations of rearrangements for the supergenes on LGs 1, 2, 7 and 12. In most cases, the inversion breakpoints could be described at high resolution because they locate within single nanopore contigs. Exceptions to this were the third breakpoint of LG01 and second breakpoint on LG07 which falls between two gadMor_Celtic contigs (Fig. 1).

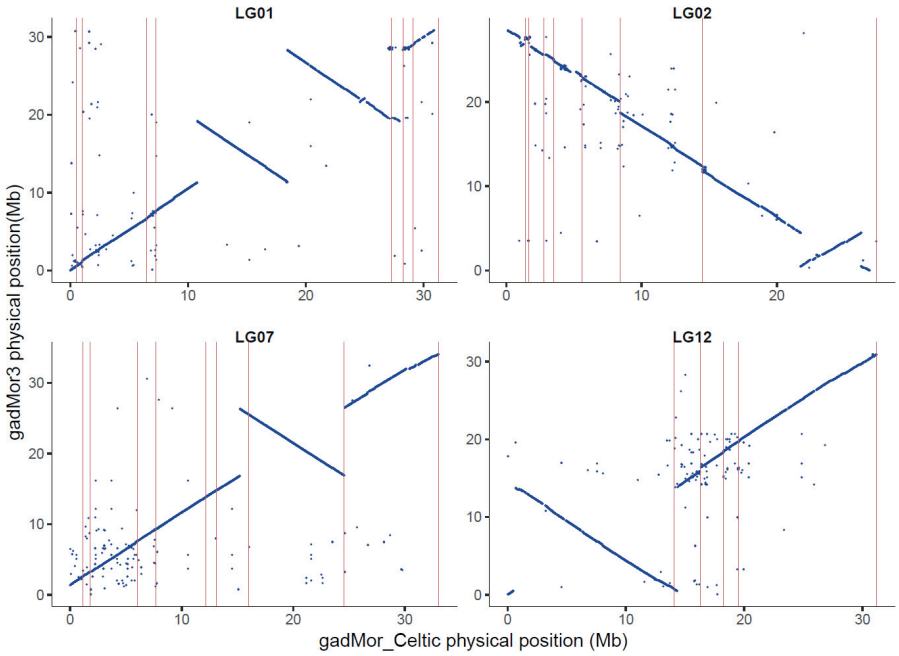

Fig. 1. *Alignment of gadMor_Celtic (x-axis) and gadMor3 (y-axis) chromosome sequences for linkage groups 1, 2, 7 and 12. Vertical lines (pink) demarcate boundaries of gadMor_Celtic contigs.*

A perfect characterization of inversion breakpoints at the sequence level using the gadMor3 and gadMor_Celtic assemblies demands that contigs from both assemblies span the breakpoints and that sequences at the breakpoints align perfectly with high confidence. As contig structure is not available in the public gadMor3 assembly, and genome alignments to some extend were confounded by repetitive sequences, we find it correct to present the inversion breakpoints as regions, or putative intervals (see Table 2).

Linkage Group	Putative interval containing breakpoint		Size (bp)	Inversion size (Mb)
	Start	End		
LG01	10,782,691	10,787,755	5,064	17.45
	18,422,802	18,425,099	2,297	
	28,225,372	28,228,130	2,758	
LG02	21,733,338	21,733,998	660	4.51
	26,233,253	26,238,098	4,840	
LG07	15,208,043	15,210,043	2,000	9.37
	24,574,346	24,575,510	1,164	
LG12	493,527	635,659	142,132	13.88
	14,330,965	14,376,973	46,008	

Table 2. Genomic regions likely containing the inversion breakpoints*. A pairwise comparison between gadMor_Celtic and gadMor3 reveals the interval*

(described as a start and stop coordinates relative to the gadMor_Celtic assembly) for each inversion breakpoint LGs 1, 2, 7, and 12.

In gadMor_Celtic the double inversion on LG01 spans a total interval of 17.45 Mb which is slightly larger than our previous estimate of 17.37 Mb (Kirubakaran et al., 2016). Our ability to detect the inversions when comparing to the gadMor3 NEAC reference suggests that Celtic cod possess the stationary (as opposed to migratory) ecotype chromosome configuration. An earlier survey of Celtic cod [25] showed that while a portion of the population migrate horizontally (from the Celtic sea to the Western English channel) they do not perform vertical migration to depths of 500 meters as has been reported for NEAC fishes [35], but instead remain at about 100 meters similar to those of the stationary populations around the Norwegian coast [36].

The inversions on LGs 2, 7 and 12 span 4.51, 9.37 and 13.88 Mb, respectively. These sizes are in relatively close agreement to earlier estimations of 5.0, 9.5, and 13 Mb, which were calculated from LD analyses and detection of regions of elevated divergence between populations [14]. In their analysis, Sodeland *et al*. (2016) used the highly fragmented gadMor1 assembly [24] and a relatively sparse set of 9,187 SNPs to define the regions, both factors that may explain the physical difference between estimates. A more recent study investigated cod populations from the Northwest Atlantic and measured LD amongst almost 3.4M SNPs detected from resequencing data, the LGs 2, 7 and 12 inversions were estimated to be 5.6, 9.3, and 11.6 Mb respectively [6]. While not identical, these regions and sizes detected in fish from both sides of the Atlantic are remarkably consistent, supporting the hypothesis that these cod have a common ancestral origin [7, 22].

Our analyses suggest the presence of putative inversions on LGs 6, 11 and 21 (see Fig. 2) which, to the best of our knowledge, have not been reported elsewhere. The inversions are smaller (1.4, 0.6, 1.78 Mb, respectively) than the rearrangements instigating the supergenes on LGs 1, 2, 7 and 12.

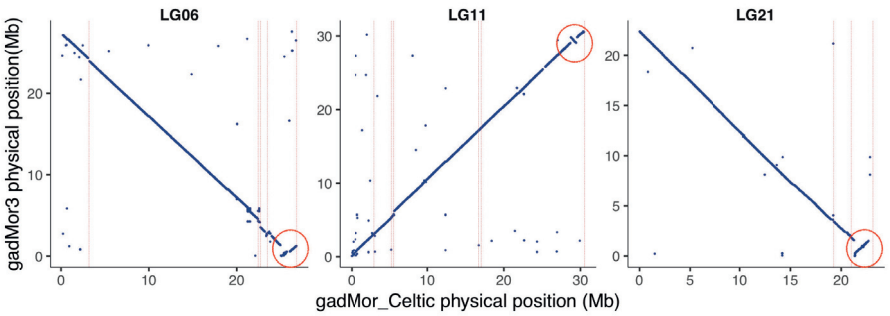

Fig.2. *Putative inversions detected on LGs 6, 11 and 21.*

Annotation of gene content and repetitive elements

There is a growing body of evidence that chromosomal inversions in fishes can capture multiple adaptive alleles and therefore act as supergenes (for example [37-39]. Defining the gene content and identifying genetic variation within rearranged regions is an important means for investigating how changes in genome organization may lead to phenotypic and adaptive divergence. Utilizing available transcript data we predict 14,292 genome wide gene models with 735, 236, 343 and 452 gene models predicted in inversions on LGs 1, 2, 7 and 12 respectively (Additional file 3). In the context of a north versus south contrast (i.e. NEAC vs Celtic) the polymorphic haemoglobin *Hbβ1* gene deserves special mention since there is good evidence for temperature-associated adaptation [40, 41]. Although the gene maps to LG02 it is located outside the inversion (approximately 3 Mb upstream) which raises questions about the mechanism maintaining its association with temperature. To document repeats in gadMor_Celtic we created a repeat library using RepeatModeler [42] which, when used with RepeatMasker [42] saw almost one third of the genome (32.26%) classified as repetitive.

Potential centromere structure and organization

Centromeres contribute to the physical linking of sister chromatids during meiosis and their location within a dyad is important for defining the chromosomal morphology (or chromosome classification) used in karyotyping studies (e.g. metacentic, acrocentric, etc). Centromeres can be relatively large and usually contain a lot of repetitive, but poorly conserved sequences [43]. Searching for known centromere repeats [43] in gadMor_Celtic assembly failed to reveal any convincing hits. We therefore used TandemRepeat finder (TRF) [44] to scan the assembly for seqences meeting characteristics typical of centromeric repeats; specifically containing more than 60% AT, longer than 80 bp, and present in all 23 LGs. We detected a 258bp sequence composed of two identical and similarly oriented 88bp repeats (one at each end) separated by an 82bp interveining sequence (see Additional file 4 for details). This expected centromeric repeat appeared 806 times (with more than 95% identity) across the genome and was found on all LGs. The location of this repeat was compared to the genetic map profiles for all 23 linkage groups (Additional File 2). We reasoned that regions of reduced recombination likely contain, or are close to, the centromere and should therefore coincide with the mapping of the centromeric repeat sequence. For most linkage groups, there was a convincing overlap between these two metrics. Most evidently, all four LGs (2, 4, 10 and 12) showing clear sigmoidal linkage profiles characteristic of a metacentric chromosome [45], contained expansion of the centromeric repeat sequence within the region of repressed recombination in the middle of the linkage group (see Fig. 3 for example).

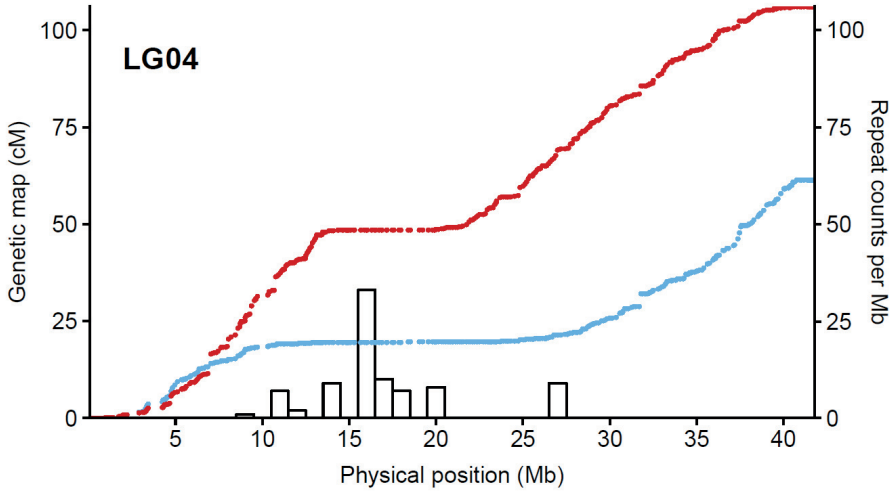

Fig. 3. *Position of potential centromere related sequence on LG04. Collinearity between LG04 genetic maps for males (red) and female (blue) and the frequency of a 258bp tandem repeat structure (histogram) predicted to be related to centromeres.*

Conclusions

In this paper we used nanopore sequencing to generate a chromosome-level genome assembly from a male Atlantic cod. The cod chosen for sequencing was captured in the Celtic Sea, where cod experience high and suboptimal summer temperatures, and represents a contrast to the current genome assemblies generated from NEAC population living in the cold Barents Sea. By generating this assembly, and comparing it against publicly available resources, we were able to characterize the chromosomal rearrangements underlying four notable supergenes displaying pronounced divergence between populations in Atlantic cod. Pairwise comparison of the two genomes also revealed additional putative rearrangements on LGs 6, 11 and 21, which has not been reported before. Identification and mapping of the centromeric repeat, combined with linkage maps, were used to study chromosomal morphology and reliably identify four characteristic metacentric chromosomes in Atlantic cod.

Methods

Sample, DNA extraction and sequencing

DNA from a single, male cod (45cm, 1009gm) fished in the Celtic Sea in January (50° 42.16N 07° 53.27W, 110m depth) was extracted from frozen blood using the Nanobind CBB Big DNA kit from Circulomics and sequenced using a PromethION instrument from Oxford Nanopore Technology (ONT). Two

sequencing libraries were generated following the ligation protocol (SQK-LSK109, ONT), one using DNA fragments >20kb, size selected using a BUF7510 High pass cassette run on a Blue Pippin (Sage Scientific), and another where no size selection was performed. Both libraries were split in two and each half sequenced successively on the same flow-cell (type R9.4.1) after nuclease flushing according to the Oxford Nanopore protocol (version: NFL_9076_v109_revF_08Oct2018). Combined data yields after quality filtering were 11.2 and 35.5 billion bases for size selected and non-size selected respectively, with median read lengths being 23.3 kb and 4.5 kb. Together this represents approximately 70X long-read genome coverage assuming an Atlantic cod genome size of 670 Mb (as is estimated for gadMor3). Short read data (2 x 250bp) was generated from non-size selected DNA using an Illumina MiSeq instrument. Libraries were prepared using a TruSeq DNA PCR free kit (Illumina) and sequenced in multiple runs to generate 71M read pairs, equalling approximately 35.5Gbp or 50X genome coverage.

Construction of the gadMor_Celtic assembly

The raw nanopore reads (n=2,868,527) was base-called using Guppy-2.2.3 (https://community.nanoporetech.com) using flip-flop model. Adapters were removed from reads using Porechop v0.2.3, 1 [46] and quality-filtered using fastp v.0.19.5.2 [47] with mean base quality greater than 7, trimming the 50bp at the 5' end of the read and removing all reads less than 4000 bp. Multiple initial assemblies applying various parameters were produced using wtdgb2 v2.3 [26]. The completeness of all assembled genomes was estimated using BUSCO v3.1.0 [48] and applying the actinopterygii (ray-finned fishes) reference gene data set. Two genome assemblies with the relative best values for contig N50, total genome size and BUSCO scores were selected (See Additional file 5) for further quality assessments.

To improve assembly contiguity, contigs showing a sequence overlap of more than 5000bp and similarity >95% were combined using quickmerge [27]. This consensus assembly was error corrected by performing two successive rounds of processing by Racon v2.3 [28] using only quality filtered nanopore reads. Raw MiSeq reads were quality filtered using Trimmomatic v0.32, before being used by Pilon v1.23 to further improve per-base accuracy in the consensus sequence. Completeness of the final polished contigs was performed as described above using BUSCO.

Linkage mapping and construction of chromosome sequences

The linkage map was constructed using 9,178 high-quality SNPs (Additional file 1) genotyped in farmed cod (n=2951) sampled from 88 families of the National cod breeding program maintained by Nofima in Tromsø, Norway, and from eight families of the CODBIOBANK at the Institute of Marine Research in Bergen,

Norway. The genotypes were produced on a 12K SNP-array created as a part of the Cod SNP Consortium (CSC) in Norway and being used in numerous previous studies [7, 13, 14, 19, 20, 22, 49].The SNPs on this array were carefully chosen to tag as many contigs as possible in the gadMor1 assembly, are thus expected to be well distributed in the genome and builds a good foundation for anchoring sequences to chromosomes. Linkage mapping was performed with the Lep-MAP software in a stepwise procedure [50]. First, SNPs were assigned to linkage groups with the 'SeparateChromosomes' command using increasing LOD thresholds until the observed number of linkage groups corresponded with the expected haploid chromosome number of 23. Additional SNPs were subsequently added to the groups with the 'JoinSingles' command at a more relaxed LOD threshold, and finally SNPs were ordered in each linkage group with the 'OrderMarkers' command. Numerous iterations were performed to optimise error and recombination parameters. Following this, sequence flanking each marker was used to precisely position all genetic markers to contigs in the gadMor_Celtic assembly using megablast [51], and thereby associate sequence with linkage groups. This analysis revealed 2 chimeric contigs containing at markers from each of different linkage groups that were selectively 'broken' using alignments with the gadMor2 assembly [23]. After breakage of the two contigs, linkage information was used to order, orientate and concatenate contigs into 23 chromosomes. Finally, SNPs were positioned in the chromosome sequences using megablast and linkage maps constructed using a fixed order in Lep-MAP to produce the final linkage maps presented in Additional file 1 and Additional file 2.

Detection of repetitive elements

RepeatModeler version 1.0.8 [42] was used to generate a repeat library, subsequently RepeatMasker version 4.0.5 [42] was run on the finished gadMor_Celtic with default options to identify the repeats in the genome assembly. For the purposes of detecting putative centromeric sequences, tandem repeats were identified using TandemRepeat finder (TRF) version 4.09 [44] with the following parameters: matching weight=2, mismatching penalty=7, indel penalty=7, match probability=80, indel probability=10, minimum score to report=30 and maximum period size to report=500. The output was processed using custom perl and unix scripts to identify repeats specifically containing more than 60% AT, longer than 80 bp, and present in all 23 LGs.

Gene annotation

Data from various public sources was used to build gene models including (i) 3M transcriptome reads generated using GS-FLX 454 technology and hosted at NCBI's SRA (https://www.ncbi.nlm.nih.gov/sra/?term = SRP013269), (ii) >250 K ESTs hosted by NCBI (https://www.ncbi.nlm.nih.gov/nucest) (iii) 4.4 M paired-end mRNA MiSeq sequences from whole NEAC larvae at 12 and 35 dph

(https://www.ebi.ac.uk/ena, PRJEB25591) and (iv) 362 M Illumina reads from 1 and 7 dph (https://www.ebi.ac.uk/ena, PRJEB25591). To enable model building, MiSeq reads and short illumina reads were mapped to the gadMor_Celtic assembly using STAR v2.3.1z [52], while 454 transcriptome reads were mapped using gmap v2014-07-28 [53] with '–no-chimeras' parameter in addition to default parameters. stringtie v1.3.3 [54] was used to assemble the reads into transcript models. Transcript models were merged using stringtie merge [54]. Gene models were tested by performing (i) open reading frame (ORF) prediction using TransDecoder [55] using both pfamA and pfamB databases for homology searches and a minimum length of 30 amino acids for ORFs without pfam support, and (ii) BLASTP analysis (evalue <1e-10) for all predicted proteins against zebrafish (*Danio rerio*) (v9.75) and three-spined stickleback (*Gasterosteus aculeatus*) (BROADS1.75) annotations from Ensembl. Only gene models with support from at least one of these homology searches were retained. Functional annotation of the predicted transcripts was done using blastx against the SwissProt database. Results from TransDecoder and homology support filtering of putative protein coding loci are shown in Additional file 6.

Availability of data and materials

The datasets generated and used during the current study, gadMor_Celtic, repeat library and all additional files are stored at figshare: doi.org/10.6084/m9.figshare.10252919. The raw nanopore reads and illumina MiSeq reads used to generate gadMor_Celtic are available at European Nucleotide Archive under accession ID PRJEB35290.

Supplementary information

Additional file 1: Linkage map of gadMor_Celtic: SNPs, position in gadMor_Celtic, genetic linkage of male, female in centimorgan (cM) and SNP flank sequence from gadMor1 (NEAC).

Additional file 2: Plots showing collinearity between genetic maps for males (red) and female (blue) and the frequency of a 258bp tandem repeat structure (histogram) predicted to be related to centromeres in all 23 chromosomes.

Additional file 3: Excel file with list of genes and positions in LGs 1, 2, 7 and 12.

Additional file 4: The putative 258bp centromere repeat sequence.

Additional file 5: wtdbg2 parameters used to generate the two initial genome assemblies.

Additional file 6: Predicted function of open reading frames were found with TransDecoder and homology search using blastp against zebrafish and stickleback protein databases.

Acknowledgements
The authors are grateful to Dr Brendan O'Hea at the Fisheries Ecosystems Advisory Service for providing the fish samples, and to our colleagues in the Cod SNP Consortium (CSC; a collaboration between CIGENE, CEES, IMR and Nofima) from where the genotypes used for the linkage analyses were derived. Funding for T.G.K. was provided by the Norwegian University of Life Sciences (NMBU). Storage resources were provided by the Norwegian National Infrastructure for Research Data (NIRD, project NS9055K).

References

1. Morris DJ, Pinnegar JK, Maxwell DL, Dye SR, Fernand LJ, Flatman S, Williams OJ, Rogers SI: **Over 10 million seawater temperature records for the United Kingdom Continental Shelf between 1880 and 2014 from 17 Cefas (United Kingdom government) marine data systems**. *Earth System Science Data* 2018, **10**(1):27-51.

2. Righton DA, Andersen KH, Neat F, Thorsteinsson V, Steingrund P, Svedang H, Michalsen K, Hinrichsen HH, Bendall V, Neuenfeldt S *et al*: **Thermal niche of Atlantic cod *Gadus morhua*: limits, tolerance and optima**. *Marine Ecology Progress Series* 2010, **420**:1-U344.

3. Mieszkowska N, Genner MJ, Hawkins SJ, Sims DW: **Effects of climate change and commercial fishing on Atlantic Cod *Gadus Morhua***. *Advances in Marine Biology* 2009, **56**:213-273.

4. Drinkwater KF: **The response of Atlantic cod (*Gadus morhua*) to future climate change**. *ICES Journal of Marine Science* 2005, **62**(7):1327-1337.

5. Neat F, Righton D: **Warm water occupancy by North Sea cod**. *Proceedings of the Royal Society B: Biological Sciences* 2007, **274**(1611):789-798.

6. Barney BT, Munkholm C, Walt DR, Palumbi SR: **Highly localized divergence within supergenes in Atlantic cod (*Gadus morhua*) within the Gulf of Maine**. *BMC Genomics* 2017, **18**(1):271.

7. Berg PR, Star B, Pampoulie C, Bradbury IR, Bentzen P, Hutchings JA, Jentoft S, Jakobsen KS: **Trans-oceanic genomic divergence of Atlantic cod ecotypes is associated with large inversions**. *Heredity* 2017, **119**(6):418-428.

8. Kirubakaran TG, Grove H, Kent MP, Sandve SR, Baranski M, Nome T, De Rosa MC, Righino B, Johansen T, Ottera H *et al*: **Two adjacent inversions maintain**

genomic differentiation between migratory and stationary ecotypes of Atlantic cod. *Molecular Ecology* 2016, **25**(10):2130-2143.

9. Karlsen BO, Klingan K, Emblem A, Jorgensen TE, Jueterbock A, Furmanek T, Hoarau G, Johansen SD, Nordeide JT, Moum T: **Genomic divergence between the migratory and stationary ecotypes of Atlantic cod**. *Molecular Ecology* 2013, **22**(20):5098-5111.

10. Hemmer-Hansen J, Nielsen EE, Therkildsen NO, Taylor MI, Ogden R, Geffen AJ, Bekkevold D, Helyar S, Pampoulie C, Johansen T *et al*: **A genomic island linked to ecotype divergence in Atlantic cod**. *Molecular Ecology* 2013, **22**(10):2653-2667.

11. Bradbury IR, Hubert S, Higgins B, Borza T, Bowman S, Paterson IG, Snelgrove PVR, Morris CJ, Gregory RS, Hardie DC *et al*: **Parallel adaptive evolution of Atlantic cod on both sides of the Atlantic Ocean in response to temperature**. *Proceedings of the Royal Society B: Biological Sciences* 2010, **277**(1701):3725-3734.

12. Bradbury IR, Hubert S, Higgins B, Bowman S, Borza T, Paterson IG, Snelgrove PVR, Morris CJ, Gregory RS, Hardie D *et al*: **Genomic islands of divergence and their consequences for the resolution of spatial structure in an exploited marine fish**. *Evolutionary Applications* 2013, **6**(3):450-461.

13. Berg PR, Jentoft S, Star B, Ring KH, Knutsen H, Lien S, Jakobsen KS, Andre C: **Adaptation to low salinity promotes genomic divergence in Atlantic cod (*Gadus morhua L.*)**. *Genome Biology and Evolution* 2015, **7**(6):1644-1663.

14. Sodeland M, Jorde PE, Lien S, Jentoft S, Berg PR, Grove H, Kent MP, Arnyasi M, Olsen EM, Knutsen H: **"Islands of Divergence" in the Atlantic Cod genome represent polymorphic chromosomal rearrangements**. *Genome Biology and Evolution* 2016, **8**(4):1012-1022.

15. Barth JMI, Villegas-Rios D, Freitas C, Moland E, Star B, Andre C, Knutsen H, Bradbury I, Dierking J, Petereit C *et al*: **Disentangling structural genomic and behavioural barriers in a sea of connectivity**. *Molecular Ecology* 2019, **28**(6):1394-1411.

16. Clucas GV, Kerr LA, Cadrin SX, Zemeckis DR, Sherwood GD, Goethel D, Whitener Z, Kovach AI: **Adaptive genetic variation underlies biocomplexity of Atlantic Cod in the Gulf of Maine and on Georges Bank**. *PLoS One* 2019, **14**(5).

17. Clucas GV, Lou RN, Therkildsen NO, Kovach AI: **Novel signals of adaptive genetic variation in northwestern Atlantic cod revealed by whole-genome sequencing**. *Evolutionary Applicatons* 2019, **12**(10):1971-1987.

18. Puncher GN, Rowe S, Rose GA, Leblanc NM, Parent GJ, Wang YJ, Pavey SA: **Chromosomal inversions in the Atlantic cod genome: Implications for management of Canada's Northern cod stock**. *Fisheries Research* 2019, **216**:29-40.

19. Kess T, Bentzen P, Lehnert SJ, Sylvester EVA, Lien S, Kent MP, Sinclair-Waters M, Morris CJ, Regular P, Fairweather R *et al*: **A migration-associated supergene reveals loss of biocomplexity in Atlantic cod**. *Science Advances* 2019, **5**(6):eaav2461.

20. Barth JMI, Berg PR, Jonsson PR, Bonanomi S, Corell H, Hemmer-Hansen J, Jakobsen KS, Johannesson K, Jorde PE, Knutsen H *et al*: **Genome architecture enables local adaptation of Atlantic cod despite high connectivity**. *Molecular Ecology* 2017, **26**(17):4452-4466.

21. Berg PR, Star B, Pampoulie C, Sodeland M, Barth JMI, Knutsen H, Jakobsen KS, Jentoft S: **Three chromosomal rearrangements promote genomic divergence between migratory and stationary ecotypes of Atlantic cod**. *Scientific Reports* 2016, **6**.

22. Sinclair-Waters M, Bradbury IR, Morris CJ, Lien S, Kent MP, Bentzen P: **Ancient chromosomal rearrangement associated with local adaptation of a postglacially colonized population of Atlantic Cod in the northwest Atlantic**. *Molecular Ecology* 2017, **27**(2):339-351.

23. Torresen OK, Star B, Jentoft S, Reinar WB, Grove H, Miller JR, Walenz BP, Knight J, Ekholm JM, Peluso P *et al*: **An improved genome assembly uncovers prolific tandem repeats in Atlantic cod**. *BMC Genomics* 2017, **18**(1):95.

24. Star B, Nederbragt AJ, Jentoft S, Grimholt U, Malmstrom M, Gregers TF, Rounge TB, Paulsen J, Solbakken MH, Sharma A *et al*: **The genome sequence of Atlantic cod reveals a unique immune system**. *Nature* 2011, **477**(7363):207-210.

25. Neat FC, Bendall V, Berx B, Wright PJ, Cuaig MO, Townhill B, Schon PJ, Lee J, Righton D: **Movement of Atlantic cod around the British Isles: implications for finer scale stock management**. *Journal of Applied Ecology* 2014, **51**(6):1564-1574.

26. Ruan, Li: **Fast and accurate long-read assembly with wtdbg2**. *BioRxiv 2019,doi:101101/530972* 2019.

27. Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ: **Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage**. *Nucleic Acids Research* 2016, **44**(19).

28. Vaser R, Sovic I, Nagarajan N, Sikic M: **Fast and accurate de novo genome assembly from long uncorrected reads**. *Genome Research* 2017, **27**(5):737-746.

29. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng QD, Wortman J, Young SK *et al*: **Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement**. *PLoS One* 2014, **9**(11).

30. Bowden R, Davies RW, Heger A, Pagnamenta AT, de Cesare M, Oikkonen LE, Parkes D, Freeman C, Dhalla F, Patel SY *et al*: **Sequencing of human genomes with nanopore technology**. *Nature Communications* 2019, **10**.

31. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT *et al*: **Nanopore sequencing and assembly of a human genome with ultra-long reads**. *Nature Biotechnology* 2018, **36**(4):338.

32. Kadobianskyi M, Schulze L, Schuelke M, Judkewitz B: **Hybrid genome assembly and annotation of Danionella translucida**. *Scientific Data* 2019, **6**.

33. Chen ZL, Omori Y, Koren S, Shirokiya T, Kuroda T, Miyamoto A, Wada H, Fujiyama A, Toyoda A, Zhang SY *et al*: ***De novo* assembly of the goldfish (Carassius auratus) genome and the evolution of genes after whole-genome duplication**. *Science Advances* 2019, **5**(6).

34. Harris RS: **Improved pairwise alignment of genomic DNA**. The Pennsylvania State University.; 2007.

35. Godo OR, Michalsen K: **Migratory behaviour of north-east Arctic cod, studied by use of data storage tags**. *Fisheries Research* 2000, **48**(2):127-140.

36. Hobson VJ, Righton D, Metcalfe JD, Hays GC: **Vertical movements of North Sea cod**. *Marine Ecology Progress Series* 2007, **347**:101-110.

37. Pettersson ME, Rochus CM, Han F, Chen J, Hill J, Wallerman O, Fan G, Hong X, Xu Q, Zhang H *et al*: **A chromosome-level assembly of the Atlantic herring genome-detection of a supergene and other signals of selection**. *Genome Research* 2019, **29**(11):1919-1928.

38. Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S *et al*: **The genomic basis of adaptive evolution in threespine sticklebacks**. *Nature* 2012, **484**(7392):55-61.

39. Pearse DE, Barson NJ, Nome T, Gao G, Campbell MA, Abadía-Cardoso A, Anderson EC, Rundio DE, Williams TH, Naish KA *et al*: **Sex-dependent dominance maintains migration supergene in rainbow trout**. *BioRxiv* 2018. **doi:** https://doi.org/10.1101/504621.

40. Frydenberg O, Moller D, Naevdal G, Sick K: **Haemoglobin Polymorphism in Norwegian Cod Populations**. *Hereditas-Genetisk A* 1965, **53**(1-2):257.

41. Andersen O: **Hemoglobin polymorphisms in Atlantic cod - a review of 50 years of study**. *Marine Genomics* 2012, **8**:59-65.

42. Smit AFA, Hubley R. **RepeatModeler Open-1.0.** 2008–2015. (http://www.repeatmasker.org).

43. Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, Sebra R, Peluso P, Eid J, Rank D *et al*: **Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution**. *Genome Biology* 2013, **14**(1).

44. Benson G: **Tandem repeats finder: a program to analyze DNA sequences**. *Nucleic Acids Research* 1999, **27**(2):573-580.

45. Ghigliotti L, Fevolden SE, Cheng CH, Babiak I, Dettai A, Pisano E: **Karyotyping and cytogenetic mapping of Atlantic cod (*Gadus morhua* Linnaeus, 1758)**. *Animal Genetics* 2012, **43**(6):746-752.

46. Wick RR, Judd LM, Holt KE: **Deepbinner: Demultiplexing barcoded Oxford Nanopore reads with deep convolutional neural networks**. *PLoS Computational Biology* 2018, **14**(11):e1006583.

47. Chen S, Zhou Y, Chen Y, Gu J: **fastp: an ultra-fast all-in-one FASTQ preprocessor**. *Bioinformatics* 2018, **34**(17):i884-i890.

48. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs**. *Bioinformatics* 2015, **31**(19):3210-3212.

49. Knutsen H, Jorde PE, Hutchings JA, Hemmer-Hansen J, Gronkjaer P, Jorgensen KM, Andre C, Sodeland M, Albretsen J, Olsen EM: **Stable coexistence of genetically divergent Atlantic cod ecotypes at multiple spatial scales**. *Evolutionary Applicaions* 2018, **11**(9):1527-1539.

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

51. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic Local Alignment Search Tool**. *Journal of Molecular Biology* 1990, **215**(3):403-410.

52. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner**. *Bioinformatics* 2013, **29**(1):15-21.

53. Wu TD, Watanabe CK: **GMAP: a genomic mapping and alignment program for mRNA and EST sequences**. *Bioinformatics* 2005, **21**(9):1859-1875.

54. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL: **StringTie enables improved reconstruction of a transcriptome from RNA-seq reads**. *Nature Biotechnology* 2015, **33**(3):290-295.

55. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M *et al*: **De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis**. *Nature Protocols* 2013, **8**(8):1494-1512.

Paper III

SCIENTIFIC REPORTS

OPEN

Characterization of a male specific region containing a candidate sex determining gene in Atlantic cod

Received: 13 August 2018
Accepted: 25 November 2018
Published online: 15 January 2019

Tina Graceline Kirubakaran[1], Øivind Andersen[1,2], Maria Cristina De Rosa[3], Terese Andersstuen[1], Kristina Hallan[1], Matthew Peter Kent[1] & Sigbjørn Lien[1]

The genetic mechanisms determining sex in teleost fishes are highly variable and the master sex determining gene has only been identified in few species. Here we characterize a male-specific region of 9 kb on linkage group 11 in Atlantic cod (*Gadus morhua*) harboring a single gene named *zkY* for zinc knuckle on the Y chromosome. Diagnostic PCR test of phenotypically sexed males and females confirm the sex-specific nature of the Y-sequence. We identified twelve highly similar autosomal gene copies of *zkY*, of which eight code for proteins containing the zinc knuckle motif. 3D modeling suggests that the amino acid changes observed in six copies might influence the putative RNA-binding specificity. Cod zkY and the autosomal proteins zk1 and zk2 possess an identical zinc knuckle structure, but only the Y-specific gene *zkY* was expressed at high levels in the developing larvae before the onset of sex differentiation. Collectively these data suggest *zkY* as a candidate master masculinization gene in Atlantic cod. PCR amplification of Y-sequences in Arctic cod (*Arctogadus glacialis*) and Greenland cod (*Gadus macrocephalus ogac*) suggests that the male-specific region emerged in codfishes more than 7.5 million years ago.

The origin and evolution of sex chromosomes from autosomes, and the mechanism of sex determination have long been subjects of interest to biologists. In eutherian mammals and birds, the sex chromosomes are highly dimorphic and have degenerated with extensive gene losses[1–3]. However in teleost fishes, cytogenetically different sex chromosomes are found in less than 10% of the species examined[4,5], and their recent origin in several lineages makes them good models to study the early stages of divergence. The frequent turnover of sex chromosomes seems to be associated with the variety of sex determining genes, even in closely related species[4–6]. For example, in medaka fishes (genus *Oryzias*), the Y-chromosomal *dmY* or *dmrt1bY* copy determines the sex in *O. latipes* and *O. curvinotus*, while testicular differentiation in *O. dancena* and in *O. luzonensis* is initiated by male-specific regulatory elements upstream of *sox3* and *gsdf* (gonadal soma-derived factor), respectively[7–10]. The various mechanisms of genetic sex determination in fish range from sex–specific alleles of the anti-Müllerian hormone (Amh) in Nile tilapia (*Oreochromis niloticus*) and the Amh type 2 receptor (*Amhr2*) in tiger pufferfish (*Takifugu rubripes*)[11,12] to complex polygenic regulation involving several genomic regions as shown in the cichlid *Astatotilapia burtoni*[13]. The diversity of mechanisms is further increased by the insertion of X- and Y- sequences modulating the expression of the neighboring master sex determinant gene, as found in the sablefish (*Anoplopoma fimbria*)[14].

Sexual differentiation in vertebrates is initially marked by highly increased proliferation of the germ cells in the presumptive ovaries compared to that in testes[6,15–19]. The time course of the sex-dimorphic germ cell proliferation varies greatly among teleost species and occurs around hatching in medaka[20], at the start-feeding stage in Atlantic cod[21] and in late juvenile stages in sea bass (*Dicentrarchus labrax*)[22]. Male germ cell divisions are inhibited by Dmy in the medaka *O. latipes*[23], and the sex determinants Gdsf, Amh and Amhr2 of the TGFβ signaling pathway might play similar roles in inducing sex differentiation in other species[6]. However, the requirement of germ cells for gonadal development appears to vary among teleost species[24–26] reflecting that sex determination might be triggered by alternative mechanisms. Most sex determinants found in vertebrates are thought to have

[1]Centre for Integrative Genetics (CIGENE), Department of Animal and Aquacultural Sciences (IHA), Faculty of Life Sciences (BIOVIT), Norwegian University of Life Sciences (NMBU), PO Box 5003, 1433, Ås, Norway. [2]Nofima, PO Box 5010, N-1430, Ås, Norway. [3]Institute of Chemistry of Molecular Recognition - CNR c/o Institute of Biochemistry and Clinical Biochemistry, Catholic University of Rome, 00168, Rome, Italy. Matthew Peter Kent and Sigbjørn Lien jointly supervised this work. Correspondence and requests for materials should be addressed to M.P.K. (email: matthew.peter.kent@nmbu.no) or S.L. (email: sigbjorn.lien@nmbu.no)

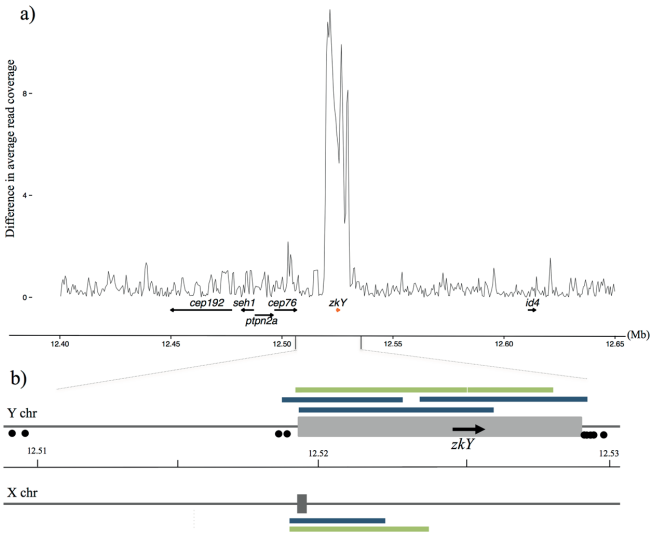

Figure 1. Genomic organization of the Y- and X-sequences on LG11 in Atlantic cod. (**a**) Comparison of read depth between male (n = 49) and female (n = 53) samples reveals an excess of Y-specific reads across a 9 kb interval on LG11 in the final and public gadMor2.1 assembly. Positions of cod *zkY* and the neighboring genes are indicated. (**b**) The Y- and X-sequences (grey boxes), including the Y-specific *zkY* gene, and the nine polymorphisms heterozygous in all males and homozygous in all females (black dots). Long reads confirming assembly integrity were generated using Oxford Nanopore (green) and Pacific Biosciences (blue) sequencing technologies.

acquired this role by being recruited from conserved downstream regulators of gonadal differentiation, although the function of some actors differs among lineages[6,27–29]. An exception is the novel salmonid sdY; a truncated male-specific copy of the interferon regulatory factor 9 (irf9), which is an immune-related gene not associated with sex[30]. Knowledge about how genes are functionalized and incorporated at the top of the sex regulatory cascade should broaden our understanding of the genetic regulation of sex determination and elucidate whether there are constraints on the types of genes that can be co-opted as master control switches.

Atlantic cod is an economically important cold-water marine species widely distributed in the North Atlantic Ocean. Due to the decline in many cod stocks, cod farming has attracted interest but production is hampered by precocious early sexual maturation, particularly in males. This could partly be solved by production of all-female triploid stocks[4,31,32]. Gynogenetic and sex-reversed cod populations have demonstrated a XX-XY sex determination system[21,33,34], but karyotyping of Atlantic cod and the closely related European hake (*Merluccius merluccius*) failed to reveal sex-linked chromosome heteromorphism[35,36]. Recently, whole genome sequence data from wild Atlantic cod was used to identify genotypes segregating closely with a XX-XY system in six putative sex determining regions distributed across five linkage groups[37]. Here we use whole genome re-sequencing data from 49 males and 53 females, together with long-read sequence data and Sanger sequencing of targeted PCR products, to characterize a Y-sequence of 9,149 base pairs on LG11 harboring a single gene which we have named zinc knuckle on the Y chromosome gene *(zkY)*. Gene expression data from early development stages and modeling of the zinc knuckle structure offer circumstantial evidence consistent with a function in Atlantic cod sex determination.

Results and Discussion

Identification of male specific sequence on linkage group 11. Two independent approaches were used to identify a male specific genomic region in Atlantic cod. Firstly, we searched whole genome sequence data from males and females for SNPs that segregate according to an XX-XY system. Illumina short-reads (approximately 10X coverage per individual) generated from 49 males and 53 females were mapped to an initial in-house developed gadMor2.1 genome assembly, followed by variant detection. When applying stringent criteria demanding that gender specific variants should be heterozygous in all 49 males and homozygous in all 53 females, we detected 9 variants all distributed within a 15 kb region on LG11 (Supplementary File 1 and Fig. 1b). BLAST alignments revealed that these variants fell inside the 55 kb region previously reported as showing most evidence for being involved in sex determination by Star *et al.*[37]. Secondly, we analyzed resequencing data to identify positions in the genome that displayed significant differences in read depth between males and females. Two regions showed an almost complete absence of female reads while displaying >500 reads from males, i.e. Y-specific characteristics. These regions were separated by an intervening sequence displaying roughly 2-times coverage in females versus males, i.e. a possible X-sequence. Both the Y- and X-sequence regions fell within an interval flanked by the nine sex-linked markers on LG11 lending further support to the significance of this region in sex

determination. Closer examination of the architecture of this region using IGV[38] revealed that despite a high read-depth, there was no evidence of single reads bridging, or read-pairs spanning, these differential read-depth sub-regions. Collectively these anomalies are suggestive of assembly errors in the initial gadMor2.1 assembly, and a hybridization of X- and Y-sequences, and underscores the importance of developing high-quality reference genomes.

In order to resolve this complex region we performed nested, long-range PCR to generate a Y-specific fragment, which was then sequenced using an Illumina MiSeq. Assembly construction using Gap Filler generated a 6 kb sequence, which was integrated into the initial gadMor2.1 assembly to produce a complete Y-sequence of 9,149 bp (Fig. 1 and Supplementary File 2). The sequencing revealed an imperfect repeated element positioned close to the end of the Y-sequence (see underlined sequence in Supplementary File 2). An X-sequence of 425 bp was constructed by Sanger sequencing PCR fragments generated from females using primers flanking the Y-sequence. To confirm the integrity of these manually curated X- and Y-sequences we aligned publicly available long-read data (Pacific BioSciences) derived from the male used to generate the reference (accession numbers at ENA (http://www.ebi.ac.uk/ena), ERX1787826-ERX1787972) and long reads generated in-house using an Oxford Nanopore MinION device, also from a male. Although overall coverage was low, we were able to identify long-reads that aligned specifically to either the Y- or X-sequences and, in combination, spanned their entire lengths (Supplementary File 3 and Fig. 1b).

The Y-sequence contains a single gene. X- and Y-sequences of the public gadMor2.1 assembly were annotated using standard gene-model predictive software and RNA-Seq data produced from early larval stages together with publicly available RNA-Seq data from Atlantic cod (See Material and Methods). This revealed a single gene (which we have named *zkY*) in the Y-sequence (Supplementary File 2 and Fig. 1), which is inserted in a synteny block that is highly conserved in teleosts and in spotted gar (*Lepisosteus oculatus*) (Supplementary Figure 1). The intronless cod *zkY* codes for a zinc knuckle protein characterized by the zinc knuckle consensus sequence Cys-X2-Cys-X4-His-X4-Cys (X = any amino acid). This motif is mainly found in the RNA-binding retroviral nucleocapsid (NC) proteins, but also in various eukaryotic proteins binding single-stranded nucleotide targets[39–42]. The flanking basic residues bind nucleic acids non-specifically through electrostatic interactions with the phosphodiester backbone of nucleic acids[43].

Zinc knuckle proteins are members of the large family of zinc finger proteins possessing a versatility of tetrahedral Cys- and His-containing motifs that bind to DNA and RNA target sites[44,45]. The DNA-binding domain of the DMRT transcription factors, including the sex determinant DmY in medaka, consists of intertwined CCHC and HCCC motifs binding in the minor groove of DNA[46]. The zinc finger protein ZFAND3 is essential for spermatogenesis in mice and the polymorphic tilapia *zfand3* was recently mapped in the sex determining locus[47,48]. Hence, the recruitment of zinc finger proteins as the sex determinant seems to have occurred several times in teleosts.

Multiple autosomal copies of cod zkY. The sex determining gene in several teleosts exhibits an autosomal copy that differs from the male-specific gene in spatio-temporal expression patterns and/or functionality[28,49–52]. We identified 12 autosomal copies of cod *zkY* mapping to seven different linkage groups and one unassembled scaffold. Premature stop codons and single base indels were found in four of the genes encoding putatively non-functional proteins lacking the zinc knuckle domain. The remaining eight genes code for zinc knuckle proteins named zk1 to zk8, which differ in length from 143 to 633 amino acids (Supplementary File 4). Sequence alignment revealed several amino acid substitutions in the zinc knuckle domain in addition to the variable size of an imperfect repeat of basic residues preceding the domain (Fig. 2A, Supplementary File 5 and Supplementary Figure 2).

Functional implications of the amino acid substitutions in the zinc knuckle domain of the autosomal proteins was inferred by exploring the predicted 3D structure and the interactions with a putative RNA oligonucleotide target. The suitability of the d(ACGCC) sequence in the structure modeling was supported by the stacking of the conserved Trp442 between the two bases G3 and C4 in agreement with the interactions between specific base moieties and Trp at the corresponding position in NC proteins[53–55]. The three cysteines Cys443, Cys436 and Cys446 together with His441 coordinate the tetrahedral binding of the Zn^{2+} ion in the modelled cod zinc knuckles. While the replacement of Met433 in zkY, zk1 and zk2 with the Ile residue in the four zk3-zk7 proteins maintains the hydrophobic interactions with Cys446, the Ile433 residue is predicted to interact with the G3 base of the pentanucleotide (Fig. 2C). The RNA-binding specificity of human HIV-1 was consistently altered by the Met- > Val and Met- > Lys substitutions in the same position of the second zinc knuckle[56]. The Asn435Lys change in the zk8 protein may affect the tetrahedral Zn^{2+} ion coordination by interacting with Cys446, while Gln443 is predicted to form a hydrogen bond with the phosphate group connecting A1 and C2 (Fig. 2D). Additionally, electrostatic interactions are predicted between Arg437 and His434, which are moved away from its C2 contacts observed in the other variants.

Altogether, the predicted alterations in the contacts between the replaced amino acids and the interactions with the oligonucleotide sequence template suggest functional differences in the putative RNA-binding activity of the zinc knuckle proteins examined. In addition, the functional role played by the flanking basic residues in the non-specific binding of nucleic acids[44] suggests that these interactions might be altered by the extended flanking repeat in the zk6 and zk7 proteins. Based on the identical zinc knuckle domain and the similar flanking sequences in the three proteins zkY, zk1 and zk2, we decided to compare the larval expression of these coding genes.

Increased transcription of cod *zkY* prior to onset of sex differentiation. A prerequisite for being a sex determining gene is to show pronounced transcription levels prior to onset of sex differentiation, which in Atlantic cod likely occurs at start feeding around 35 days post hatching (dph) based on the high number of germ

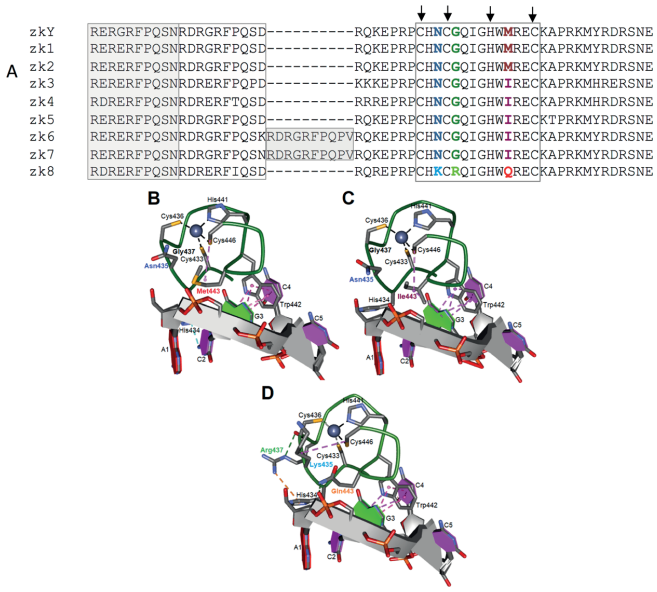

Figure 2. Amino acid substitutions in the zinc knuckle domain in cod zkY and its autosomal copies. (**A**) Sequence alignment of the zinc knuckle domain (boxed) and flanking regions in cod zkY and the eight autosomal protein variants. The characteristic Cys-Cys-His-Cys residues of the zinc knuckle are arrowed, substituted amino acids are indicated by colors and correspond to the labelled positions 435, 437, and 443 in (**B**-**D**). Repeated segments adjacent to the zinc knuckle are shaded. (**B**–**D**) Modeled structure of the three different zinc knuckle domains of cod zkY and the autosomal proteins. The Zn^{2+} ion is displayed together with the oligonucleotide d(ACGCC) template (see Methods). (**B**) Met443 variant of zkY, zk1, zk2, (**C**) Ile443 variant of zk3-zk7, (**D**) Lys435-Arg437-Gln443 variant of zk8. Hydrogen bonds are indicated by dotted green lines, Pi-donor hydrogen bonds in light blue, hydrophobic interactions by dotted magenta lines, and electrostatic bonds by dotted orange lines.

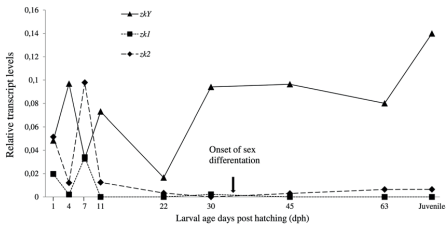

Figure 3. Larval expression of *zkY, zk1* and *zk2* from hatching to juvenile stage. Expression levels are given as relative transcript read numbers. The onset of sex differentiation is indicated[21].

cells in females compared to males together with the female-biased expression of *cyp19a1a*[21,57]. RNA-Seq analysis of cod *zkY* and the autosomal *zk1* and *zk2* copies showed that the three genes were expressed at variable levels from hatching until 22 dph at which stage the transcription levels of *zkY* increased and stabilized at relatively high levels from 30 dph onwards (Fig. 3). In contrast, the larval expression of *zk1* and *zk2* was almost undetectable or very low from 11 dph.

The increased larval expression of the male-specific *zkY* gene prior to the onset of sex differentiation is consistent with a possible function in sex determination. Multiple germline-specific RNA regulatory proteins are essential for germ cell specification, maintenance, migration and proliferation in various organisms[58]. While this regulatory network seems to be evolutionary conserved in nematodes, flies and mammals, the key role played by different sex determinants in the control of male germ cell proliferation in fish suggests the involvement of lineage-specific regulators[6]. Knockdown of the germline zinc knuckle helicases *glh1-2* in *C. elegans* and the *vasa* homolog in mice resulted in male infertility[59–61]. Intriguingly, the loss of zinc knuckles in the RNA-binding Vasa

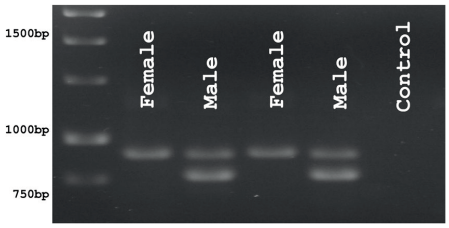

Figure 4. Agarose gel separation of PCR products from male and female Atlantic cod. (Cropped from a larger gel image presented in Supplementary Figure 4).

in vertebrates and insects was suggested to have coincided with the emergence of hitherto unknown zinc-knuckle cofactors conferring target specificity[62].

X- and Y-sequences in other gadoid species. To elucidate the emergence of this putative sex determining region in other cod fishes we scanned draft genome assemblies from 12 species within the family *Gadidae*[63] for sequences matching the Atlantic cod Y- sequence. In its entirety, the 9,149 bp male-specific region showed only partial hits suggesting that the Y- sequence is absent or incorrectly assembled in these draft assemblies. Therefore, to find evidence for the sex determination region in other codfish species we were obliged to use a PCR based approach. Although the multi-species sequence alignments were highly fragmented, a close examination revealed isolated regions within and immediately outside the male-specific region of Atlantic cod which were, to some extent, conserved across species. These regions allowed us to develop primers, which amplified a specific fragment of the Y-sequence in both Greenland cod (*Gadus macrocephalus ogac*) and Arctic cod (*Arctogadus glacialis*), in addition to Atlantic cod (see alignment in Supplementary File 6). The X-sequence was efficiently amplified and sequenced in Greenland cod, Arctic cod, polar cod (*Boreogadus saida*), haddock (*Melanogrammes aeglefinus*) and burbot (*Lota lota*) (see alignment in Supplementary File 7). This result documents that the Y-specific sequence was present prior to the separation of Arctic cod from Atlantic cod more than 7.5 MYA[63,64] while the presence of X-sequence in burbot suggests that this arrangement predates the *Lotinae – Gadinae* split 45 MYA (Supplementary Figure 3).

A diagnostic test for distinguishing male and female Atlantic cod is potentially useful for researchers and the emerging aquaculture industry. To efficiently determine gender, we developed a simple PCR reaction including two different forward primers (one Y-sequence specific and the other matching a common sequence upstream of the X- and Y- regions) and a common reverse primer (annealing to a sequence downstream of the X- and Y-regions) resulting in a single band in females and double in males (Fig. 4).

Sex determination is a complex process involving the coordinated actions of multiple factors, and a greater understanding of what genes are involved in the sex determination cascade is needed. Furthermore, this understanding must be developed in multiple species to learn whether these processes are species or taxa specific and to improve our knowledge of the evolution of sex determination. The main focus of this paper has been to improve our knowledge of the genomic basis for sex determination in Atlantic cod. Our results indicate that the *zkY* gene present within the male specific region on LG11 is involved in Atlantic cod sex determination, and documents the presence of X- and Y- sequences in other gadoids.

Materials and Methods

Ethics statement. To produce whole genome sequence data, fin clips were collected non-destructively in strict accordance with the welfare rules given by the National Animal Research Authority (NARA). Samples were collected as a part of an ongoing, long term project authorized and managed by Nofima AS which seeks to maintain a biobank of samples from parents of the National cod breeding program nucleus. For RNA sequencing (larvae 1–35 dph), permission for sampling biological material is not required for fish eggs and larvae collected before start-feeding according to NARA.

Sampling and Illumina whole-genome sequencing. Tissue was collected from 49 males and 53 females from the National Atlantic cod breeding program maintained by Nofima, Tromsø, Norway (available at https://www.ebi.ac.uk/ena, PRJEB12803 and Supplementary File 8). The fish were parents in year classes 2005 or 2006 and represented the second generation of cod produced in captivity. The original broodstock in the base population were sampled from different geographical areas along the Norwegian coast[65]. DNA was extracted using the DNeasy kit manufactured by QIAGEN (Germany) according to manufacturer's instructions. Libraries were prepared using the Truseq Library prep kit (Illumina, San Diego, USA), and sequenced using an Illumina HiSeq 2500 instrument. A total of 13.1 billion paired-reads (2 × 100 PE) were produced, averaging 129 million paired reads per individual with coverage from 12.4 to 19.0X.

Construction of the initial gadMor2.1 assembly and variant detection. The initial gadMor2.1 assembly was constructed by integrating a dense linkage map with two public draft assemblies (NEWB454 and CA454ILM https://figshare.com/articles/Transcript_and_genome_assemblies_of_Atlantic_cod/3408247) generated from sequencing the same male. The assemblies were constructed using different combinations of short-read sequencing data and different assembly programs (Newbler and Celera respectively) and displayed

different qualities with NEWB454 having longer scaffold N50s and more gaps than CA454ILM[66]. The linkage map containing 9354 SNP markers was produced after genotyping a pedigree of 2951 fishes. Chimeric scaffolds were broken at junctions between contigs containing SNPs from different LGs. Overlapping scaffolds were identified by comparing SNPs mapping to both assemblies and were merged using coordinates from alignment with LASTZ[67] generating scaffolds that were used to build the final chromosome sequences. Finally, all scaffolds were oriented, ordered and concatenated into a new chromosome sequence based on information from the linkage map.

Variants were detected by first filtering raw reads from each individual using Trimmomatic v0.32[68], and subsequently aligning reads to the unmasked initial gadMor2.1 assembly using Bowtie2 v2.3.2 with the parameter 'sensitive'[69]. The resulting BAM files were merged by gender using Sambamba v0.6.5[70] and GATK HaplotypeCaller v3.8[71] was used to call variants with the following parameters: gt_mode DISCOVERY, minPruning 3. A Perl script was used to report SNPs, and their positions, that display heterozygous genotypes in all males and homozygous genotypes in all females.

Read depth differences between males and females. For each sex, quality filtered reads from males (n = 49) and females (n = 53) were combined to generate two gender-specific bed files. Bedtools genomecov version 2.22.0[72] was then used to count read depth at each position in the public gadMor2.1 assembly. A Perl script was then used to calculate the average read depth per individual across successive 1000 bp intervals with an overlap of 500 bp (positions with 0 coverage in males and females were ignored). For each average, the absolute difference between the larger and smaller number was calculated and plotted. Across the genome, the region with the most extreme read depth differences was seen on LG11 (Fig. 1a).

Construction of Y- and X-sequences. A nested PCR strategy was used to amplify the Y-specific sequence lacking from the initial gadMor2.1 assembly. The initial reaction used the following primers and conditions, (Fwd; 5′-CCTAAACCACAGTCCTGGGC-3′, Rev; 5′-ACATTGTGCACACACATTGTATC-3′, PrimeSTAR GXL DNA Polymerase (Takara, Japan), cycling 30 times with 10 s at 98 °C, 15 s at 57 °C, 3 min at 68 °C, final extension 72 °C for 10 min. PCR-product was used as template in a subsequent reaction using the same conditions but with the nested PCR primers (Fwd; 5′-CTCTGTAGTTTGTGGTGGGGT-3′, Rev; 5′-ACATGACGAATGGCCTCCTTT-3′). The resulting PCR product was purified using a QIAquick PCR purification kit from QIAGEN (Germany) and quantified. DNA was prepared for sequencing using a Nextera XT kit from Illumina (USA) according to manufacturer's instructions and the resulting library sequenced using 2 × 250 nt sequencing chemistry using a MiSeq (Illumina, USA). After quality trimming the resulting reads were anchored to the existing flanking Y-sequences using Gap filler[73]. The resulting sequence contained a single gap that was filled by Sanger sequencing a PCR product generated using the following PCR primers and conditions (Fwd; 5′-ACACAACGCAGAGTCTGTCC-3′, Rev: 5′-TCAGCTAGTCTCGCAATGGC-3′, denaturation 15 min at 95 °C, then cycling 30 times with 10 s at 98 °C, 15 s at 98 °C, 60 s at 68 °C, final extension 10 min at 72 °C). An X-sequence was constructed from Sanger sequencing a PCR product produced using primers annealing up- and down-stream of the Y-sequence.

Generation X- and Y-sequences in other gadoids. Sequence alignments of male specific region in Atlantic cod with public assembly data from 12 gadid species[63], using MUMmer version 3.23[74], identified four rather short regions that showed good conservation across species, including two regions within the Y-sequence (B; 12,520,541–12,520,838 and C; 12,527,782–12,5280,27) and two regions (A; 12,518,827–12,519,231 and D; 12,528,480–12,529,285) immediately flanking the X- and Y-sequences. Primers were designed using Primer3[75] to amplify from within the Y-sequence conserved sequence C to flanking sequence D (i.e. a Y-specific PCR) and from flanking sequence A to D (an X-sequence specific PCR). Products from these reactions were Sanger sequenced and aligned using MAFFT v7.402[76] (Supplementary Files 6 and 7).

Diagnostic sex-test. A sex distinguishing duplex PCR was designed for Atlantic cod using the following primers and conditions (Fwd A; 5′-ACACACGGTCTGCTGTAGTG-3′, Fwd C; 5′-GGAGGGGAATTGTACAAACACG-3′, Rev D; 5′-GTGTGCCAAATGGATGCCAA-3′), denaturation for 15 min at 95 °C, cycling 35 times with 30 s at 94 °C, 60 s at 55 °C, 60 s at 72 °C, final extension 72 °C for 10 min (Supplementary File 9).

Nanopore Sequencing. High-molecular weight DNA was extracted using a phenol-chloroform method[77]. Sequencing libraries were prepared using SQK-RAD003 and SQK-LSK108 kits and protocols from Oxford Nanopore Technology (Oxford, UK) and sequenced on a MinION device generating a combined total of 3.9 Gb sequence data with an N50 read length of 4.2 kb. Reads were aligned to the public gadMor2.1 assembly using GraphMap aligner v0.5.2[78].

Transcript profiling of cod larvae. Cod larvae were hatched at the National Atlantic Cod Breeding Centre in Tromsø, Norway, after the incubation of fertilized eggs in seawater rearing tanks at 4.5 °C. For 1 and 7 dph larvae, RNA was extracted from a pool of 10 individuals using the QIAGEN AllPrep DNA/RNA/miRNA Universal kit (QIAGEN; Germany). Samples were prepared for sequencing using TruSeq Stranded mRNA kit from Illumina (USA) and sequenced using an Illumina HiSeq 2000 to produce 362 million reads. For the samples 12 and 35 dph samples, RNA was extracted using an RNeasy kit (QIAGEN, Germany) prepared for sequencing using TruSeq Stranded mRNA kit from Illumina (USA) and sequenced using an Illumina MiSeq (2 × 250nt) to produce 4.4 million reads. All reads were trimmed using Trimmomatic v0.32[68] before further analysis.

Total available short read data (http://www.ebi.ac.uk/ena, PRJEB18628; and https://www.ncbi.nlm.nih.gov/sra/?term = SRP056073) was binned based on days before hatching (dph) before being aligned to the public gadMor2.1 assembly using star aligner STAR v2.3.1z12 as described previously. Potential transcripts were constructed

using stringtie v1.3.3[79], while cuffmerge v2.2.1[80] was used to produce a GTF file containing key metrics. FPKM values for zkY, zky1, zky2, were calculated using only those reads with a mapping quality of ≥30.

Gene annotation. Data from various public sources was used to build gene models including (i) 3M transcriptome reads generated using GS-FLX 454 technology and hosted at NCBI's SRA (https://www.ncbi.nlm.nih.gov/sra/?term = SRP013269), (ii) >250 K ESTs hosted by NCBI (https://www.ncbi.nlm.nih.gov/nucest) (iii) 4.4 M paired-end mRNA MiSeq sequences from whole NEAC larvae at 12 and 35 dph (https://www.ebi.ac.uk/ena, PRJEB25591), (iv) Pacbio reads from (https://www.ebi.ac.uk/ena, PRJEB18628), (v) 362 M Illumina reads from 1 and 7 dph (https://www.ebi.ac.uk/ena, PRJEB25591) and (vi) approximately 1.7B Illumina reads from 4–63 dph as well as juvenile samples (https://www.ncbi.nlm.nih.gov/sra/?term = SRP056073). To enable model building, short Illumina reads (<250nt) were mapped to the public gadMor2.1 assembly using STAR v2.3.1z12. Because Illumina reads from 4 dph - juvenile were generated using an unstranded library, the parameter' outSAMstrandField intronMotif' was used in alignment. Long reads from PacBio were mapped using STARlong v2.5.2a[81] while 454 transcriptome reads were mapped using gmap v 2014-07-28[82] with '–no-chimeras' parameter in addition to default parameters. Cufflinks v2.2.1[80] was used to assemble the reads into transcript models for all alignments except for data from 4 dph - juvenile stage samples where stringtie v1.3.3[79] was used. Transcript models were merged using cuffmerge v2.2.1[80].

Gene models were tested by performing (i) open reading frame (ORF) prediction using TransDecoder[83] using both pfamA and pfamB databases for homology searches and a minimum length of 30 amino acids for ORFs without pfam support, and (ii) BLASTP analysis (evalue <1e-10) for all predicted proteins against zebrafish (*Danio rerio*) (v9.75) and three-spined stickleback (*Gasterosteus aculeatus*) (BROADS1.75) annotations from Ensembl. Only gene models with support from at least one of these homology searches were retained. Functional annotation of the predicted transcripts was done using blastx against the SwissProt database.

Modeling the zinc knuckle domain. The three-dimensional structure of the three different variants of the zinc knuckle domain were built using a threading approach which combines three-dimensional fold recognition by sequence alignment with template crystal structures and model structure refining. The query-template alignment was generated by HHPRED (https://toolkit.tuebingen.mpg.de) and then submitted to the program MODELLER[84] as implemented in Discovery Studio (Dassault Systèmes BIOVIA). Query coverage and e-value score were considered to define a suitable template structure. The structure of nucleocapsid protein NCp10 of retrovirus MoMuLV, which contains a single Cys-X2-Cys-X4-His-X4-Cys zinc knuckle domain bound to the oligonucleotide d(ACGCC) was selected as template (https://www.rcsb.org/structure/1a6b). Zinc ions and oligonucleotides were explicitly considered through molecular modeling steps. Fifty models, optimized by a short simulated annealing refinement protocol available in MODELLER, were generated and their consistency was evaluated on the basis of the probability density function violations provided by the program. Stereochemistry of selected models was checked using the program PROCHECK[85]. Visualization and manipulation of molecular images were performed with Discovery Studio (Dassault Systèmes BIOVIA).

Data Availability

The gadMor2.1 genome assembly, long-range PCR MiSeq Illumina data are available at https://figshare.com/s/313f8fe1fdcc82571a99. The 102 individual samples read data are available at ENA, with the study accession number PRJEB12803. The MiSeq whole NEAC larvae at 12,35 dph and 1,7 dph illumina samples are available at ENA, with the study accession number PRJEB25591. All data generated or analyzed during this study are included in this published article and its Supplementary Information files.

References

1. Graves, J. A. Sex chromosome specialization and degeneration in mammals. *Cell* **124**, 901–914 (2006).
2. Bellott, D. W. *et al.* Convergent evolution of chicken Z and human X chromosomes by expansion and gene acquisition. *Nature* **466**, 612–U613 (2010).
3. Peichel, C. L. Convergence and divergence in sex-chromosome evolution. *Nat Genet* **49**, 321–322 (2017).
4. Devlin, R. H. & Nagahama, Y. Sex determination and sex differentiation in fish: an overview of genetic, physiological, and environmental influences. *Aquaculture* **208**, 191–364 (2002).
5. Volff, J. N., Nanda, I., Schmid, M. & Schartl, M. Governing sex determination in fish: Regulatory putsches and ephemeral dictators. *Sex Dev* **1**, 85–99 (2007).
6. Kikuchi, K. & Hamaguchi, S. Novel sex-determining genes in fish and sex chromosome evolution. *Dev Dyn* **242**, 339–353 (2013).
7. Nanda, I. *et al.* A duplicated copy of DMRT1 in the sex-determining region of the Y chromosome of the medaka. *Oryzias latipes. PNAS* **99**, 11778–11783 (2002).
8. Matsuda, M. *et al.* DMY is a Y-specific DM-domain gene required for male development in the medaka fish. *Nature* **417**, 559–563 (2002).
9. Myosho, T. *et al.* Tracing the emergence of a novel sex-determining gene in medaka. *Oryzias luzonensis. Genetics* **191**, 163–170 (2012).
10. Takehana, Y., Nagai, N., Matsuda, M., Tsuchiya, K. & Sakaizumi, M. Geographic variation and diversity of the cytochrome b gene in Japanese wild populations of medaka. *Oryzias latipes. Zoolog Sci* **20**, 1279–1291 (2003).
11. Kamiya, T. *et al.* A trans-species missense SNP in Amhr2 is associated with sex determination in the tiger pufferfish, *Takifugu rubripes* (fugu). *Plos Genet* **8**, e1002798 (2012).
12. Li, M. *et al.* A tandem duplicate of anti-mullerian hormone with a missense SNP on the Y chromosome is essential for male sex determination in nile tilapia. *Oreochromis niloticus. Plos Genet* **11**, e1005678 (2015).
13. Roberts, N. B. *et al.* Polygenic sex determination in the cichlid fish *Astatotilapia burtoni*. *BMC Genomics* **17**, 835 (2016).
14. Rondeau, E. B. *et al.* Genomics of sablefish (*Anoplopoma fimbria*): expressed genes, mitochondrial phylogeny, linkage map and identification of a putative sex gene. *BMC Genomics* **14**, 452 (2013).
15. Nakamura, Y. *et al.* Migration and proliferation of primordial germ cells in the early chicken embryo. *Poult Sci* **86**, 2182–2193 (2007).
16. Zust, B. & Dixon, K. E. Events in germ-cell lineage after entry of primordial germ-cells into genital ridges in normal and uv-irradiated xenopus-laevis. *J Embryol Exp Morph* **41**, 33–46 (1977).

17. Nakamura, M., Kobayashi, T., Chang, X. T. & Nagahama, Y. Gonadal sex differentiation in teleost fish. *J Exp Zool* **281**, 362–372 (1998).
18. Siegfried, K. R. & Nusslein-Volhard, C. Germ line control of female sex determination in zebrafish. *Developmental Biology* **324**, 277–287 (2008).
19. Saito, D. *et al*. Proliferation of germ cells during gonadal sex differentiation in medaka: Insights from germ cell-depleted mutant zenzai. *Developmental Biology* **310**, 280–290 (2007).
20. Kobayashi, T. *et al*. Two DM domain genes, DMY and DMRT1, involved in testicular differentiation and development in the medaka. *Oryzias latipes*. *Dev Dyn* **231**, 518–526 (2004).
21. Haugen, T. *et al*. Sex differentiation in Atlantic cod (*Gadus morhua* L.): morphological and gene expression studies. *Reprod Biol Endocrin* **10**, 47 (2012).
22. Saillant, E. *et al*. Sexual differentiation and juvenile intersexuality in the European sea bass (*Dicentrarchus labrax*). *J Zool* **260**, 53–63 (2003).
23. Herpin, A. *et al*. Inhibition of primordial germ cell proliferation by the medaka male determining gene Dmrt I bY. *BMC Dev Biol* **7**, 99 (2007).
24. Fujimoto, T. *et al*. Sexual dimorphism of gonadal structure and gene expression in germ cell-deficient loach, a teleost fish. *PNAS* **107**, 17211–17216 (2010).
25. Wargelius, A. *et al*. Dnd knockout ablates germ cells and demonstrates germ cell independent sex differentiation in Atlantic salmon. *Sci Rep-Uk* **6**, 21284 (2016).
26. Goto, R. *et al*. Germ cells are not the primary factor for sexual fate determination in goldfish. *Dev Biol* **370**, 98–109 (2012).
27. Cutting, A., Chue, J. & Smith, C. A. Just how conserved is vertebrate sex determination? *Dev Dyn* **242**, 380–387 (2013).
28. Matsuda, M. & Sakaizumi, M. Evolution of the sex-determining gene in the teleostean genus. *Oryzias*. *Gen Comp Endocr* **239**, 80–88 (2016).
29. Herpin, A. & Schartl, M. Plasticity of gene-regulatory networks controlling sex determination: of masters, slaves, usual suspects, newcomers, and usurpators. *EMBO Rep* **16**, 1260–1274 (2015).
30. Yano, A. *et al*. The sexually dimorphic on the Y-chromosome gene (sdY) is a conserved male-specific Y-chromosome sequence in many salmonids. *Evol Appl* **6**, 486–496 (2013).
31. Pandian, T. J. & Koteeswaran, R. Ploidy induction and sex control in fish. *Hydrobiologia* **384**, 167–243 (1998).
32. Penman, D. J. & Piferrer, F. Fish gonadogenesis. part I: Genetic and environmental mechanisms of sex determination. *Rev Fish Sci* **16**, 16–34 (2008).
33. Ottera, H. *et al*. Induction of meiotic gynogenesis in Atlantic cod, *Gadus morhua* (L.). *J Appl Ichthyol* **27**, 1298–1302 (2011).
34. Whitehead, J. A., Benfey, T. J. & Martin-Robichaud, D. J. Ovarian development and sex ratio of gynogenetic Atlantic cod (*Gadus morhua*). *Aquaculture* **324**, 174–181 (2012).
35. Ghigliotti, L. *et al*. Karyotyping and cytogenetic mapping of Atlantic cod (*Gadus morhua* Linnaeus, 1758). *Anim Genet* **43**, 746–752 (2012).
36. Garcia-Souto, D., Troncoso, T., Perez, M. & Pasantes, J. J. Molecular Cytogenetic Analysis of the European Hake *Merluccius merluccius* (Merlucciidae, Gadiformes): U1 and U2 snRNA Gene Clusters Map to the Same Location. *Plos One* **10**, e0146150 (2015).
37. Star, B. *et al*. Genomic characterization of the Atlantic cod sex-locus. *Sci Rep-Uk* **6**, 31235 (2016).
38. Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178–192 (2013).
39. Summers, M. F. *et al*. Nucleocapsid zinc fingers detected in retroviruses: EXAFS studies of intact viruses and the solution-state structure of the nucleocapsid protein from HIV-1. *Protein Sci* **1**, 563–574 (1992).
40. Williams, M. C., Gorelick, R. J. & Musier-Forsyth, K. Specific zinc-finger architecture required for HIV-1 nucleocapsid protein's nucleic acid chaperone function. *PNAS* **99**, 8614–8619 (2002).
41. Guerrerio, A. L. & Berg, J. M. Design of single-stranded nucleic acid binding peptides based on nucleocapsid CCHC-box zinc-binding domains. *J Am Chem Soc* **132**, 9638–9643 (2010).
42. Benhalevy, D. *et al*. The human CCHC-type zinc finger nucleic acid-binding protein binds G-rich elements in target mRNA coding sequences and promotes translation. *Cell Rep* **18**, 2979–2990 (2017).
43. Mitra, M. *et al*. The N-terminal zinc finger and flanking basic domains represent the minimal region of the human immunodeficiency virus type-1 nucleocapsid protein for targeting chaperone function. *Biochemistry* **52**, 8226–8236 (2013).
44. Michalek, J. L., Besold, A. N. & Michel, S. L. J. Cysteine and histidine shuffling: mixing and matching cysteine and histidine residues in zinc finger proteins to afford different folds and function. *Dalton T* **40**, 12619–12632 (2011).
45. Laity, J. H., Lee, B. M. & Wright, P. E. Zinc finger proteins: new insights into structural and functional diversity. *Curr Opin Struc Biol* **11**, 39–46 (2001).
46. Zhu, L. Y. *et al*. Sexual dimorphism in diverse metazoans is regulated by a novel class of intertwined zinc fingers. *Gene Dev* **14**, 1750–1764 (2000).
47. de Luis, O., Lopez-Fernandez, L. A. & del Mazo, J. Tex27, a gene containing a zinc-finger domain, is up-regulated during the haploid stages of spermatogenesis. *Exp Cell Res* **249**, 320–326 (1999).
48. Ma, K. *et al*. Charactering the ZFAND3 gene mapped in the sex-determining locus in hybrid tilapia (*Oreochromis spp*.). *Sci Rep-Uk* **6**, 25471 (2016).
49. Bradley, K. M. *et al*. An SNP-based linkage map for zebrafish reveals sex determination loci. *G3-Genes Genom Genet* **1**, 3–9 (2011).
50. Reichwald, K. *et al*. Insights into sex chromosome evolution and aging from the genome of a short-lived fish. *Cell* **163**, 1527–1538 (2015).
51. Rondeau, E. B., Laurie, C. V., Johnson, S. C. & Koop, B. F. A. PCR assay detects a male-specific duplicated copy of anti-mullerian hormone (amh) in the lingcod (*Ophiodon elongatus*). *BMC Res Notes* **9**, 230 (2016).
52. Hattori, R. S. *et al*. A Y-linked anti-mullerian hormone duplication takes over a critical role in sex determination. *PNAS* **109**, 2955–2959 (2012).
53. Meric, C. & Goff, S. P. Characterization of Moloney murine leukemia virus mutants with single-amino-acid substitutions in the Cys-His box of the nucleocapsid protein. *J Virol* **63**, 1558–1568 (1989).
54. Urbaneja, M. A., McGrath, C. F., Kane, B. P., Henderson, L. E. & Casas-Finet, J. R. Nucleic acid binding properties of the simian immunodeficiency virus nucleocapsid protein NCp8. *J Biol Chem* **275**, 10394–10404 (2000).
55. Hashimoto, H. *et al*. Crystal structure of zinc-finger domain of Nanos and its functional implications. *EMBO Rep* **11**, 848–853 (2010).
56. Mark-Danieli, M. *et al*. Single point mutations in the zinc finger motifs of the human immunodeficiency virus type 1 nucleocapsid alter RNA binding specificities of the gag protein and enhance packaging and infectivity. *J Virol* **79**, 7756–7767 (2005).
57. Johnsen, H., Tveiten, H., Torgersen, J. S. & Andersen, O. Divergent and sex-dimorphic expression of the paralogs of the Sox9-Amh-Cyp19a1 regulatory cascade in developing and adult atlantic cod (*Gadus morhua L.*). *Mol Reprod Dev* **80**, 358–370 (2013).
58. Cinalli, R. M., Rangan, P. & Lehmann, R. Germ cells are forever. *Cell* **132**, 559–562 (2008).
59. Gruidl, M. E. *et al*. Multiple potential germ-line helicases are components of the germ-line-specific P granules of *Caenorhabditis elegans*. *PNAS* **93**, 13837–13842 (1996).
60. Kuznicki, K. A. *et al*. Combinatorial RNA interference indicates GLH-4 can compensate for GLH-1; these two P granule components are critical for fertility in *C-elegans*. *Development* **127**, 2907–2916 (2000).

61. Tanaka, S. S. *et al*. The mouse homolog of Drosophila Vasa is required for the development of male germ cells. *Gene Dev* **14**, 841–853 (2000).
62. Gustafson, E. A. & Wessel, G. M. Vasa genes: emerging roles in the germ line and in multipotent cells. *Bioessays* **32**, 626–637 (2010).
63. Malmstrom, M. *et al*. Evolution of the immune system influences speciation rates in teleost fishes. *Nat Genet* **48**, 1204–1210 (2016).
64. Bakke, I. & Johansen, S. D. Molecular phylogenetics of gadidae and related gadiformes based on mitochondrial DNA sequences. *Mar Biotechnol* **7**, 61–69 (2005).
65. Bangera, R., Odegard, J., Nielsen, H. M., Gjoen, H. M. & Mortensen, A. Genetic analysis of vibriosis and viral nervous necrosis resistance in Atlantic cod (*Gadus morhua* L.) using a cure model. *J Anim Sci* **91**, 3574–3582 (2013).
66. Torresen, O. K. *et al*. An improved genome assembly uncovers prolific tandem repeats in Atlantic cod. *BMC Genomics* **18**, 95 (2017).
67. Harris, R. S. *Improved pairwise alignment of genomic DNA* PhD degree thesis, The Pennsylvania State University (2007).
68. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
69. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).
70. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).
71. McKenna, A. *et al*. The Genome Analysis Toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297–1303 (2010).
72. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
73. Nadalin, F., Vezzi, F. & Policriti, A. GapFiller: a de novo assembly approach to fill the gap within paired reads. *BMC Bioinformatics* **13**(Suppl 14), S8 (2012).
74. Kurtz, S. *et al*. Versatile and open software for comparing large genomes. *Genome Biol* **5** (2004).
75. Untergasser, A. *et al*. Primer3-new capabilities and interfaces. *Nucleic Acids Res* **40** (2012).
76. Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform* (2017).
77. Russell, J. S. A. D. *Molecular cloning: A laboratory manual third edition*. 2344 pp (Cold Spring Harbor Laboratory Press, 2001).
78. Sovic, I. *et al*. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nat Commun* **7**, 11307 (2016).
79. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290–295 (2015).
80. Trapnell, C. *et al*. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511–515 (2010).
81. Dobin, A. *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
82. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
83. Haas, B. J. *et al*. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8**, 1494–1512 (2013).
84. Sali, A. & Blundell, T. L. Comparative protein modeling by satisfaction of spatial restraints. *J Mol Biol* **234**, 779–815 (1993).
85. Morris, A. L., MacArthur, M. W., Hutchinson, E. G. & Thornton, J. M. Stereochemical quality of protein structure coordinates. *Proteins* **12**, 345–364 (1992).

Acknowledgements

Thanks to Kim Præbel at The Arctic University of Norway (UiT) for providing DNA samples from haddock, Arctic cod, Greenland cod, polar cod and burbot. Thanks also to Nicola Barson for her valuable, critical review of the manuscript.

Author Contributions

S.L., M.P.K., O.A. designed the research. T.G.K. performed the genome assembly and analysis. S.L., M.P.K. and O.A. evaluated the analysis. S.L. and M.P.K. designed the experiments. T.A., K.H., M.P.K. carried out the experiments. M.C.D.R. performed protein modeling. T.G.K., O.A., M.P.K. and S.L. wrote the paper. All authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-36748-8.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Errata list

PhD candidate: Graceline Tina Kirubakaran
Thesis: Thesis number: 2019:102
Date: 3.12.2019

Side	Line	Original text	Corrected text
Paper 2, Page 12	Availability of data and materials	The datasets generated and used during the current study, gadMor_Celtic, repeat library and all additional files are stored at NMBU dataverse.no and can be accessed using the following link https://doi.org/10.18710/EVWJVK. The raw nanopore reads used to generate gadMor_Celtic are available at European Nucleotide Archive under accession ID PRJEB35290.	The datasets generated and used during the current study, gadMor_Celtic, repeat library and all supplementary files files are stored at figshare: doi.org/10.6084/m9.figshare.10252919. The raw nanopore reads and illumina MiSeq reads used to generate gadMor_Celtic are available at European Nucleotide Archive under accession ID PRJEB35290.
Paper 2, Page 13	Acknowledgements		Storage resources were provided by the Norwegian National Infrastructure for Research Data (NIRD, project NS9055K).

ISBN: 978-82-575-1667-3
ISSN: 1894-6402

Norwegian University
of Life Sciences

Postboks 5003
NO-1432 Ås, Norway
+47 67 23 00 00
www.nmbu.no

ANDVORDGRAFISK.NO