NVD
Rit

SERRITGERÐ
Thesis

TØKNIFRÁGREIÐING
Technical Report

UNDIRVÍSINGARTILFAR
Teaching Material

UPPRIT
Notes

PhD Thesis

Sequencing the North Atlantic herring (*Clupea harengus*) genome and development of a genetic stock management tool

Sunnvør Klettskarð í Kongsstovu

Amplexa Genetics®

NVD*Rit* 2019:14

NÁTTÚRUVÍSINDADEILDIN
Faculty of Science and Technology

FRÓÐSKAPARSETUR FØROYA
University of the Faroe Islands

Heiti / Title	***Sequencing the North Atlantic herring (*****Clupea harengus*****) genome and development of a genetic stock management tool*** Kortlegging av silda arvastronginum og menning av arvafrøðiligum stovnsumsitingaramboð
Høvundar / Authors	**Sunnvør Klettskarð í Kongsstovu**
Vegleiðari / Supervisor	Svein-Ole Mikalsen, Fróðskaparsetur Føroya Hans Atli Dahl, Amplexa Genetics
Ábyrgdarvegleiðari / Responsible Supervisor	Svein-Ole Mikalsen, Fróðskaparsetur Føroya
Report Type / Ritslag	*Industrial PhD Thesis, Molecular Life Sciences* Vinnu PhD ritgerð, mýlalívsvísindi
Latið inn / Submitted	08.11.2019

NVD*Rit* 2019:14

Útgevari / Publisher Náttúruvísindadeildin, Fróðskaparsetur Føroya
Bústaður / Address Vestara Bryggja 15, FO 100 Tórshavn, Føroyar
☎ • @ +298 352 550 • nvd@setur.fo

Table of Contents

Abstract

In Faroese waters, quantities of several pelagic fish species have increased over the past few years. One example is the Atlantic herring (*Clupea harengus* L.). Its increase has affected both the fishing industry and economy of the Faroe Islands, as well as increased interest in herring biology. Furthermore, the costs of genomic methods such as genome sequencing have reduced drastically the past few years, resulting in increased use of large-scale sequencing in research. In this study, we have investigated aspects of herring biology and genomics using large-scale genomics and bioinformatical methods. This project has paved the way for the use of sequencing and bioinformatics in research in the Faroe Islands, and its results are summarised below.

The herring genome was sequenced and assembled and then compared with the existing herring assembly, which was published after the initiation of this project. The results indicated that we were able to reproduce the herring assembly, and through merging the two assemblies we generated an improved herring genome assembly. A unified nomenclature and better gene predictions would improve future annotations and would make interspecies comparisons easier.

A manual analysis of the connexin gene family in nine teleosts, including herring, indicated that the annotation of this gene family was poor in teleosts. There were wrongly predicted connexins, nonpredicted connexins, and truncated genes, and furthermore, the naming of the genes was highly inconsistent. Our analysis showed that the genes follow a similar pattern in teleosts and mammals, and by following the rules set by naming committees we suggested new naming for the connexins in teleosts.

Furthermore, we sequenced individual herring at low coverage and identified genetic variations in putative populations. These variations were used in a genome-wide association analysis, where we identified regions on the herring genome that were associated with sex. These regions indicated that herring have a male heterogametic sex determination system. This was the first time a specific sex determination system has been suggested for herring. However, we could not identify a specific sex regulatory gene.

We further used the individual variation to investigate the herring population structure in the Northeast Atlantic Ocean. We investigated four herring stocks that spawn in Faroese, Icelandic, Norwegian, and Shetland waters. These stocks were the Faroese autumn-spawning, Icelandic summer-spawning, Norwegian spring-spawning, and North Sea autumn-spawning herring. The results indicated that these stocks were genetically distinct populations with the exception of the Faroese stock, which could not be clearly distinguished from the Icelandic population. In addition, a genetic panel was developed to assign individual herring to one of these four stocks. The panel was tested but exhibited somewhat mixed results. The Norwegian and North Sea stocks could be distinguished from the other stocks with high accuracy (> 90%). However, the distinction between the Faroese and Icelandic herring was problematic. When the Icelandic and Faroese stocks were combined, we were able to assign the test individuals with an accuracy of 89%. Although further validation of the panel is still required, the panel could be useful in stock management, herring fishery monitoring, and keeping the herring fishery sustainable.

In conclusion, this study produced more knowledge about herring genetics and evolution, and could be useful for keeping herring fisheries sustainable.

Samandráttur á føroyskum (Faroese abstract)

Seinastu árini hava størri mongdir av uppsjóvarfiski verið í føroyskum sjógvi enn áður, til dømis sild (*Clupea harengus* L.). Hetta hevur ført við sær, at stórar broytingar hava verið í føroysku fiskivinnuni. Avreiðingarvirðið á hesum fiski er høgt og samfelagsbúskapurin er vaksin nógv, tí er áhugi í at vita meira um sild.

Í hesi verkætlan hava vit hugt nærri at lívfrøðini hjá sild við at brúka arvafrøðilig háttaløg. Arvafrøðilig háttaløg (so sum sekvensering) eru vorðin bíligari seinastu árini, og tí hava tey fingið størri rúmd í granskingarheiminum. Henda verkætlan hevur verið við til at skapa grundarlag fyri slíkari gransking í Føroyum. Í høvuðsheitum komu vit fram á fylgjandi úrslit.

Arvastrongurin (genomið) hjá sild var lisin og settur saman. Okkara úrslit vóru samanborin við úrslit hjá øðrum granskarum, sum eisini høvdu sett sildaarvastrongin saman. Samanberingarnar vístu, at vit høvdu endurskapt samansetingina av arvastronginum hjá sild. Við at sameina hesar báðar samansetingar, fingu vit eina enn betri samanseting av arvastronginum.

Góðskan á samansetingini varð kannað við millum annað at greina connexin ílegufamiljuna. Í hesum sambandi komu vit fram á annað áhugavert úrslit. Vit funnu ósamsvar millum frámerking av hesi ílegufamilju í ymiskum fiskasløgum. Tí hugdu vit nærri at frámerking av hesi ílegufamilju í 9 ymiskum fiskasløgum. Henda kanning vísti, at tað eru nógvir feilir í frámerkingini av hesi ílegufamiljuni í fiski, og at tað er ikki samsvar ímillum, hvussu ílegurnar vóru navngivnar í teimum ymisku fiskasløgunum. Við at fylgja reglum hjá navnagevingarnevndum, komu vit við einum tilmæli um, hvussu hesar ílegur áttu at verið navngivnar.

Arvastrongurin hjá einstøkum sildum var eisini lisin og arvalig avvik (SNPar) funnin. Síðan varð leitað eftir sambandi ímillum lívfrøðiligar funktiónir og hesar SNPar. Fyrst leitaðu vit eftir sambandi ímillum SNPar og kyn, tá funnu vit seks øki á sildaarvastronginum, ið kunnu setast í samband við kynið á fiskinum. Tað vísti seg, at um sildin hevði einsykin avvik á hesum økjum, vóru tær kvennkyn, meðan kallfiskarnir høvdu hinsykin avvik. Tó var ikki møguligt at siga, hvør ílega á hesum økjum ávirkar kynið á sildini.

Eisini nýttu vit arvalig avvik til at finna arvaligan mun millum fýra sildastovnar, ið gýta í føroyskum, íslendskum, norskum og hetlendskum sjógvi. Hesir sildastovnar eru tann føroyska heystgýtandi sildin, íslendska summargýtandi sildin, norðhavssildin og norðsjóvarsildin. Úrslitini vístu, at íslendski, norski og hetlendski stovnarnir eru arvaliga ólíkir, meðan tann

føroyski stovnurin líkist sera nógv íslendska stovninum, men ikki hinum. Vit framleiddu eitt ‘SNP panel’ til at áseta hvørjum stovni ein sild kemur úr. Hetta panelið var roynt, men neyvleikin var ikki góður, tí trupult var at síggja mun á føroysku sildunum og íslendsku sildunum. Tá vit løgdu íslendsku og føroysku sildirnar í sama bólk, kundu vit við 89% neyvleika áseta, hvørjum av teimum trimum stovnunum ein sild var úr. Hetta arbeiði er ikki liðugt, men úrslitini kunnu nýtast í stovnsumsiting av sild og skipan av burðardyggum sildafiskiskapi við Føroyar.

Acknowledgements

I would not have been able to complete this project without help, and therefore, I want to thank everyone who has helped me in one way or other during this project. In the Faroes, we have a saying that goes, ‘*Ongin nevndur, ongin gloymdur*’ (Nobody mentioned, nobody forgotten), which I think is suitable here. However, a few people have been especially helpful, and they deserve a personal thanks.

Firstly, special thanks go to my supervisors Svein-Ole Mikalsen and Hans Atli Dahl for all their hard work, help, and guidance. In addition, I thank them for taking time out of their busy schedules to give me feedback and advise.

Secondly, I want to thank our collaborators at the Faroe Marine Research Institute (FAMRI), Eydna í Homrum and Jan Arge Jacobsen, for their helpful discussions, feedback, and advice, as well as for providing herring samples. I also want to thank Jens Arni Thomassen and Poul Vestergaard from FAMRI for handling the samples.

For providing herring samples, I want to thank Niklas R. Jacobsen, Hilmar Johannesen and the rest of the crew onboard ‘Katrin’, Mannbjørn í Grund and the rest of the crew onboard ‘Sildin’, Bartal Vang and the rest of the crew onboard ‘Grani’, Leanna Henderson, Paul Macdonald and Mark Hamilton from the NAFC Marine Centre in Shetland, and Guðmundur J. Óskarsson from the Marine and Freshwater Research Institute Iceland.

Additionally, I wish to thank Paul Flicek and the Flicek research group from EMBL-EBI for my stay there, bioinformatical help, advice, guidance, and access to computer resources. For their help with proofreading, statistics, population genetics, and laboratory work, I thank Gunnhild Dahl Niclasen, Hannes Gislason, Thomas Damm Als, and Line Larsen.

Furthermore, I want to extend my gratitude to Granskingarráðið, Fiskivinnugransking, Felagið Nótaskip, Statoil Føroyar, and Innovationsfonden for funding the project.

Lastly, I want to thank my wonderful family and friends for their support and understanding, but most of all I want to thank my boyfriend, Tummas T. Tómasson for his unlimited support and love, as well as for keeping me sane during these past years.

List of abbreviations

BUSCO	Benchmarking universal single-copy orthologs
cDNA	Complementary deoxyribonucleic acid
CLA	Chromosome level assembly
DKK	Danish krone
DNA	Deoxyribonucleic acid
EEZ	Exclusive economic zone
ESD	Environmental sex determination
EU	European Union
FAMRI	Faroe Marine Research Institute
FASH	Faroese autumn-spawning herring
FRC	Feature response curves
Gb	Gigabases
GSD	Genetic sex determination
GWAS	Genome wide association study
ICES	International Council for the Exploration of the Sea
IMR	Institute of Marine Research, Norway
ISPH	Icelandic spring-spawning herring
ISSH	Icelandic summer-spawning herring
kb	Kilobases
L.	Linnaeus
Mb	Megabases
mRNA	Messenger ribonucleic acid
MSR	Master sex regulation
NASH	Norwegian autumn-spawning herring
NEAFC	North East Atlantic Fisheries Commission
NGS	Next generation sequencing
NSAH	North Sea autumn-spawning herring
NSSH	Norwegian spring-spawning herring
OLC	Overlap-Layout-Consensus
PacBio	Pacific Biosciences
PCR	Polymerase chain reaction
RNA	Ribonucleic acid
SNP	Single nucleotide polymorphisms
SR	Sex region
TAC	Total allowable catches
TGS	Third generation sequencing
VPA	Virtual population analysis

1. Introduction

1.1. Atlantic herring biology

Atlantic herring (*Clupea harengus* L.) is one of the most abundant fish species in the world. The species is an important source of human food and a crucial part of the ecosystem in the Atlantic Ocean. In the Faroe Islands, Atlantic herring are known as 'sild', and constitute approximately 5% of the total export value of the Faroe Islands [1].

Atlantic herring belong to the class Actinopterygii, and more specifically the infraclass Teleostei, the order Clupeiformes, and the family Clupeidae [2]. There are 198 species in the Clupeidae family; the most abundant of which are the species belonging to the genus Clupea. These are the Atlantic herring *(C. harengus),* Pacific herring *(C. pallasii),* and Araucanian/Chilean herring *(C. bentincki)*. These three herring species are vital resources for commercial fisheries. They have a vast habitat range; as the common names suggest, they can be found in the Atlantic and Pacific Oceans. In this thesis, the focus is on the Atlantic herring (hereinafter 'herring'). On the eastern side of the North Atlantic, the herring habitat ranges from Svalbard south to the northern Bay of Biscay, and from South Greenland to Novaya Zemlya in Russia, including the Baltic Sea. On the western side of the North Atlantic, it ranges from the southwest of Greenland to South Carolina [3, 4].

Herring are pelagic fish that can grow up to 40–45 cm long. They have a blue back and silver belly, as depicted in Figure 1.1. Herring have stable separation of sexes in different individuals (gonochorous) with a sex ratio of 50:50, but their sex determination system is unknown (see Section 1.5). Herring are migratory fish that gather in large schools [6]; they migrate between spawning and feeding grounds with the older fish leading the way [7, 8]. Herring are

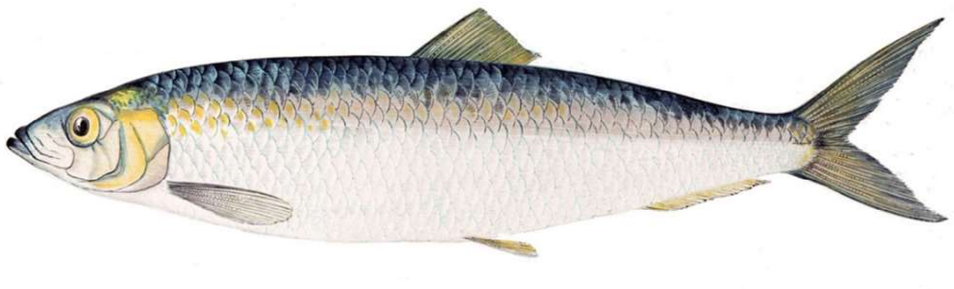


Figure 1.1. Illustration of an Atlantic herring (*Clupea harengus*). Source: [5].

zooplankton filter-feeders, but can switch to actively hunting zooplankton such as copepods depending on prey concentrations [6].

Herring mature at approximately 3–4 years of age and spawn once a year at natal spawning grounds. These spawning grounds vary in bottom substrates, salinity, and temperature [9, 10]. Different spawning grounds and times have given rise to several populations of herring in the North Atlantic [10] (see Section 1.2). During spawning, female herring separate from the school and release their eggs, and the males then release a cloud of milt to fertilise the eggs. Once the eggs are fertilised, they sink to the sea bed and stick to the bottom substrate where they mature [10]. Hatching takes approximately 10–15 days depending on the temperature [11]. The resulting larvae are pelagic and drift with the current to nursery grounds; they are approximately 5–9 mm long when they hatch and have a yolk sac that acts as their energy source for the first days, until they are roughly 12 mm long. Next, they develop mouth parts and can actively feed. At approximately 25–45 mm, they metamorphose into juvenile herring and begin to have the appearance of adult herring. They actively swim and migrate towards shores where they gather in schools, feeding and growing until they reach maturity, when the cycle starts again [7].

1.2. Herring populations in the Northeast Atlantic

Here in Chapter 1 and all subsequent sections and subsections, the word ‘population’ should be understood as genetically distinct or putative genetically distinct populations. In other words, not all populations mentioned have been shown to be genetically distinct but are believed to be.

The population structure of herring consists of populations with specific spawning grounds and times. Some populations are large and migratory, such as the Norwegian spring-spawning herring (NSSH), whereas others are small and local, such as the Faroese autumn-spawning herring (FASH). Thus, herring have a complex population structure with high plasticity, and little is known about the genetic background for the different biological behaviours of spawning and migration.

In 1919, a large herring population named the Atlanto-Scandian herring found in the Norwegian sea was described by Johansen [12]. This large population spawned in several places along the coast of Norway from Lindesnes to Lofoten, as well as on the banks east of the Faroe Islands [13, 14]. This population migrated to feeding areas in the Norwegian Sea after spawning and wintered in an oceanic area between Iceland and the Faroe Islands, before returning to the spawning areas. However, in the 1960s, Atlanto-Scandian herring populations decreased because of heavy fishing and poor recruitment, as well as deteriorating climatic conditions in the Northeast Atlantic [15]. Eventually the population crashed and abandoned their traditional feeding and wintering areas. Figure 1.2 presents the herring catches from 1950 to 2014, and clearly shows the heavy fishing and collapse of the population during the late 60s. After the collapse, the herring remained close to the Norwegian shore after spawning and spent their winters in the fjords, mainly Vestfjorden in northern Norway. Moreover, the spawning area shrunk to only include the coast of Trøndelag and Møre in Norway. A total ban on herring fisheries was implemented and the stock slowly recovered over the next 30 years. The Atlanto-Scandian herring population was a combination of mainly the present NSSH, Icelandic herring populations, and a small population spawning on the banks east of the Faroes [14]. The last two populations have not recovered since the collapse.

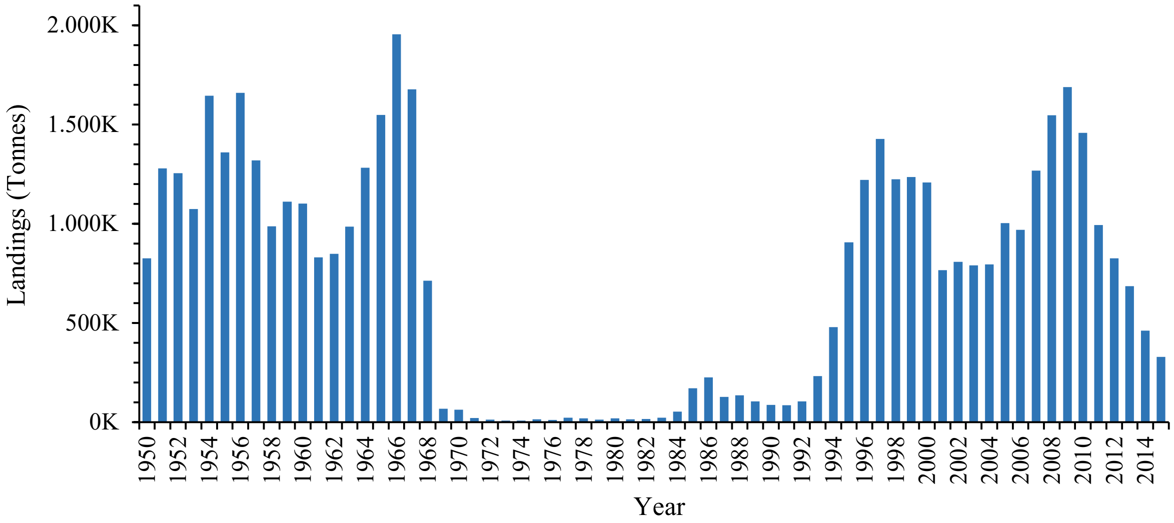


Figure 1.2. Landings of Atlantic herring from the NSSH population. Data from ICES fish assessments, accessed through the SJØMIL database at the Institute of Marine Research (IMR), Norway.

NSSH is the largest population in the Northeast Atlantic. Parts of the historical Atlanto-Scandian migration routes are now being used again by NSSH; they spawn on the coast of Norway and feed in the open ocean between Norway, the Faroe Islands, and Iceland. In addition to NSSH, a Norwegian autumn-spawning herring (NASH) population is found in Norwegian waters [16], as well as several small local populations that mainly spawn in local fjords [17].

Around Iceland, two local herring populations can be found: the Icelandic summer-spawning herring (ISSH) and the Icelandic spring-spawning herring (ISPH). In addition, the NSSH population migrates into Icelandic waters during the summer [18].

Only the local FASH population, also called fjord herring, spawns in Faroese waters. From time to time, small amounts of spring-spawning herring can be found spawning in some fjords and east of the Faroes, but the origin of these populations is unknown. Little is known about the FASH population; for example, the exact spawning locations have not been identified. Nevertheless, they have been found spawning in fjords and on the shelves east of the isles in early autumn, and both nursery and feeding areas are inshore. This fjord herring was observed as early as the 1780s [19]. However, the population is small and there is only a minor fishery on this population. Occasionally, autumn-spawning herring can be found on the banks and shelf area east of the Faroe Islands, but these are not believed to be part of the FASH population. In 1990 and 1991 Jacobsen investigated these herring and concluded, based on biological characteristics such as age composition, growth rate, and vertebrae counts, that they were most likely from the North Sea autumn-spawning (NSAH) population feeding in Faroese waters during summer [20, 21]. Figure 1.3 presents an overview of the herring stocks in the Northeast Atlantic and their migrations.

In addition, herring components exist consisting of several population in the North Sea and Baltic Sea [22, 23]. The Baltic herring are classified as a subspecies of the Atlantic herring.

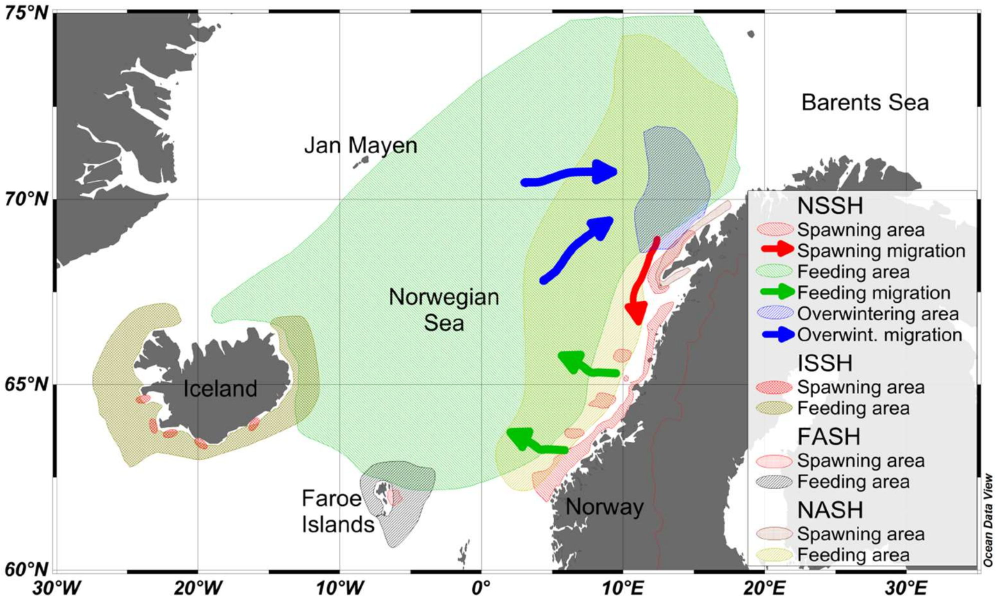


Figure 1.3. Atlantic herring populations in the Northeast Atlantic, their migrations, and interactions. NSSH = Norwegian spring-spawning herring, ISSH = Icelandic summer-spawning herring, FASH = Faroese autumn-spawning herring, and NASH = Norwegian autumn-spawning herring. Figure reprinted with permission from [24] ©Inter-Research 2015.

1.3. Distinguishing between herring populations

During their feeding migrations, the NSSH, ISSH, FASH, and NSAH populations can mix to some degree (Figure 1.3). This can result in mixed stock fisheries where nontarget herring populations are also caught. Distinguishing between the different populations that are caught can be problematic. Morphological, physiological, and biological characteristics are examined to assign individuals to a population, but such observations can be subjective, and the different investigators can assign the same fish to different populations, thereby casting doubt on such methods.

1.3.1. Phenotypic methods

Several phenotypic methods have been used to distinguish between populations, such as the observation of vertebrae, otoliths, and gonads.

Herring vertebral count is negatively influenced by temperature and positively influenced by salinity during the incubation period, and therefore reflects the spawning time and environment. Because the vertebral count of herring that are spawned and incubated at different temperatures and salinity differs, it can be used to distinguish the different herring populations [25]. However, because this method is highly sensitive to temperature variations, it is not always accurate [26]. Nevertheless, it does give an indication of the spawning environment and can be useful together with other methods [27]. The method only gives an average vertebral count from a sample and cannot be used to determine the origin/type of a specific herring in the sample.

Otoliths are bones found in the inner ear of herring and other teleost fish; they grow as the fish grows. Because summer and winter growth differ, a narrow hyaline winter band and a wider opaque summer band are laid down in the otolith during a year's growth; thus, the annuli seen in otoliths represent years. The centre represents the first year of life, and the second ring the second year of life, and so on [28]. Because the centre or the nuclei of the otolith is laid down in the first year, it can tell us about the environment at spawning; fish spawned early in the year (spring spawners) have an opaque nuclei, whereas fish spawned late in the year (summer and autumn spawners) have a hyaline nuclei [29]. These differences can be seen in Figure 1.4. The structure of the growth rings (microstructure) reveals the growth rate of the fish; larger increments indicate fast growth and smaller increments indicate slow growth. Because populations can have different growth rates, otolith microstructures can be used to differentiate populations [30]. The first summer growth (the width of the first winter ring, Figure 1.4a and b) can be used to determine the origin of herring found in Faroese waters during summer. In Figure 1.4a, the narrower width indicates a cooler environment and slower growth in the first summer after spawning, which is typical for an NSSH herring growing in the Barents Sea. This can be compared with the otolith in Figure 1.4b, which has a

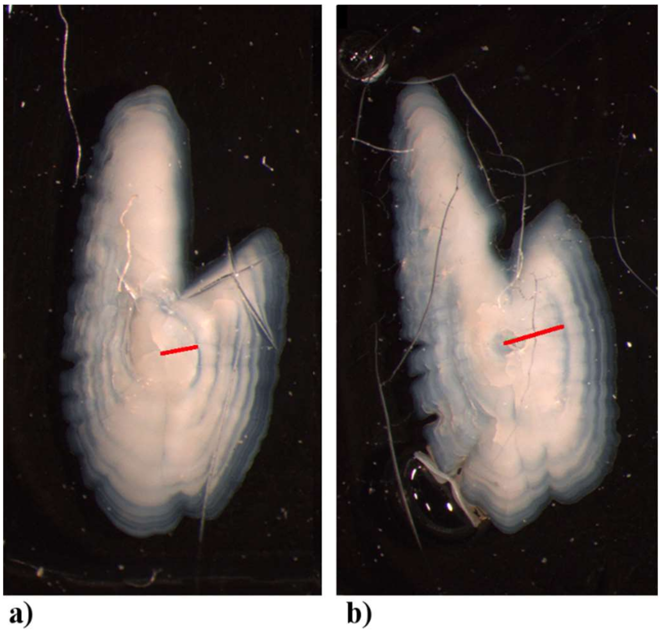


Figure 1.4. Atlantic herring otoliths with a) an opaque nucleus and b) a hyaline nucleus. The spawning type can be determined by the otolith nucleus and by the width of the first winter ring, shown as red lines. Pictures courtesy of the Faroe Marine Research Institute.

wider and faster growth during the first summer, typical for a herring growing in a Faroese fjord.

Otolith chemistry can also reveal the nursery grounds individuals are from, because the local elements and compounds are incorporated into the otolith as the fish grows [31]. In addition, the outline of the otolith can reveal different populations of herring [32] because it is affected by environmental factors (temperature, body growth, and food quantity) and genetic factors [33-36].

Another phenotypic method for discriminating populations is to investigate the maturity stage of the gonads. When herring mature, their gonads enter a maturation cycle with eight stages: stage 1 = immature; stages 2–5 = maturing or pre-spawning; stage 6 = spawning; stage 7 = spent; and stage 8 = resting [37]. The maturity stage at the time of being caught can be compared with the spawning time of the expected populations to assign the herring to a population. However, there are times of the year when the maturity stages of herring from different populations can appear the same, making it difficult to assign them to a population. For example, in late summer, the summer spawners have finished spawning and entered the resting stage, and the spring spawners have not started maturing yet and are also in the resting stage [38]. At the Faroe Marine Research Institute (FAMRI), the maturity stage method is used in conjunction with the opaque/hyaline otolith nuclei method to assess mixed catches (personal communication, Jan Arge Jacobsen and Eydna í Homrum).

These phenotypic methods have been able to distinguish between populations to a varying degree. A downside of these characteristics is that they are affected by the environment and the observer, and furthermore, the environment being in a state of continuous change reduces the accuracy of some of these methods.

1.3.2. Genetic methods

According to the Dictionary of Biology [59], genetics is the study of heredity and variation. The beginning of genetics is often said to be in 1866 when Gregor Mendel published his findings about how traits are inherited in peas [39], although it took decades before the results were noted among scientists. However, the term genetics was not officially used until 1905, when William Bateson coined this term [40]. During the next century, many important discoveries were made, which resulted in genetics becoming a discipline with techniques and methods used far beyond its original borders. A few of these discoveries are listed in Box 1:

Box 1. Some important discoveries in the field of genetics post-Mendel.

- 1910: T. H. Morgan showed that chromosomes carry genetic information [41].
- 1931: B. McClintock and H. Creighton showed that crossing over is the cause of recombination [42].
- 1941: G. W. Beadle and E. L. Tatum showed that genes code for proteins [43].
- 1944: O. Avery, C. MacLeod, and M. McCarty confirmed DNA as the genetic material [44].
- 1951: B. McClintock discovered transposons, showing that DNA is dynamic [45].
- 1953: R. Franklin, J. Watson, and F. Crick showed that the DNA structure is a double helix [46].
- 1961: The understanding of the triplet nature of the genetic code [47].
- 1972–1974: Primitive DNA sequencing methods were developed [4, 48].
- 1971: The principle of targeted DNA amplification (later known as PCR) was demonstrated [49].
- 1977: Dideoxy sequencing of DNA was developed by F. Sanger [50].
- 1985: A. Jeffreys published the DNA fingerprinting method [51].
- 1997: Dolly the sheep was cloned at the Roslin Institute by I. Wilmut and his colleagues [52].
- 2001: First draft sequences of the human genome were released simultaneously by the Human Genome Project (HUGO) and Celera Genomics [53, 54].
- 2008: First human sequenced with Next-Generation sequencing (NGS) technology [55].
- 2009: Single molecule long read sequencing (Third-Generation sequencing; TGS) was developed by Pacific Biosciences (PacBio) [56].

Today, the use of genetics and genetic tools has expanded enormously. We can study evolution, starting from ancient organisms to every branch of the tree of life using genetics and genomics, as long as DNA can be obtained from these organisms [57]. Moreover, we can identify disease-causing mutations in an individual's genome and treat them accordingly [58]. Animals can be genetically modifying to express desirable traits, and this year the first (claimed) genetically modified humans were born [59], which created great controversy in the scientific community.

In this genomic era, population genetics has also evolved. With high throughput sequencing and genotyping technology becoming cheaper every year, large studies with many populations and samples are both possible and affordable. These studies produce a significant amount of data, which must be processed and analysed. This requires advanced statistics, powerful computers, and computer skills (*i.e.,* bioinformatics; see Section 1.6).

Using genetics to differentiate between herring populations is possible because they have different spawning times and places, which results in genetic difference. Some populations have also adapted to special environments, such as the low salinity in the Baltic. These phenotypic adaptions are based on genetic adaption and can explain the genetic differences between populations [22].

Microsatellites are loci in an organism's genome that consist of short tandem repeats of nucleotides. The repeat units in these tandem repeats vary in size from one nucleotide to six, depending on the loci. The number of times the repeat units are repeated varies between individuals, which is inherited; therefore, these microsatellites have a similar number of repeats in closely related individuals [60]. However, two alleles in the same individual can show great variations, depending on the parents. Thus, microsatellites can be used to distinguish between populations. Microsatellites have been used in the study of herring population structures; for example, to identify the population structures of Alaskan Pacific herring [61] and Atlantic herring in the North and Baltic Seas as well as the Skaggerak [62].

Single nucleotide polymorphisms (SNPs) are variations of only one nucleotide, as the name suggests. These SNPs can start out as germ line *de novo* mutations that are then passed on to offspring. After several generations of genetic drift, selection, and possibly some bottleneck events, this mutation could be present in a measurable part of the population and regarded as a SNP. As of today, a specific single nucleotide variation must be present in > 1% of the population to be classified as a SNP. If it is present in < 1% of the population, it is generally regarded as a rare genetic variation. SNPs can affect genes or gene expression, or have no affect at all, depending on their nature and position in the genome [63]. Moreover, SNPs can have a higher or lower frequency in a specific subpopulation compared with a different subpopulation. This makes them useful markers in the study of population structure. One SNP might not be enough to distinguish between populations, but they are easy and cheap to genotype, as well as numerous throughout the genome of all organisms. SNPs have, for example, been used to assign herring from mixed fisheries to their origin population [64] and used in the study of the spawning time of herring [65].

With the help of genetics, scientists are starting to unravel the herring population structure. It has not always been possible to establish significant differences between populations, but significant differences have been revealed between Atlantic herring in the Northeast and Northwest Atlantic, as well as among spawning groups in the Northwest Atlantic [66]. Studies have also shown that both the Baltic herring and North Sea herring are genetically distinct from herring in the Northeast Atlantic [67, 68]. In addition, several distinct populations have been found in the Baltic Sea [69-71], but distinguishing between most spawning aggregations in the North Sea has proven difficult (apart from the English Channel population) [68, 72]. One study showed that the Landvikvannet herring (a local Norwegian fjord) was distinct from NSSH, but differences could not be found between other local fjord populations [24]. A few studies have included the small FASH population and none have been able to distinguish it from the other Northeast herring [24, 38, 64]. Only one study [73] showed a difference between ISSH and NSSH, although others have not been able to replicate this [24].

Using genetic markers to distinguish between populations is a powerful tool. Because the genomes of organisms are sequenced and assembled, the specific locations of these markers can also help explain the molecular mechanism behind the local adaption of different populations. However, producing these high-quality assemblies has its challenges (see subsection 1.6.1).

1.4. Fisheries management of Atlantic herring

In fisheries management, fish populations are divided into stocks, and the catch advice or total allowable catch (TAC) is usually given by stock. These stocks are usually self-contained biological populations, but sometimes may be combinations of biological populations for practical reasons. Reasons for this include that it is simply more practical to manage two populations jointly or that the population structure is not known within the stock [74].

1.4.1. Fisheries management in the Faroe Islands

In the Faroe Islands, a licence is required for commercial fishing. This licence consists of a harvesting licence (veiðiloyvi) and a fishing licence (fiskiloyvi). The harvesting licence allows a vessel to fish in Faroese and international water, whereas the fishing licence specifies the species, quantity, and where and when a vessel is allowed to fish. A vessel can have several fishing licences, but they only last 1 year or season. The conditions of the fishing licenses vary from year to year, mostly by the quantity allowed to be caught [75].

This quantity is controlled by two different systems: the fishing day system and the quota system. The fishing day system is used for vessels fishing demersal species in Faroese waters, such as saithe, cod, haddock, blue whiting, redfish, tusk, and ling. The vessels are organised into groups that receive a certain number of fishing days, which are split between the vessels in the group. The total number of fishing days is determined by the Faroese parliament every year. The quota system is used for all other fish species and fishing areas. The total quota, in tonnes, for each species and stock is set by the Minister of Fisheries, usually in collaborations with other countries in case of straddling stocks [75]. The Faroese Fisheries Inspection (Vørn) is responsible for monitoring the fishing industry by inspecting catches and landings of individual vessels and the weighing-in of catches [76].

FAMRI, or *Havstovan* as it is known in Faroese, is administratively part of the Ministry of Fisheries. Its role is '..to make studies of the Faroese marine environment and its living resources, and to inform and advise the Faroese authorities and public about these conditions' [77]. FAMRI undertakes annual fisheries surveys and analyses catches from commercial fisheries that are submitted to assessment working groups under the International Council for the Exploration of the Sea (ICES). With these data, FAMRI informs the Faroese government about the state of the fish stocks the Faroese fishing industry utilises, both in Faroese and foreign waters. The government then uses this information when assigning the aforementioned fishing days and quotas. FAMRI also participates within ICES in international scientific assessments of shared fish populations of importance for Faroese fisheries [77].

1.4.2. Management of shared fish stocks

Shared stocks, or straddling stocks, are fish stocks that migrate through more than one country's exclusive economic zone (EEZ). If a straddling stock only occurs in national waters, the TACs are set by the countries in whose EEZs the stock occurs (the Coastal States of that stock), and they manage the stock jointly. If a straddling stock also enters international waters, a regional management body, is also a part of the management. In the Northeast Atlantic the North East Atlantic Fisheries Commission (NEAFC) is the regional management body. Annual meetings are held where the Coastal States set the TAC for the straddling stock and agree on quotas for each Coastal State and set aside a quota for the NEAFC parties. The NEAFC quotas are then distributed by the NEAFC to countries that have a historical claim to the stock. These can be Coastal States and non-Coastal States. Coastal State quotas are fished in national waters, whereas NEAFC quotas are fished in international waters.

These arrangements are sometimes broken for various reasons. An example is the agreement between the EU, Norway, Iceland, Russia and the Faroe Islands concerning the joint management of their shared fish species and stocks (redfish, blue whiting, Atlanto-Scandian herring, mackerel, and Rockall haddock) [78]. Currently, no agreement exists on the shared large pelagic stocks in the Northeast Atlantic (herring, mackerel, and blue whiting), and consequently, the total annual catch exceeds the TAC advised by ICES for these species by up to one third [79-81].

1.4.3. Stock assessment

Advising decision-makers on how much of a stock can be fished is a complex task. Numerous factors, both biological and economic, must be considered. The first step is to provide reliable estimates of the catches, disaggregated into age groups, and the second is to assess the state of the stock. Stock assessments can be compared to accounting, where there are income, expenditure, and a balance (Figure 1.5). In a stock assessment, the income consists of the recruitment of young fish to the stock and growth of the fish in the stock; the expenditure is the fish that die (mortality) or emigrate from the stock; and the mortality is split into two groups, fishing and natural mortality, such as from predation and disease. The result is the stock biomass, which should be in balance if the stock is not overfished [74, 77].

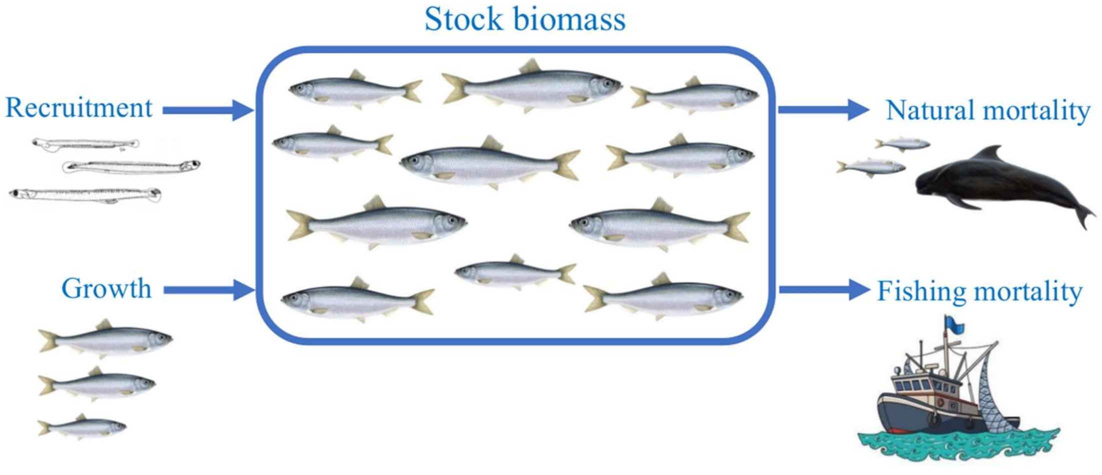


Figure 1.5. Stock assessment. The recruitment and growth of fish increase the stock biomass, whereas fishing mortality and natural mortality (*e.g.,* from predation and disease) decrease the stock biomass.

The stock assessment is based on fish age group distribution, as well as the average weight and length at each age group. Samples from landings and research vessel surveys are investigated; the age group distribution, and the average weight and length at each age group are found. The total weight of fish landed or caught in research vessel surveys are recorded and the age groups and number of individuals are estimated from the sampled fish. The proportion of mature fish in landings and research vessel surveys is also used in stock assessments [77, 82]. Using these data, the fishing mortality and stock biomass can be estimated. Several methods exist for calculating these estimations, but the most used is virtual population analysis (VPA) [83]. The VPA method gives an overview of the state of the stock back in time. The estimates are less accurate for the first 3 years back, but the further back one goes, the more accurate they become. To increase the accuracy, the VPA results are combined with data from research vessel surveys [82]. Based on these stock assessments, advice on how to manage the stock in the future can be provided to decision-makers for setting new quotas.

These stock assessments are estimations based on available data of stocks, therefore, there are uncertainties. For example, unexpected natural phenomena can sometimes cause high recruitment or low natural mortality, resulting in a higher stock biomass than estimated. A lower stock biomass than expected could also result from nonreported fishing mortality or poor recruitment because of, for example, unfavourable conditions during spawning or the early growth period. Nevertheless, these stock assessments are crucial tools in the sustainable management of fish stocks.

1.5. Sex determination in herring and other fish

As mentioned in subsection 1.3.2, genetic variation can be used to investigate biological questions such as the population structure of species. Genetic variation such as SNPs can also be used as markers to identify regions on the genome that have a particular biological function. In this study, we wanted to investigate whether the SNPs used to investigate population structure could also be used to answer other biological questions; therefore, we chose to study sex determination in Atlantic herring.

Herring have stable separation of sexes in different individuals (gonochorous). This is not the case for all species; some have individuals with both sexes (hermaphrodites), and others change sex dependent on age, environmental, and/or social cues [84, 85]. Sex determination systems are highly diverse [86], and Figure 1.6 illustrates a simplified version of this diversity [87].

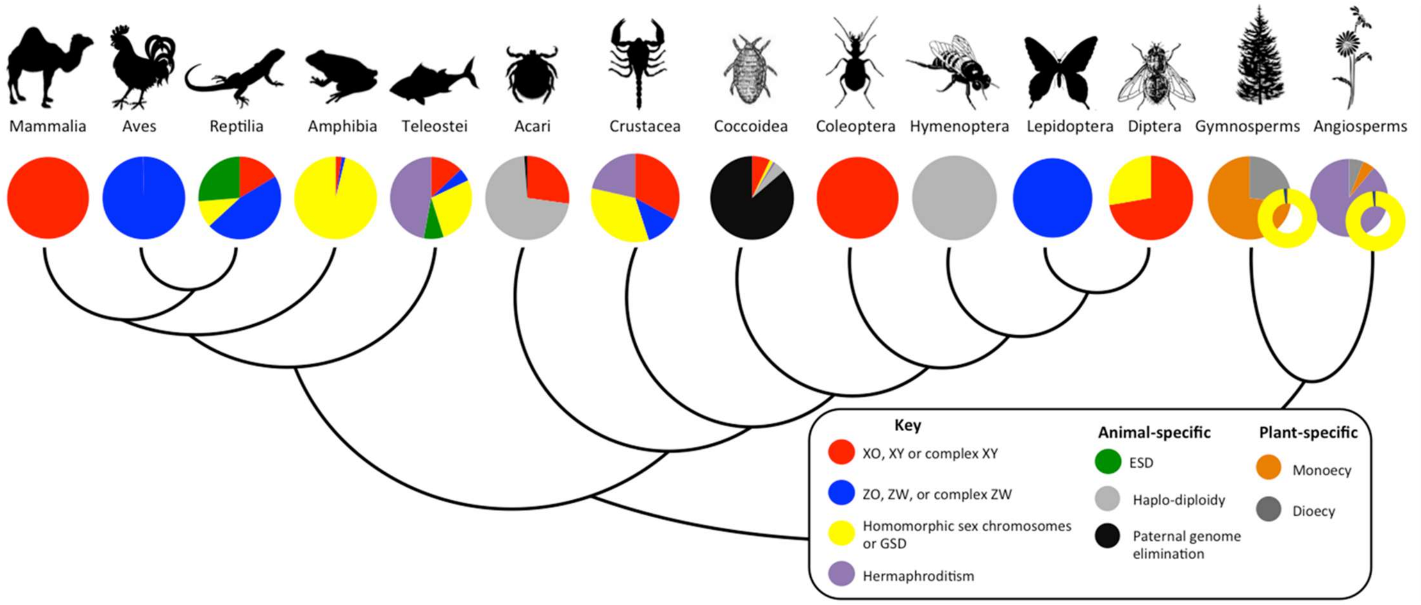


Figure 1.6. Diversity of sex determination systems for representative plant and animal clades. The bubble insert graph for the plant clades represents the relative proportion of species with documented sex chromosomes within plants with separate sexes. Figure from [87].

There are systems where sex is determined by the environment (ESD), but the most common systems are genetic sex determination (GSD) systems. The best-known GSD systems are those with heteromorphic sex chromosomes; in other words, the sex chromosomes differ in size. These systems can be either male heterogametic (XY), female heterogametic (ZW), or more complex with more or fewer sex chromosomes, such as XXY or Z0. Moreover, sex chromosomes can also be homomorphic, meaning they are morphologically identical. With these systems, the sex can be determined by (for example) small regions or SNPs on the sex chromosomes that are specific to the sexes. All mammals have the XY system, whereas all birds have the ZW system [88]. By contrast, teleost fish have a variety of sex determination systems, such as ESD systems, the XY system, ZW system, and more complex polygenic systems (Figure 1.6) [84, 89].

The sex determination system for herring is not known; however, the sex determination system of a few species from the Clupeidae family have been studied. These species are the toli shad (*Tenualosa toli*) and longtail shad (*T. macrura*), which are hermaphrodites, and the Brazilian menhaden (*Brevoortia aurea*), which is male heterogametic with X_1X_2Y sex chromosomes [86, 90]. The Argentine menhaden (*B. pectinate*), Gulf menhaden (*B. patronus*), yellowfin menhaden (*B. smithi*), and Atlantic menhaden (*B. tyrannus*) are also gonochoristic and homomorphic [86, 91], but their sex determination systems are not known. Identifying the sex determination system of Atlantic herring would reveal more about the evolution of sex determination in the Clupeidae family and teleost fish in general.

1.6. Bioinformatics

Bioinformatics is a multidisciplinary field where mathematics, statistics, and computer science are used to analyse large amounts of data to understand biological phenomena [92]. The cost of producing sequencing data has been in continuous reduction the past 20 years, exceeding the expected reductions by Moore's law by several orders of magnitude [93]. Consequently, the number of large sequencing projects has been increasing, with national and international projects focusing on almost every aspect of the tree of life. A few examples include the 1000 Genomes Project [94], the 100,000 Genomes Project [95], the Fish-T1K project [96], and the Earth BioGenome Project [97]. This trend has also reached the Faroe Islands, where the FarGen project aims to sequence the whole population of the Faroe Islands [98]. These large projects, together with all the smaller ones now feasible, create a high demand for bioinformatics and bioinformaticians.

1.6.1. Genome assembly

Sequencing a genome is just the first step, after which much bioinformatical work must be done to use the genome. A desirable aim may be to generate a *de novo* assembly of the genome. Sequencing results in billions of short, unordered DNA fragments (reads) from random positions in the genome, that need to be assembled correctly to represent the sequenced genome. This is a vast and complicated task which is often compared to a jigsaw puzzle. Fortunately, we have powerful computers and assembly software that can perform the assembly task.

The earliest assemblers use an approach called greedy extension. In this greedy approach, read overlaps are found, and the two reads with the best overlap are joined. This process is repeated until a minimum overlap quality threshold is reached [100]. The assembler Phrap used this greedy approach [103], and it was the main assembler used in the Human Genome Project [54].

This method works well for data sets with long and few sequencing reads (*i.e.,* Sanger sequencing data) but becomes problematic when used for data sets with short and numerous sequencing reads (*i.e.,* NGS data).

Most assemblers that use NGS data are based on assembly graphs. Figure 1.7 presents an overview of such a process. In short, the sequencing reads are compared with each other to find their overlaps (Figure 1.7c). The overlaps are recorded in an assembly graph (Figure 1.7d) and continuous sequences or contigs are found by 'walking' along the graph, passing through every note therein (Figure 1.7e). Finally, scaffolding is performed with the help of mate-pair data or other long-range information (Figure 1.7f). This gives the order of the contigs in a scaffold with gaps of known size in between [99]. Assembly graphs have lines called edges that represent the overlaps between reads and can have reads or *k*-mers as the nodes in the graph (Figure 1.7d). *K*-mers are subsequences with the length *k* from (for example) a sequencing read. There are different types of assembly graphs, for example, overlap-layout-consensus (OLC) graphs or De Bruijn graphs. OLC graphs have reads as nodes in the assembly graph and overlaps as edges between the nodes. By contrast, De Bruijn graphs have *k*-mers as nodes. To build a De Bruijn graph, the reads are split into *k*-mers and each unique *k*-mer is added to the assembly graph as nodes. The neighbouring *k*-mers are simultaneously added to the graph and edges are drawn between them. This *k*-mer method does not require the read overlap finding step present in the OLC method because this information is found as the *k*-mers are added to the graph. Therefore, it is much faster than the OLC methods, but does require a large amount of memory. Another disadvantage of the OLC method is that the processing time increases as the coverage increases because there are more reads to compare. This is not the case with De Bruijn graphs because only unique *k*-mers are added [100]. AllPaths-LG and the Celera Assembler are examples of assemblers that use De Bruijn and OLC graphs, respectively [101, 102].

One of the challenges of using short reads for *de novo* assemblies is the repeats in the genome. If repeats are longer than the reads, which is often the case when using NGS data, then the assembly software will have trouble deciphering them. If the software cannot find a solution from the assembly graph, then gaps are introduced at the location of these repeats. In addition to gaps, repeats can also cause misassemblies, such as collapsed repeats or rearrangements, because of misinterpretations of the assembly graph (Figure 1.8) [104].

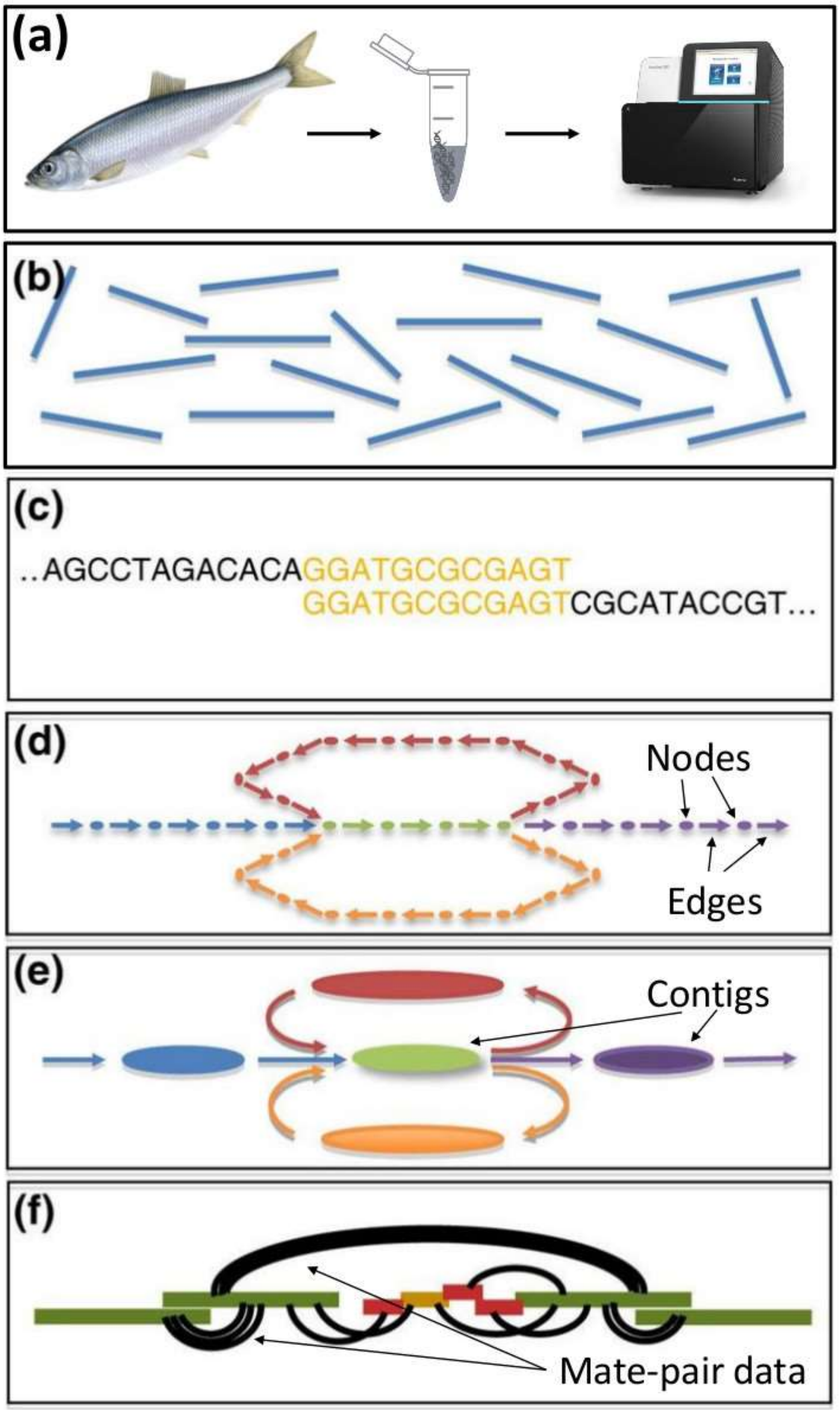


Figure 1.7. Schematic overview of genome assembly. (a) DNA is collected from the biological sample and sequenced, resulting in billions of short reads (b). (c) The short fragments are compared and overlaps are found. (d) The overlaps are captured in a large assembly graph shown as nodes, with edges drawn between. (e) The assembly graph is refined and contigs found. (f) Mate-pair data and other long-range information are used to order and orient the initial contigs into large scaffolds. Figure reprinted and modified from [99].

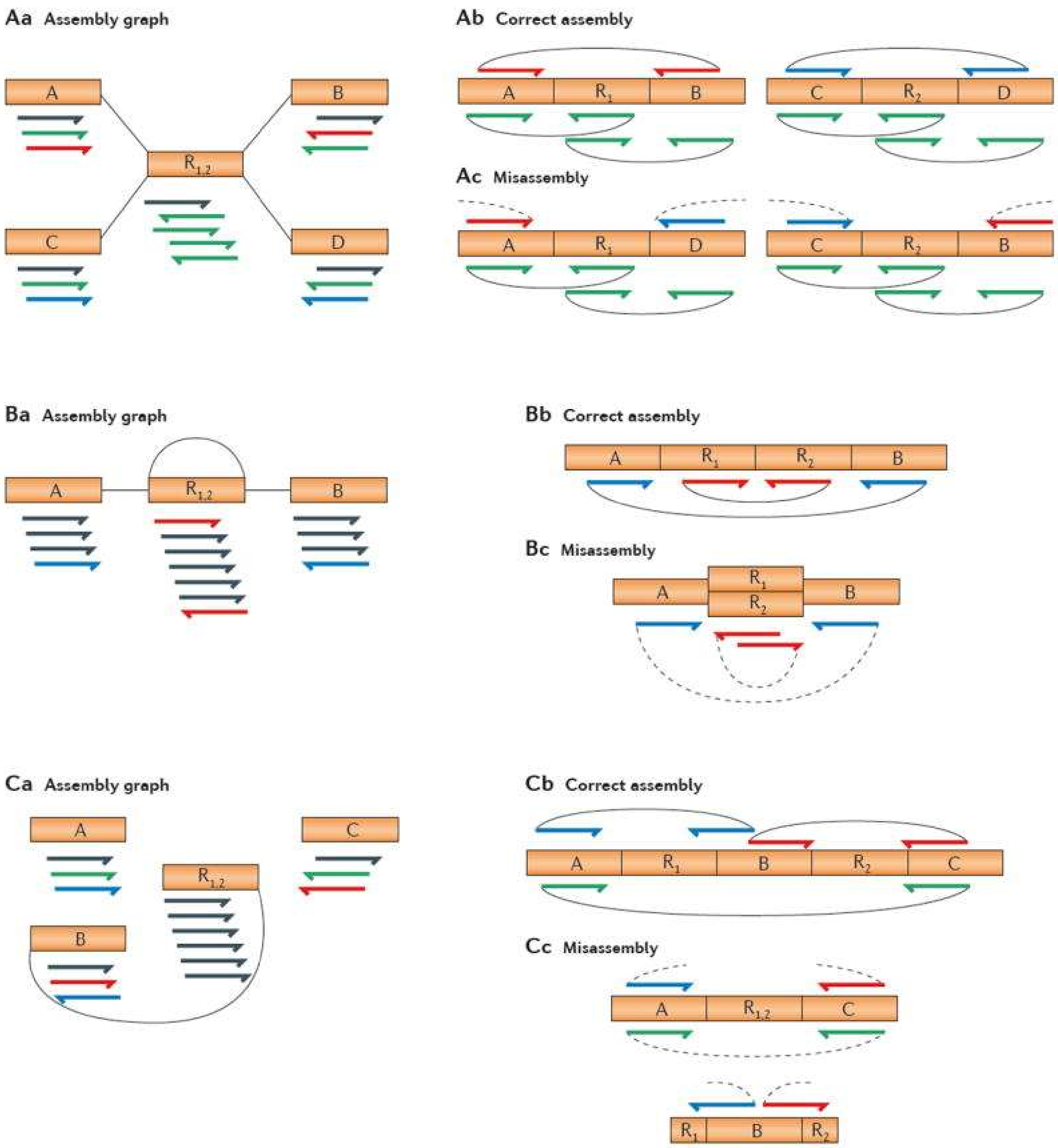


Figure 1.8. Assembly errors caused by repeats. A) Rearrangement assembly error caused by repeats. **Aa)** An example assembly graph involving six contigs, two of which are identical (R_1 and R_2). The arrows shown below each contig represent the reads that are aligned to it. **Ab)** The true assembly of two contigs, showing mate-pair constraints for the red, blue and green paired reads. **Ac)** Two incorrectly assembled chimeric contigs caused by the repetitive regions R_1 and R_2. Note that all reads align perfectly to the misassembled contigs, but the mate-pair constraints are violated. **B)** A collapsed tandem repeat. **Ba)** The assembly graph contains four contigs, where R_1 and R_2 are identical repeats. **Bb)** The true assembly, showing mate-pair constraints for the red and blue paired reads, which are oriented correctly and spaced the correct distance apart. **Bc)** A misassembly that is caused by collapsing repeats R_1 and R_2 on top of each other. Read alignments remain consistent, but mate-pair distances are compressed. A different misassembly of this region might reverse the order of R_1 and R_2. **C)** A collapsed interspersed repeat. **Ca)** The assembly graph contains five contigs, where R_1 and R_2 are identical repeats. **Cb)** In the correct assembly, R_1 and R_2 are separated by a unique sequence. **Cc)** The two copies of the repeat are collapsed onto one another. The unique sequence is then left out of the assembly and appears as an isolated contig with partial repeats on its flank. Reprinted by permission from Springer Nature, Nature Reviews Genetics [104], ©Springer Nature 2011.

Sequencing errors cause branching in the assembly graph, making it more complex. However, NGS data are highly accurate (0.1% errors [105]), and tools are available for the error correction of sequencing reads (*e.g.,* Quake [106]); and many assemblers include an error correction step in their pipeline (*e.g.,* SOAPdenovo [107]). Other challenges of genome assemblies include organism ploidy, gene or whole genome duplications, and the heterozygosity of the genome. Heterozygosity causes the sequencing graph to become more complex. Polyploidy causes similar problems as repeats, but the rearrangements occur between the different copies of the chromosomes. Gene duplications and whole genome duplications are essentially repeats in the genome.

1.6.2. Assembly quality evaluation

Numerous genome assemblers are available, and they can give different results for the same genome, even the same data. Therefore, the ability to compare the quality of assemblies and choose the best assembler for one's genome and data is crucial. If a reference genome is available, the new assembly could be compared to this. However, this is not always the case, and thus, assessing the quality of assemblies can be tricky.

Simple metrics exist such as N50, the number of contigs/scaffolds, and contig/scaffold size that evaluate the size and fragmentation of the assembly. However, they do not necessarily indicate the quality or correctness of the assembly. A study has in fact showed that N50 is negatively correlated with assembly quality [108]. In studies where several assemblies have been compared, such as Assemblathon 1, Assemblathon 2, and GAGE [109-111], several metrics have been used to describe the quality of the assemblies. The results of these studies have shown that different metrics indicate different strengths and weaknesses of the assemblies. Therefore, to evaluate and compare assemblies, it is necessary to use several metrics that indicate the size, fragmentation, completeness and correctness of the assemblies.

To assess the completeness of an assembly, the presence or absence of genes in the assembly can be investigated. This annotation can be performed experimentally by sequencing mRNA, assembling the genes that the mRNAs code for, and aligning this to the assembly, to find the location (as well as introns and exons) of the gene in the assembly. Genes in the assembly can also be predicted by algorithms that scan the assembly for signatures of genes, such as open reading frames and intron–exon boundaries. The success of these predictions is limited by the sensitivity of the algorithms. The software BUSCO is another option, which searches for benchmarking universal single-copy orthologues in the assembly [112].

The correctness of an assembly can be assessed by identifying possible misassemblies; for example, the compressions or expansions caused by repeats (described in Figure 1.8B and C). This can be investigated by aligning paired-end and mate-pair data to the assembly. Paired reads mapping on different scaffolds and coverage differences can indicate possible misassemblies (often called features). Feature response curves (FRC) that capture different types of features can be calculated for assemblies and easily compared to obtain an overall picture of the assemblies' correctness. Software for these calculations exists, such as FRC^{bam} [108]. Software packages also exist for comparisons of genome assemblies that incorporate several different metrics; for example, QUAST-LG and REAPR [113, 114].

1.6.3. The use of genetic variation and bioinformatics to identify subpopulations

Once the genome assembly and gene annotation are available, many biological questions can be investigated. For example, finding genetic variations between individuals from the same species that explain phenotypic traits such as sex, colour, or height. These genetic variations can be found by sequencing individuals, aligning the reads to the genome assembly, and investigating where in the genome the different individuals exhibit differences. This is of course not done manually if the whole genome is being investigated, but using software such as FreeBayes [115] and GATK [116], which call genetic variation.

Additionally, the population structure can be investigated in a similar manner by identifying variations that are specific to subpopulations [117]. These variations can later be used in the opposite direction to assign individuals to subpopulations. This can be highly useful when subpopulations are difficult or labour-intensive to establish from phenotypic traits, or when the phenotypic traits are no longer available; for example, as fish fillets in a shop. Genetic methods can have the advantage of allowing many individuals to be investigated simultaneously. This is dependent on the method used, but most methods can be set up for high throughput using robots; this standardises the assignment to a subpopulation, thereby minimising human errors and variability.

1.6.4. The herring genome

The herring genome has been estimated to be approximately 850 Megabases (Mb) and consist of 26 chromosome pairs [118-121]. At the start of this study, no assembly of the herring genome was available; however, 5 months later, the first draft of the herring genome was published [122]. This assembly was based on a Baltic herring, a subspecies of the Atlantic herring. The assembled size was 808 Mb arranged in 6,915 scaffolds and 73,682 contigs, with an N50 of 1,860 Kilobases (kb) [122].

1.7. Aims of this study

As previously mentioned, Atlantic herring is a highly migratory species with a vast geographical distribution and several populations mixing during parts of the year. This behaviour has made it difficult to elucidate the population structure, which is a critical parameter in the proper management of populations/stocks, as well as a means to avoid overexploitation. Herring is a crucial national and international resource, as well as a part of the ecosystem; therefore, it is imperative to keep the fisheries sustainable. Neglecting to account for population structure in fisheries management can result in overexploitation and the loss of genetic diversity [123]. Knowledge of population structure is required to ensure that the intended population is targeted by fisheries, and to make realistic regulations for fisheries management. Furthermore, population structure is important in the fight against illegal, unreported, and unregulated fishing, as well as the forensic identification of fish and fish products throughout the food processing chain.

According to the fisheries industry, mixtures of herring populations are often present in catches, which was confirmed by [64], and this causes problems related both to economic profit and the sustainable management of herring populations. In addition, the nations that participate in herring fishery in the North Atlantic disagree on the distribution and size of quotas. To protect fragile populations and manage the mixed fishery, a reliable population detection method is required for the industry.

We undertook this study to identify a possible genetic solution to this problem in the industry, as well as to answer interesting biological questions regarding herring biology and evolution.

The aims of this study were as follows:

1. To generate a *de novo* assembly of the Atlantic herring (*Clupea harengus*) genome and evaluate the assembly quality.
2. To identify genetic differentiations (SNPs) between individual herring and four putative herring populations in the Northeast Atlantic.
3. To use these SNPs to test for a simple physiological property (*e.g.,* sex) that supposedly is directly linked with specific variations.
4. To select markers, and create and validate a panel of genetic markers, enabling a cost-efficient and reliable method for discriminating herring populations in catches from the North Atlantic.

2. Summary of included manuscripts

2.1. Manuscript 1: Using long and linked reads to improve an Atlantic herring (*Clupea harengus*) genome assembly. Published in Scientific Reports 9, 17716 (2019).

Manuscript 1 describes a *de novo* assembly of the Atlantic herring genome using short reads, and progressive improvements of the assembly with the help of long reads and linked reads. The assembly was compared with the previously published draft herring genome assembly, which showed that these two assemblies were similar but with improvements in the new assembly. These results showed that the herring genome assembly was reproduceable, thereby validating the herring genome assembly.

2.2. Manuscript 2: Phylogeny of teleost connexins reveals highly inconsistent intra- and interspecies use of nomenclature and misassemblies in recent teleost chromosome assemblies. Revised version accepted in BMC Genomics.

In Manuscript 1, the gap junction protein gene family (also called connexin genes) was used as one of the quality controls in the assemblies. We noted a highly inconsistent use of nomenclature for this gene family and several wrong gene predictions, which can be a problem for automatic annotations. In Manuscript 2, we undertook a broader investigation of the naming of the connexin genes in teleosts. The publicly available connexin gene sequences from teleosts, covering the range of divergence times, were collected and compared. The results showed that the gene family pattern of connexin genes were similar across the analysed teleosts, but the naming of the connexin genes did not reflect this pattern; for example, several nomenclature systems are used, several distinct genes have the same name in a species, and some genes have directly wrong names. This showed that the clear rules for naming orthologous genes in fish and mammals, outlined by nomenclature committees, are not followed.

Ohnologous genes in teleosts were indicated, and a more consistent nomenclature that follows the outlined rules from the nomenclature committees was suggested. Furthermore, we showed that connexin sequences can indicate some errors in two high-quality chromosome assemblies that recently became available.

2.3. Manuscript 3: Identification of male heterogametic sex determining regions on the Atlantic herring *Clupea harengus* genome. Submitted.

Manuscript 3 describes how we identified six regions on the Atlantic herring genome that are associated with sex, using low-coverage whole genome sequencing and a genome-wide association study (GWAS). The majority of SNPs associated with sex were homozygous in female fish and heterozygous in male fish. This indicated male heterogametic sex determination in herring. Possible sex determination genes were investigated but evidence was insufficient for indicating a single gene on these sex regions.

2.4. Manuscript 4: Atlantic herring (*Clupea harengus*) population structure in the Northeast Atlantic Ocean.

In Manuscript 4, the herring population structure was investigated using low coverage sequencing of herring from four herring stocks in and around Faroese waters. SNPs were called and used for population structure analyses and individual assignment. The results showed that all four stocks are genetically differentiated, but cluster-analysis only identified three clusters. The Faroese and Icelandic stocks could not confidently be distinguished, but some evidence existed that these two stocks were not completely panmictic. Assignment of new herring individuals to the putative populations was successful for two of the populations (assignment accuracy > 90%) but less successful for the Faroese and Icelandic stocks (assignment accuracies of 47% and 43%, respectively). However, when samples from these two problematic stocks were pooled, the overall assignment accuracy was 89%.

3. Manuscripts

3.1. Using long and linked reads to improve an Atlantic herring *(Clupea harengus*) genome assembly

í Kongsstovu, S., Mikalsen, S., Homrum, E.í. *et al.* Using long and linked reads to improve an Atlantic herring (*Clupea harengus*) genome assembly. *Sci Rep* 9, 17716 (2019). https://doi.org/10.1038/s41598-019-54151-9

Published as an open access research article in Scientific Reports on the 27th of November 2019.

SCIENTIFIC REPORTS
natureresearch

OPEN

Using long and linked reads to improve an Atlantic herring (*Clupea harengus*) genome assembly

Sunnvør í Kongsstovu[1,2,4*], Svein-Ole Mikalsen[2], Eydna í Homrum[3], Jan Arge Jacobsen[3], Paul Flicek[4] & Hans Atli Dahl[1]

Atlantic herring (*Clupea harengus*) is one of the most abundant fish species in the world. It is an important economical and nutritional resource, as well as a crucial part of the North Atlantic ecosystem. In 2016, a draft herring genome assembly was published. Being a species of such importance, we sought to independently verify and potentially improve the herring genome assembly. We sequenced the herring genome generating paired-end, mate-pair, linked and long reads. Three assembly versions of the herring genome were generated based on a *de novo* assembly (A1), which was scaffolded using linked and long reads (A2) and then merged with the previously published assembly (A3). The resulting assemblies were compared using parameters describing the size, fragmentation, correctness, and completeness of the assemblies. Results showed that the A2 assembly was less fragmented, more complete and more correct than A1. A3 showed improvement in fragmentation and correctness compared with A2 and the published assembly but was slightly less complete than the published assembly. Thus, we here confirmed the previously published herring assembly, and made improvements by further scaffolding the assembly and removing low-quality sequences using linked and long reads and merging of assemblies.

Atlantic herring (*Clupea harengus*) is one of the most abundant fish species in the world and is an important economical and nutritional resource. In 2016, a total of 1,639,760 tons of Atlantic herring were fished worldwide[1]. Herring is especially crucial to the Faroe Islands, where 108,244 tons were fished in 2017, constituting 7.5% of the total value of exported goods that year[2].

The species is a pelagic, highly migratory fish, with a vast geographical distribution. Several populations of Atlantic herring have been identified, spawning in different seasons and sites in the North Atlantic Ocean[3]. Some of the populations mix to a varying degree during their feeding migrations and are only distinguished by morphological, physiological, and biological characteristics, which can be open to interpretations[4]. Identifying populations and the extent of mixed fisheries is vital to keep the fisheries sustainable. Thus, knowledge of the population structure is necessary. Disregard of population structure in fisheries management can lead to overexploitation and result in the loss of genetic variation[5], which may be vital for adaptation in an ocean affected by climate change. Furthermore, knowledge of the population structure can be used to forensically identify fish and fish products throughout the food processing chain, and it assists in the fight against illegal, unreported, and unregulated (IUU) fishing. Genetics is a useful tool in the fight against IUU, as shown in Nielsen *et al.*[6]. Most of the commercially fished species are not model organisms, and therefore, limited genetic information is available for them. A few studies have been performed on herring population genetics, but the ability to distinguish some of the subpopulations has only been partially accomplished[4,7,8]. The availability of the assembled genome for the species in question is the ultimate basis for developing population genetic markers, to be able to map microsatellites, single nucleotide polymorphisms (SNPs), and other polymorphisms. Generally, more variations are expected in the noncoding regions than in coding regions. Therefore, assembling the whole genome rather than just the transcriptome means that more detailed population genetic markers can be developed, increasing the power for separating closely related populations.

[1]Amplexa Genetics A/S, Hoyvíksvegur 51, FO-100, Tórshavn, Faroe Islands. [2]University of the Faroe Islands, Department of Science and Technology, Vestara Bryggja 15, FO-100, Tórshavn, Faroe Islands. [3]Faroe Marine Research Institute, Nóatún 1, FO-100, Tórshavn, Faroe Islands. [4]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK. *email: skik@amplexa.com

	Raw reads		Reads after QC		
Sequencing technology and library type	No. of reads	Bases ≥ Q30	No. of reads	Bases ≥ Q30	Coverage after QC
Illumina - Paired end	668,361,981	78.5%	490,582,474	91.1%	150.0x
Illumina - Mate pair with insert size 4.5 kb*	591,526,598	68.1%	156,135,780	90.4%	26.3x
Illumina - Mate pair with insert size 7 kb*	116,602,405	79.2%	44,016,755	94.4%	8.8x
Illumina - 10x Genomics	363,163,358	61.3%	—	—	78.5x
MinION	1,135,273	—	985,281	—	2.4x

Table 1. Summary of sequencing results. Coverage refers to the coverage of the estimated 850 Mb Atlantic herring genome. Quality control (QC) for paired-end data consisted of quality trimming and adapter removal. QC for mate-pair data consisted of quality trimming and sorting of reads based on presence of adapter in reads. No QC was performed on the 10x Genomics reads as recommended by 10x Genomics. The QC for the MinION reads consisted of alignment to the draft assembly and only aligned reads longer than 500 bp were kept. *When the mate pair library data were investigated bioinformatically, both libraries seemed to have an insert size of 2 kb.

The size of the herring genome is estimated to be approximately 850 megabases (Mb), and it consists of 26 pairs of chromosomes[9–12]. In 2016, Martinez Barrio *et al.* published the first draft of the herring genome[13]. The assembled size was 808 Mb, arranged in 73,682 contigs and 6,915 scaffolds, with a scaffold N50 of 1,860 kilobases (kb). Studies have shown that different assembly approaches may yield different assembly results[14–16]. Furthermore, combining several sequencing technologies can improve genome assemblies[17–19]. Thus, being a species of such ecological, economical, and nutritional importance, we undertook a second assembly using a different combination of sequencing technologies to verify and improve the herring genome assembly and obtain more definitive genomic information of this species. This knowledge is critical for the further study of the herring population structure and genetic variation.

Here, we sequenced the herring genome on an Illumina platform, generating paired-end, mate-pair, and linked (10x Genomics) reads. Long reads were also generated using the Oxford Nanopore Technologies platform, MinION. A *de novo* herring genome was assembled from the short reads and scaffolded using the long and linked-read data. In the last stage, our assembly was merged with the previously published assembly by Martinez Barrio *et al.*[13] (GCF_000966335.1_ASM96633v1; here referred to as the published draft assembly) to create a more accurate genome assembly, shown by comparing the assemblies with multiple quality parameters.

Results

Sequencing and assembly.

A paired-end library and two mate-pair libraries (both approximately 2 kb when investigated bioinformatically) were sequenced along with long (MinION) and linked (10x Genomics) reads. The same individual was sequenced with Illumina technology and on one MinION run. However, the DNA from this individual was too degraded to obtain long reads. Therefore, three additional MinION runs were performed using a fresh sample from a second individual, which resulted in longer reads and higher output. The total output for the four runs was 985,281 reads with an N50 of 8,119 bp. A third individual was sequenced using 10x Genomics technology, to obtain input fragments that were as long as possible. Table 1 presents a summary of the sequencing results.

To generate an improved herring genome assembly, we first generated *de novo* assemblies from the short-read data using the AllPaths-LG and SGA assemblers[20,21] with different parameters (Supplementary Table S1). The assembly with the best summary statistics (*i.e.*, number of contigs, number of scaffolds, and N50) was named A1. This assembly was improved using gap closing software and long and linked reads for scaffolding (see Materials and Methods) resulting in the A2 assembly. Lastly, the A2 assembly was merged with the published draft assembly to obtain the best assembly possible (A3). Table 2 presents the characteristics of these assemblies. For comparison, we generated an alternative assembly using the hybrid assembler, MaSuRCA, which Zimin *et al.*[22] claimed to have equal or superior performance to AllPaths-LG. This resulted in a highly fragmented assembly (74,436 scaffolds and N50 of 28 kb). Thus, in our hands, MaSuRCA did not perform better than AllPaths-LG combined with SSPACE-LongRead[23] and ARCS[24]. The MaSuRCA assembly was not further used in this study.

Did scaffolding with linked and long reads improve the assembly?.

To assess the level of improvement obtained through gap-closing and scaffolding with long and linked reads, we compared the assemblies using QUAST[25]. QUAST is a tool for assessing the quality of genome assemblies and can be used both with and without a reference assembly. Without a reference assembly, QUAST calculates several descriptive summary statistics for the assemblies, which are mostly based on the size and fragmentation of the assemblies (*e.g.*, the number of scaffolds, length of the assembly, N50, and NG50). GC content, Ns per 100 kbp, and predicted rRNA genes are also found by QUAST. Table 2 presents selected QUAST results, and as expected, both the fragmentation and size of the assembly were improved when A1 was scaffolded with long and linked reads, resulting in A2. The number of scaffolds decreased by roughly 38%; both the total length and length of the largest scaffold increased and N50 almost doubled (Table 2). The same trend could be seen in the number and length of contigs. Moreover, the completeness of the assembly improved, and Ns per 100 kbp decreased by 1,077. There were 60 complete rRNA genes in A2, compared with 52 in A1, and 12 partials in A2 compared with 15 in A1 (Table 2).

The completeness of the assemblies was further assessed using Benchmarking Universal Single-Copy Orthologs (BUSCO), which searches for near-universal single-copy orthologs based on evolutionarily-informed expectations of gene content[26]. Different BUSCO sets are used for different groups of organisms, and presently

Metric	A1	A2	A3	Draft
# scaffolds (>= 0 bp)	15,378	9,444	2,419	6,915
# scaffolds (>= 1,000 bp)	15,188	9,334	2,419	6,915
# scaffolds (>= 5,000 bp)	10,057	6,348	1,709	2,267
# scaffolds (>= 10,000 bp)	8,049	5,378	1,573	1,964
# scaffolds (>= 25,000 bp)	5,166	3,798	1,319	1,481
# scaffolds (>= 50,000 bp)	3,252	2,678	1,043	1,131
Total length of scaffolds (>= 0 bp)	702,694,152	729,318,454	790,426,535	807,711,962
Largest scaffold (bp)	2,291,227	3,948,801	13,043,132	13,053,552
Scaffold N50 (bp)	177,425	332,253	1,971,137	1,897,858
# contigs (>= 0 bp)	131,323	112,927	61,451	67,061
Total length of contigs (>= 0 bp)	524,819,960	551,688,118	711,593,948	725,034,955
Largest contig (bp)	169,324	179,560	251,421	245,657
Contig N50 (bp)	6,450	8,441	25,590	25,381
GC (%)	43.07	43.06	44.13	44.11
# Ns per 100 kbp	25,665	24,588	9,995	10,314
# predicted rRNA genes	52 + 15 part	60 + 12 part	57 + 10 part	57 + 10 part

Table 2. Comparison of assemblies A1, A2 and A3 from this study and the published draft assembly. Results from the QUAST analysis, all statistics are based on contigs of size >= 3,000 bp, unless otherwise noted; for example, # contigs (>= 0 bp) includes all contigs.

the set for ray-finned fish includes 4,584 genes. The BUSCO analysis showed the same trend as the QUAST analysis when progressing from assembly A1 to A2. The number of complete BUSCOs increased by 251, fragmented BUSCOs decreased by 172, and missing BUSCOs decreased by 79 (Table 3), indicating a more complete assembly.

The summary statistics in Table 2 are commonly used metrics to compare assemblies, but they only show how fragmented the assemblies are and say little about the completeness and correctness of the assemblies. Furthermore, these traditional metrics do not necessarily indicate which assembly is of the highest quality. In fact, N50 has been shown to be negatively correlated with the quality of an assembly[27].

To assess the assembly correctness, a feature response curve (FRC) was calculated for each assembly. FRC is a metric that, according to the authors Narzisi and Mishra[28], captures the trade-offs between quality and contig size. The analysed features and underlying logics were described by Phillippy *et al.*[29]. In short, a steeper curve indicated an assembly of higher quality. The results from this comparison can be seen in Fig. 1. The FRCs for A1 and A2 diverged at a higher feature threshold, with A2 being steeper.

FRC^{bam} outputs 14 categories of features based on both paired-end and mate-paired data[27]. Features are areas on the assembly that show indications of assembly errors based on the alignment of sequencing reads. Through examination of the different features separately it became obvious that the assemblies had different types of features (*i.e.*, different strengths and weaknesses). We ranked the assemblies for all 14 types of features so that the assembly with the steepest FRC for the specific feature obtained the best ranking (1st), we then summed over all the features types to obtain a ranking of the assemblies based on overall features. This ranking is shown in Table 4, and overall A2 (2nd) was ranked better than A1 (4th). FRCs for the specific features can be seen in Supplementary Figs. S1–S14. As mentioned earlier, the FRC also accounts for the contig size. However, examining only the total number of features, we saw that A1 had 564,464 features, whereas A2 had 544,122, showing a reduction of 3.6%.

BUSCOs (Table 3) did provide an indication of the level of completeness, but we wanted to further inspect the completeness by looking at the connexin (gap junction protein) gene family. Generally, bony fish have approximately 40 recognised connexin genes[30,31]. Most of these genes have their coding sequence in a single exon, greatly facilitating a manual analysis. Additionally, these genes have two conserved regions that are easily recognised across the gene family. From other species, including different bony fish, it is known that some of these genes are located close to each other[30,32]. In this context, the conserved regions might be considered repetitive sequences, which could make these genes more prone to assembly errors.

In our manual analysis of the connexin genes we first identified 51 herring connexin genes from the draft assembly by Martinez Barrio *et al.*[13]. Of these, 49 connexin genes were already predicted and annotated by Martinez Barrio *et al.* and available in GenBank. In addition to the 49 connexin genes, one connexin gene was predicted as a *KAT6B-like* gene, and one connexin gene (called *Cx39.2* or *GJD2like*) was not predicted but found in our searches. Some of the genes found in the draft assembly were believed to be duplicates or triplicates, based on the 98–100% sequence identities (see Table 5 and Supplementary Table S2). Thus, these genes were either very recently duplicated or arose through erroneous assembly, and we consider 46 as a more likely number of functional connexin genes in herring. More details on the analysis of connexin genes in herring and other teleosts can be found elsewhere (Mikalsen SO, Tausen M and í Kongsstovu S, submitted).

Furthermore, we investigated the presence of the connexins in our progressive assemblies A1, A2 and A3 (the latter is described in more details below). There were 3 connexins lacking in A1 (*Cx32.2like*_XM_012828709, *GJA5like*_ XM_012816449, and *GJD3like*_XM_012837668), one of which was found in A2 (*GJA5like*_ XM_012816449). In addition, the *GJD2like*_ XM_012838313 and *GJA5like*_XM_012840593 genes were fragmented in A1 (Table 5). The fragmentation of *GJA5like*_ XM_012840593 was still present in A2, whereas the

BUSCOs	A1	A2	A3	Draft
Complete BUSCOs	3,598	3,849	4,258	4,348
Complete and single-copy BUSCOs	3,473	3,706	4,085	4,176
Complete and duplicated BUSCOs	125	143	173	172
Fragmented BUSCOs	409	237	177	105
Missing BUSCOs	577	498	149	131
Total BUSCO groups searched	4,584	4,584	4,584	4,584

Table 3. Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis of the A1, A2, and A3 assemblies and the previously published draft herring genome assembly.

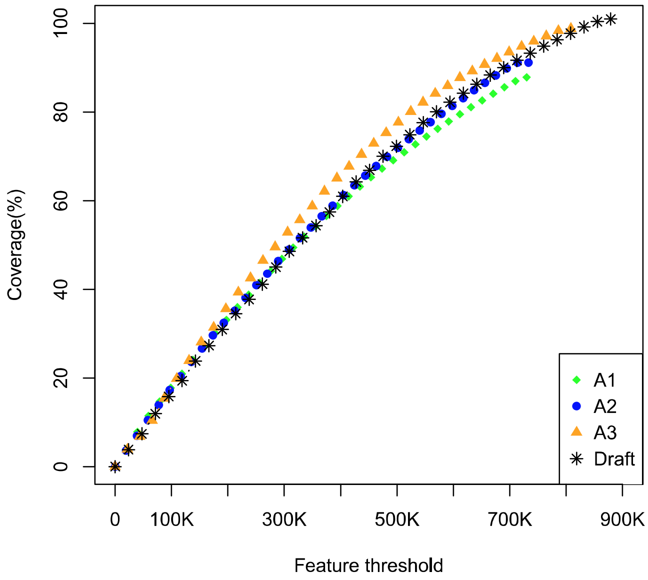


Figure 1. Feature response curves for the A1, A2, A3 and draft assembly. The FRCs were generated using FRC^{bam}[27] and plotted in R v3.4.3[51].

*GJD2like*_ XM_012838313 was found as a single complete coding sequence in A2, but parts of the gene were now triplicated (Table 5). Thus, the duplications indicated in the published draft assembly were not present in A1 or A2. They were also not present in other bony fish, such as Japanese eel (diverged before herring), Atlantic cod (diverged after herring) or zebrafish, which is probably the most heavily investigated teleost, and is supposed to have common divergence with herring from the remaining teleosts[33,34]. As this study came to an end, a chromosome level assembly of the herring genome (GCA_900700415.1) was made available along with a preprint paper[35]. The connexin duplications were also absent in this new assembly. Thus, we consider it likely that these duplications are caused by erroneous assembly.

Merging the assembly from this study with the draft assembly. Even though A2 was an improvement on A1, it was shorter and more fragmented than the published draft assembly as well as less complete (Tables 2 and 3). To generate the best possible herring assembly from the available data, A2 and the previously published assembly were merged, giving rise to A3. As can be seen in Table 2, A3 had fewer scaffolds (2,419 compared with 6,915), higher N50 (1,971,137 compared with 1,897,858), and 319 fewer Ns per 100 kbp than the draft assembly. Nevertheless, the largest scaffold was slightly shorter in A3, and there were fewer complete BUSCOs (4,258 compared to 4,348) and more fragmented BUSCOs (177 compared to 105) in A3 compared to the draft assembly (Tables 2 and 3). In addition, the total length of A3 was 17 Mb shorter than the total length of the previously published draft assembly; 3 Mb of this difference was explained by the decrease in gap length. The Metassemble[36] software package was used for merging the two assemblies. In short, the software aligns the assemblies and confirms the merging steps via mate-pair reads. In addition, unaligned sequences are removed. In the case of A3, approximately 10 Mb of sequences (3,912 short scaffolds from the draft assembly) were removed, which was the main reason for A3 being shorter than the draft assembly. Removal of these short scaffolds was another reason why the summary statistics improved. Another partial explanation was that some areas were accidentally (and probably wrongly) repeated in the draft assembly but resolved in A3. Nevertheless, 103 breakpoints and 3,224 insertions were introduced in the generation of the A3 assembly. In addition to the removal of the 3,912 short scaffolds, 202 scaffolds were joined to form 101 scaffolds. To test if the removal of scaffolds was the only reason why the summary statistics improved, the removed scaffolds were added to A3 and the summary statistics

Feature type	A1	A2	A3	Draft
COMPR_MP	1st	2nd	3rd	4th
COMPR_PE	4th	3rd	1st	2nd
HIGH_COV_PE	3rd	4th	2nd	1st
HIGH_NORM_COV	3rd	4th	2nd	1st
HIGH_OUTIE_MP	3rd	4th	1st	2nd
HIGH_SPAN_MP	4th	1st	1st	3rd
HIGH_SPAN_PE	3rd	1st	2nd	4th
LOW_COV_PE	3rd	1st	4th	2nd
LOW_NORM_COV_PE	2nd	1st	4th	3rd
STRECH_MP	3rd	4th	2nd	1st
STRECH_PE	1st	2nd	3rd	4th
Sum	30	27	25	27
Overall rank	4th	2nd	1st	2nd

Table 4. Ranking of the A1, A2, A3 and draft assemblies based on FRCs from 11 different feature types. FRC^{bam} was used for the FRC analysis. Rank: Each of the 14 features (potential assembly errors) analysed by FRC^{bam} were individually ranked (based on Supplementary Figs. S1–S14) from 1st to 4th, with 1st having the steepest FRC. The ranks were summed without weighting the features. Feature types HIGH_OUTIE_PE, HIGH_SINGLE_MP, and HIGH_SINGLE_PE were excluded because of limited data points in the FRC. Feature types are explained in the legends of Supplementary Figs. S1–S14.

for this combined assembly were calculated. This assembly had 6,331 scaffolds, an N50 of 1.96 Mb, a total length of 800 Mb, and a gap length of 80 Mb, indicating that the removed scaffolds did contribute to the improvements in the summary statistics but were not the sole reason.

The FRCs for A2 and the draft assembly were highly similar, and the main difference was the total length of the assemblies. A2 was shorter than the draft assembly, and thus the FRC only reached 91% coverage (Fig. 1). A3 showed a steeper FRC than the draft assembly but was slightly shorter. When ranking the assemblies based on overall features, the merged A3 was ranked as 1st, whereas the published draft assembly and A2 were ranked 2nd. The total number of features improved with the merging of the assemblies from 544,122 in A2 and 487,486 in the draft assembly to 473,588 in A3. These results indicate that A3 is more correct than A2 and the draft assembly.

The connexin analysis revealed duplications or triplications in six connexin genes in both A3 and the draft assembly. The same duplications/triplications were present in A3 and the draft (Table 5), suggesting that both these assemblies have some issues with repeats. Nevertheless, the missing connexins in A1 and A2 were present in both A3 and the draft.

Whole genome alignments were generated using the web tool D-Genies to investigate whether any major structural variations existed between the assemblies[37]. Figure 2 shows the alignment between A3 and the published draft assembly. The largest rearrangements are indicated by the coloured arrows, and our notion is that these indicate some of the improvements made by the merging of the assemblies.

As mentioned, a herring chromosome level assembly became available very recently[35]. A QUAST run with this assembly as a reference was conducted to compare the all available assemblies. Table 6 lists selected QUAST results. It was evident that the A2 assembly had the most misassemblies whereas A1 has the fewest, indicating that the scaffolding steps caused several misassemblies (Table 6). It was also evident that low-quality sequences were removed in A3 because A3 had the fewest number of misassembled scaffolds, fewest unaligned scaffolds, lowest duplication ratio, and longest alignment. Some of these misassemblies might be individual variations rather than actual misasseblies. However, the chromosome level assembly had 4,036 complete, 3,881 complete and single copy, 155 complete and duplicated, 174 fragmented and 374 missing BUSCOs. This indicated that the chromosome level assembly was less complete than both the previously published draft assembly and A3.

Discussion

In this study, we generated a *de novo* assembly of the herring genome and improved its fragmentation, correctness, and completeness with gap closing software and long and linked reads. The assembly was then combined with the published draft assembly[13], resulting in a less fragmented assembly that was slightly less complete but overall showed an increase in correctness, based on summary statistics, BUSCO, connexin, and FRC analyses.

Comparing two or more assemblies is not necessarily straightforward. Simple summary statistics exist, such as the number of contigs/scaffolds, N50, L50, and total assembly length. However, these metrics only evaluate the size and fragmentation; but they say very little about the quality or correctness. Studies have compared several assemblies, such as Assemblathon 1, Assemblathon 2 and GAGE[14–16], and these studies have used several metrics to get a fair comparison. A common conclusion has been that using only one metric to evaluate assemblies does not necessarily reveal the optimal assembly. Different metrics indicate different strengths and weaknesses of assemblies. We therefore chose to use several different metrics, which we believe appropriately represents the quality of the assemblies, to compare the assemblies in this study.

A comparison of A1 and A2 revealed that the long and linked reads improved the fragmentation of the assembly. The number of scaffolds decreased by 38% while the N50 almost doubled, but the gap length increased slightly. This increase in gap size was to be expected from this scaffolding step, because SSPACE-LongRead does

		A1		A2		A3		Published draft assembly	
Connexin[a]	mRNA Acc. no[b]	Scaffold[c]	Position[c]	Scaffold	Position	Scaffold	Position	Scaffold	Position
Cx32.2like	XM_012828709	—		—		38	1911178–1910384	NW_012218207	1912581–1911787
GJA5like	XM_012816449	—		5201	1–939	810	17809–16457	NW_012219501	17809–16457
GJA5like	XM_012840593	1893 1893	77725–77986^{Fr} 81019–81888^{Fr}	1668 1668	85132–86034^{Fr} 81418–81660^{Fr}	2	4108234–4107059	NW_012223947	4109526–4108351
GJB3like	XM_012818491 *XM_012818489*	1447	67762–67004	893	284447–283689	258 *258*	698572–697814 *719052–718294*	NW_012219726 *NW_012219726*	699040–698282 *719520–718762*
GJB3like	XM_012822385 *XM_012822374* *XM_012822365*	41	751478–750609	17	751444–750575	19 *19* *19*	2723482–2722843 *2725676–2724807* *2728190–2727321*	NW_012217989 *NW_012217989* *NW_012217989*	2733880–2733241 *2736074–2735205* *2738588–2737719*
GJB4like	XM_012818492 *XM_012818490*	1447	70822–70040	893	287507–286725	258 *258*	722112–721330 *701632–700850*	NW_012219726 *NW_012219726*	722580–721798 *702100–701318*
GJD2like	XM_012838313	1213 1213	110633–109796^{Fr} 93762–93246^{Fr}	1118 *1996* *1118* *1118*	117810–116668 *74407–74559* *102088–101572* *109471–108955*	35 *35* *35*	1605461–1606603 *1612109–1612625* *1617455–1617970*	NW_012223366 *NW_012223366* *NW_012223366*	1605047–1606189 *1611695–1612211* *1617041–1617556*
GJD3like	XM_012837668 *XM_012837669*	—		—		81 81 *81* *81* *81*	1079728–1080054^{e1} 1080603–1081282^{e2} *1090945–1091624* *1090176–1090488* *1091628–1091757*	NW_012223169 NW_012223169 *NW_012223169* *NW_012223169* *NW_012223169*	1079728–1080054^{e1} 1080603–1081282^{e2} *1102966–1103278*^{e1} *1090945–1091624*^{e2} *1091628–1091757*^{e2}
GJD3like	XM_012837670	4907 4907	23078–22754^{e1} 22252–21980^{e2}	216 216	557978–557512^{e1} 558804–558480^{e2}	81 81 *81*	1102952–1103276^{e1} 1103742–1104367^{e2} *1090162–1090486*	NW_012223169 NW_012223169 *NW_012223169*	1090162–1090486^{e1} 1103742–1104367^{e2} *1102952–1103276*

Table 5. Suspected assembly errors in the connexin genes of the A1, A2, and A3 assemblies and the published draft assembly. Suspected errors include regions of repetition (position written in italics) and missing connexin genes (represented as —). Fr, e1 and e2 indicate fragmented, exon 1 and exon 2, respectively. [a]The name is an abbreviation of the name given by the mentioned GenBank accession numbers. For example, 'GJB3-like' should be read as 'gap junction beta-3 protein-like, mRNA'. Please note that unique genes may have the same name. [b]GenBank nucleotide (nr) accession numbers for predicted transcripts from the published draft assembly. If the gene had several predicted transcription variants, only transcription variant 1 was included in the analyses. If several identical, or near identical (>98%) transcripts have been predicted, the other accession numbers are given in italics. [c]The positions here regarded as the coding sequence of the gene is given in normal font (the exon/intron borders are not exact), and the 'suspect repeated' regions are given in italics. The positions are given as the coding direction (*i.e.*, from the 5′) independent of whether the sequence is on the plus or minus strand.

not include the MinION read in the assembly[23]. The number and length of contigs also improved, with 18,396 fewer contigs and 26 Mb longer total contig length (Table 2) . The completeness of the assembly was also improved with scaffolding. A2 had fewer Ns per 100 kbp, an increased number of complete BUSCOs, a decreased number of fragmented BUSCOs, and increased complete predicted rRNA genes (Tables 2 and 3). Furthermore, the correctness improved. The number of total features in A2 decreased and the A2 FRC was improved. In addition, a missing connexin gene in A1 was present in A2 but new duplications in other connexin genes were introduced (Table 5). These results, as well as recent *de novo* assemblies of fish genomes[19,38] and genomes from other organisms[18,39,40], illustrate that long-read technology is highly useful in *de novo* genome assemblies.

A comparison of A2, A3 and the previously published draft assembly revealed the A3 assembly to have the best summary statistics (Table 2). Some of this improvement was because of the removed scaffolds in the merging step, but as mentioned above, even when these scaffolds were included, the summary statistics were superior to those of the draft assembly. A3 also had the fewest total features; however, the draft assembly had slightly higher level of completeness compared with A3 (4,348 complete BUSCOs compared to 4,258; Table 3). A3 was also shorter than the draft assembly. This trend of an improved versions of an assembly showing shorter assembly length was also seen in the improved cod assembly published by Tørresen *et al.*[41]. Furthermore, Holt *et al.*[42] found fewer predicted coding genes in the improved pigeon genome even though the increases in N50 and N90 were more pronounced than in the present study. The FRC for A3 was steeper than the A1, A2, and draft assembly FRCs (Fig. 1). In relation to the connexin genes, the A3 assembly had the same repeat issues as in the draft assembly (Table 5). In summary, merging A2 and the previously published assembly resulted in a mostly improved assembly, although problems probably still remain with incomplete coverage and duplications.

The A3 assembly only consists of sequences supported by alignment between A2, the draft assembly, and sequencing reads. The A3 assembly constitutes nearly 90% of the estimated herring genome[9–12]. In other words, the A3 assembly is a highly accurate and validated version of the herring genome in the sense that it highlights the regions and their accuracies found by different sequencing technologies and different assemblers. In recent years, the problem of reproducibility has been highlighted and much discussed[43]. Here, we were able to confirm the majority of the published herring genome assembly using different wet lab and *in silico* approaches, as well as generated an improved assembly that we can have strong confidence in. Nevertheless, the A3 assembly is based on four different herring individuals. Generating a genome assembly from several individuals might result in poorer assembly results because the individual variations (*e.g.*, structural rearrangements or microsatellites) may complicate the assembly process. Comparing assemblies based on different individuals is also challenging,

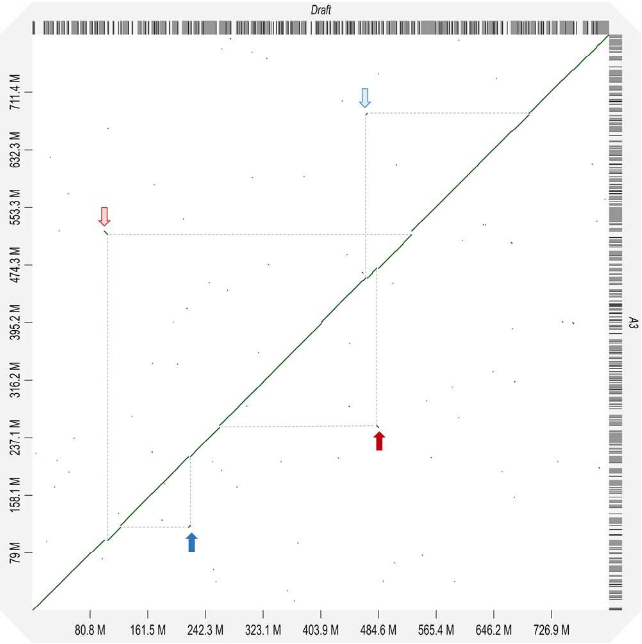


Figure 2. Dotplot showing the whole genome alignment between the published draft herring genome assembly and the A3 herring genome assembly from this study. The alignment was generated using D-Genies[37]. Examples of transpositions between the two assemblies are indicated by blue arrows and examples of inverted transpositions are shown by red arrows. The horizontal and vertical grey dotted lines indicate the positions on the two assembles that are affected.

because it might not be possible to tell if assembly differences are due to individual variation or assembly error. Using a single individual to generate an assembly is therefore preferable, but due to degraded DNA this was not possible in this study. This means that some of the corrections and differences found in this study could be individual differences between the herring used in this study and the one used by Martinez Barrio *et al.*[13]. Thus, the A3 assembly approaches an average herring genome rather than a genome from a specific herring.

As mentioned earlier, a high-quality chromosome level assembly of the herring genome was made available just as this study was coming to an end. We found that all the available assemblies had misassembly issues compared with the new chromosome level assembly. A1 had the fewest misassemblies, whereas A3 had the fewest misassembled scaffolds relative to the chromosome-level assembly. From this comparison, it was evident that scaffolding using linked and long reads can cause misassemblies. However, using more stringent scaffolding parameters and more data would reduce the number of misassemblies introduced. As mentioned above, some of these misassemblies could also be variations between the individuals used for the various assemblies and not true misassemblies. A3 and the published draft assembly were highly similar in this comparison. A3 had fewer misassembled scaffolds, fewer local misassemblies, fewer unaligned scaffolds (both full and partial alignments), shorter unaligned length, slightly lower duplication rate, the largest alignment, and higher NA50 and NGA50. By contrast, the draft assembly had fewer misassemblies, shorter misassembled scaffold length, a slightly higher fraction of the genome assembled, and a longer total aligned length (Table 6). It is also worth mentioning that the BUSCO analysis revealed both the A3 and draft assemblies to be more complete than the chromosome level assembly, at least in relation to the number of genes.

To conclude, the A3 assembly was the most complete and correct herring genome assembly with the best summary statistics. This assembly is an improvement on the previously published herring draft genome assembly in terms of correctness, and acts as a validation of the herring genome assembly. The results from this study underline how important long and linked read data are in *de novo* genome assembly. Both the long and linked reads improved the herring genome assembly in this study. Combining the assemblies from this study with the draft herring assembly resulted in an improved herring genome assembly. Additionally, this study showed, in agreement with previous studies[14–16], the importance of comparing both the correctness and completeness of genome assemblies.

Materials and Methods

Sample collection and DNA extraction.

A single Atlantic herring kidney sample was sequenced on a NextSeq500 sequencer (Illumina, San Diego, California, United States) and a MinION nanopore sequencer (Oxford Nanopore Technologies, Oxford, England). The herring was collected on a research cruise by the Faroe Marine Research Institute in the summer of 2015. The kidney sample was stored in RNAlater (ThermoFisher Scientific, Waltham, Massachusetts, United States). After 24 hours at room temperature the sample was frozen until used. DNA was extracted using an AS1000 Maxwell 16 instrument (Promega, Madison, Wisconsin, United States) and the Maxwell 16 Tissue DNA purification kit (Promega). DNA concentration was measured using a Qubit 3.0 fluorometer (ThermoFisher Scientific).

Metric	A1	A2	A3	Draft
# misassemblies	4,284	8,810	6,045	6,034
# misassembled scaffolds	2,306	2,499	572	649
Misassembled scaffolds length (bp)	326,883,342	549,092,805	621,722,316	616,769,397
# local misassemblies	19,581	30,042	55,799	55,990
#scaffold gap extensive misassemblies	369	806	436	439
# scaffold gap local misassemblies	82,670	70,490	23,309	22,741
# possible misassemblies by TEs	2,922	4,056	3,640	3,548
# unaligned misassembled scaffolds	1,292	950	892	1,157
# unaligned scaffolds (full + partial)	463 + 8,064	247 + 5,563	61 + 1,706	228 + 2,256
Unaligned length (bp)	84,214,870	97,034,653	211,181,030	217,841,770
Genome fraction (%)	59.41	61.11	66.87	66.94
Duplication ratio	1.42	1.42	1.19	1.20
# mismatches per 100 kbp	709.66	883.77	1,634.08	1,643.42
# indels per 100 kbp	110.22	109.28	127.10	127.13
Largest alignment (bp)	1,320,028	1,496,625	1,700,060	1,587,972
Total aligned length (bp)	435,132,440	456,232,649	501,922,313	503,353,702
NA50 (bp)	30,242	35,372	77,322	69,158
NGA50 (bp)	23,530	35,287	114,419	112,940
LA50 (bp)	3,156	2,771	1,498	1,600
LGA50 (bp)	3,717	2,775	1,159	1,174
K-mer-based compl. (%)	42.13	43.00	51.92	51.95
K-mer-based correct length (%)	72.39	39.82	54.43	57.40
K-mer-based misassembled length (%)	19.02	54.57	43.06	38.94
# k-mer-based misjoins	800	1,967	433	423

Table 6. QUAST generated comparisons of A1, A2, and A3 assemblies and the published draft herring assembly, using the new chromosome level assembly as reference. Thus, all results are relative to the chromosome level assembly.

The sample for another three MinION runs was caught in Haraldssund, Faroe Islands, by the local fishing boat 'Sildin'. In an attempt to obtain DNA molecules as long as possible, the DNA was extracted as soon as the boat came ashore. It was extracted from the kidney using an AS1000 Maxwell 16 instrument and the Maxwell 16 Tissue DNA purification kit. The smaller DNA fragments were excluded by a 0.8x volume of AMPureXP bead (Beckman Coulter, Brea, California, United States) clean-up, as per the manufacturer's instructions. DNA concentration was measured using the Qubit 3.0 fluorometer and the purity was measured using a NanoPhotometer™ Pearl instrument (IMPLEN, Munich, Germany).

The sample used for 10x Genomics sequencing was caught by the local fishing boat 'Grani' on Kaldbaksfjørður, Faroe Islands. The DNA from the kidney was extracted using the MagAttract HMW DNA Kit (Qiagen, Hilden, Germany).

Ethics. The herring samples were received from stock assessment cruises and commercial catches. No fish were caught for the purpose of this project, and all fish were dead when they were selected. Thus, no ethical approval was required.

Library preparation for Illumina sequencing. For the paired-end sequencing, the DNA was fragmented to roughly 300 bp using a Covaris M220 focused-ultrasonicator (Covaris, Woburn, Massachusetts, United States). The library was then prepared using the KAPA LTP Library Preparation Kit (KAPABiosystems, Wilmington, Massachusetts, United States) and quantified using the KAPA Library Quantification Kit (KAPABiosystems), following the manufacturer's instructions. The paired-end library was sequenced on a NextSeq500 (Illumina) using one Mid and one High Output v2 kit.

Two mate-pair libraries, with intended insert sizes of 4,500 bp and 7,000 bp, were prepared using the Nextera Mate-Pair Library Preparation kit (Illumina), following the manufacturer's instructions. The mate-pair libraries were quantified using the KAPA Library Quantification Kit (KAPABiosystems) and sequenced on a NextSeq500 (Illumina). However, when later investigated bioinformatically, both libraries seemed to have an insert size of approximately 2 kbp. This was most likely because of error in the library preparation and/or fragmented DNA. One of the libraries was sequenced with a High Output v2 kit while the other was sequenced with a Mid Output v2 kit.

Oxford nanopore technologies. Four different MinION runs were conducted. The library for the first run was prepared using the same DNA sample as the Illumina sequencing together with the Rapid Sequencing kit (SQK-RAD001). The library was sequenced on a FLO-MIN105 flow cell and run for 48 hours. After the run, the reads were uploaded to Metrichor v1.2.6 for base calling. To obtain longer reads, a fresh DNA sample from a different individual was used for the subsequent MinION runs. Run two was conducted by using the Rapid

Sequencing kit (SQK-RAD002) and a FLO-MIN107 flow cell. The MinION ran for 28 hours and reads were uploaded to Metrichor v1.5.7 for base calling. Runs three and four were conducted using the Ligation Sequencing kit (SQK-LSK108) and FLO-MIN107 flow cells. The MinION ran for 48 hours and the reads were base-called using Albacore v1.2.5 (Oxford Nanopore Technologies). All protocols followed the manufacturers' instructions, except for the SQK-LSK108 kit where the DNA repair step was omitted.

10x Genomics. The linked reads were generated from a 10x Genomics library prepared by the Chromium Genome Reagent Kit (10x Genomics, San Francisco, California, United States) according to the manufacturer's instructions and altered according to the technical note 'Guidelines for De Novo Assembly of Genomes Smaller than ~3 Gb using 10x Genomics® Supernova TM V1.2'[44] and personal communication with 10x Genomics staff. The library was sequenced on a NextSeq. 500 (Illumina) using a High Output v2 kit.

Data pre-processing. All the data processing, assemblies and comparisons were performed on the EMBL-EBI cluster in Hinxton, except for the manual connexin gene analysis.

Trimmomatic v0.36 was used to remove adapter sequences and trim low-quality bases with an average quality score lower than 20 (sliding window of four bases) from the paired-end data[45]. Then, AfterQC v0.4.0 was used to remove the polyG reads[46]. The mate-pair data were also subjected to the same trimming conditions as the paired-end data using Trimmomatic, but adapters were not trimmed. In addition, the data were also processed using NextClip v1.3.1, and only the reads with one or both adapter sequences were used[47].

FastQC v0.11.5 was used to assess the quality of all the sequencing data[48]. Poretools v0.6.0[49] was used to extract the FASTQ files longer than 500 bp from MinION runs one and two, whereas Albacore v1.2.5 was used for runs three and four.

The assembly process. The first assembly (A1) was generated using the Illumina data and the de Bruijn graph assembler AllPaths-LG v52488[20]. This assembler was chosen because of the size of the genome and the results from the Assemblathon 2 study[16], where it performed well on the fish genome assembly. The Illumina data were generated with this assembler in mind. Several parameters and subsets of the data were tested, and the best assembly was chosen for further use in this study. In addition, the SGA v0.10.15[21] and MaSuRCA v3.2.2[22] assemblers were tested, but did not yield as good assemblies as AllPaths-LG assembler. Supplementary Table S1 contains the different parameters and subsets of the data used for the different assembly runs.

A2 was generated by closing gaps in A1, in addition to two scaffolding steps. The GapFiller v1.10 software package was used to close gaps. In short, this software aligns sequencing reads to the assembly and then tries to extend the ends of the contigs, if enough sequencing reads support this[50]. We ran this software for 20 iterations. The resulting assembly was then scaffolded with four runs of MinION reads using the SSPACE-LongRead v1.1 software package[23]. In addition to the default parameters, the options −a 500 and −l 1 were used, indicating the length of alignment and number of links required for scaffolding. The linked reads were intended for a *de novo* assembly using the Supernova v1.2.2 assembler (10x Genomics) but because of a problematic sequencing run the data did not yield a good assembly (results not shown). Therefore, a second scaffolding step was performed using the linked reads and ARCS v1.0.5[24] (default parameters). Simply stated, ARCS and SSPACE-LongRead scaffold sequences by aligning the new data (linked and long reads, respectively) to the sequences (A1 in our case) and if these new data align to different sequences these are merged[23,24].

A3 was generated by combining A2 and the draft assembly using Metassembler v1.5[36]. The previously published draft assembly was used as the primary assembly, together with the mate-pair data from this study. A run with A2 as the primary assembly was also conducted but resulted in a poorer assembly. The merged assembly was again scaffolded using the linked reads and ARCS, as described above.

Comparisons using QUAST and BUSCO. To compare the assemblies in this study and the draft assembly, we used the genome comparison tool QUAST v5[25] with the option – large and no reference assembly. QUAST was also run with the newly available chromosome level herring assembly as a reference. QUAST can also run a BUSCO analysis using the eukaryotic database. However, we chose to run a separate standalone BUSCO analysis using the Actinopterygii database[50], to compare the completeness of the generated assemblies.

Manual connexin analysis. A manual analysis of the connexin gene family[30] was performed to assess the correctness and completeness of the assemblies. We collected all predicted connexin genes/mRNAs available in GenBank from the herring genome published by Martinez Barrio *et al.*[13]. This amounted to 49 connexin genes (before exclusion of near identical sequences). We also searched for additional (non-predicted) connexin genes in the published draft assembly using the NCBI Basic Local Alignment Search Tool (BLAST). Any hit was manually inspected, and two additional connexin sequences were found: one connexin gene predicted as *KAT6B-like* (a *gja8-like* sequence) and one previously non-predicted sequence (a *cx39.2/gjd2-like* sequence). After exclusion of five predicted sequences that showed >98.4% identity to other connexin sequences we had a set of 46 unique connexin sequences (Supplementary Table S2). We blasted the unique sequences against our unannotated assemblies and any unexpected hits were noted. Correspondingly, any unexpected hits in the published draft herring genome were noted.

FRC. FRC^{bam} v1.3.0 and the paired-end and mate-pair data from the present study were used to evaluate the correctness of the assemblies[27]. The FRC^{bam} output consists of FRCs for 14 feature types. To rank the assemblies based on the different types of features, all 14 FRCs were plotted, and for each the best assembly was given 1 point, second best 2 points, and so on. If two assemblies had very similar curves, both assemblies received the same number of points. For example, A1 had the steepest curve and received 1 point, and both A2 and A3 had the second steepest curve so both received 2 points. Then, no assembly received 3 points, but the next assembly received 4 points. If the curve only had two points, the feature was excluded. The scores were summed and the assembly with the lowest score was ranked first.

Lastly, the assemblies were aligned against each other using D-Genies[37] to determine whether any major structural variations existed.

Data availability

The sequencing reads and assemblies are available in the European Nucleotide Archive repository, under the project accession http://www.ebi.ac.uk/ena/data/view/ERP107609.

Received: 27 August 2019; Accepted: 8 November 2019;
Published online: 27 November 2019

References

1. Food and Agriculture Organization of the United Nations. *Fishery and aquaculture statistics yearbook 2016*. (Food and Agriculture Organization of the United Nations. Fishery and Aquaculture Statistics Yearbook 2016, 2018).
2. Hagstova Føroya. *Heildarfiskiveiðan skift á leiðir og fiskaslag* (*1990–2017*). (Hagstova Føroya, 2017).
3. Hay, D. *et al*. In *Herring: Expectations for a new millennium* (eds Funk, F. *et al.*) 381–454 (University of Alaska Sea Grant, Fairbanks, 2001).
4. Pampoulie, C. *et al*. Stock structure of Atlantic herring *Clupea harengus* in the Norwegian Sea and adjacent waters. *Marine Ecology Progress Series* **522**, 219–230, https://doi.org/10.3354/meps11114 (2015).
5. Smith, P., Francis, R. & McVeagh, M. Loss of genetic diversity due to fishing pressure. *Fisheries Research* **10**, 309–316, https://doi.org/10.1016/0165-7836(91)90082-Q (1991).
6. Nielsen, E. E. *et al*. Gene-associated markers provide tools for tackling illegal fishing and false eco-certification. *Nature Communications* **3**, 851, https://doi.org/10.1038/ncomms1845 (2012).
7. Teacher, A., Kähkönen, K. & Merilä, J. Development of 61 new transcriptome-derived microsatellites for the Atlantic herring (*Clupea harengus*). *Conservation Genetics Resources* **4**, 71–74, https://doi.org/10.1007/s12686-011-9477-5 (2012).
8. Bekkevold, D. *et al*. Gene-associated markers can assign origin in a weakly structured fish, Atlantic herring. *ICES Journal of Marine Science* **72**, 1790–1801, https://doi.org/10.1093/icesjms/fsu247 (2015).
9. Ida, H., Oka, N. & Hayashigaki, K.-I. Karyotypes and cellular DNA contents of three species of the subfamily Clupeinae. *Japanese Journal of Ichthyology* **38**, 289–294, https://doi.org/10.11369/jji1950.38.289 (1991).
10. Hardie, D. C. & Hebert, P. D. Genome-size evolution in fishes. *Canadian Journal of Fisheries and Aquatic Sciences* **61**, 1636–1646, https://doi.org/10.1139/f04-106 (2004).
11. Ohno, S., Muramoto, J., Klein, J. & Atkin, N. Diploid-tetraploid relationship in clupeoid and salmonoid fish. *Chromosomes today* **2**, 139–147 (1969).
12. Hinegardner, R. & Rosen, D. E. Cellular DNA content and the evolution of teleostean fishes. *The American Naturalist* **106**, 621–644, https://doi.org/10.1086/282801 (1972).
13. Martinez Barrio, A. *et al*. The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. *eLife* **5**, e.12081, https://doi.org/10.7554/eLife.12081 (2016).
14. Salzberg, S. L. *et al*. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research* **22**, 557–567, https://doi.org/10.1101/gr.131383.111 (2012).
15. Earl, D. *et al*. Assemblathon 1: a competitive assessment of *de novo* short read assembly methods. *Genome Research* **21**, 2224–2241, https://doi.org/10.1101/gr.126599.111 (2011).
16. Bradnam, K. R. *et al*. Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. *GigaScience* **2**, 10, https://doi.org/10.1186/2047-217X-2-10 (2013).
17. Mostovoy, Y. *et al*. A hybrid approach for *de novo* human genome sequence assembly and phasing. *Nature Methods* **13**, 587, https://doi.org/10.1038/nmeth.3865 (2016).
18. Ye, C., Hill, C. M., Wu, S., Ruan, J. & Ma, Z. DBG2OLC: Efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Scientific Reports* **6**, 31900, https://doi.org/10.1038/srep31900 (2016).
19. Tan, M. H. *et al*. Finding Nemo: hybrid assembly with Oxford Nanopore and Illumina reads greatly improves the clownfish (*Amphiprion ocellaris*) genome assembly. *GigaScience* **7**, gix137, https://doi.org/10.1093/gigascience/gix137 (2018).
20. Butler, J. *et al*. ALLPATHS: *de novo* assembly of whole-genome shotgun microreads. *Genome Research* **18**, 810–820, https://doi.org/10.1101/gr.7337908 (2008).
21. Simpson, J. T. & Durbin, R. Efficient construction of an assembly string graph using the FM-index. *Bioinformatics* **26**, i367–i373 (2010).
22. Zimin, A. V. *et al*. The MaSuRCA genome assembler. *Bioinformatics* **29**, 2669–2677, https://doi.org/10.1093/bioinformatics/btt476 (2013).
23. Boetzer, M. & Pirovano, W. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics* **15**, 211, https://doi.org/10.1186/1471-2105-15-211 (2014).
24. Yeo, S., Coombe, L., Warren, R. L., Chu, J. & Birol, I. ARCS: scaffolding genome drafts with linked reads. *Bioinformatics* **34**, 725–731, https://doi.org/10.1093/bioinformatics/btx675 (2017).
25. Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D. & Gurevich, A. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* **34**, i142–i150, https://doi.org/10.1093/bioinformatics/bty266 (2018).
26. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
27. Vezzi, F., Narzisi, G. & Mishra, B. Reevaluating assembly evaluations with feature response curves: GAGE and assemblathons. *Plos One* **7**, e52210, https://doi.org/10.1371/journal.pone.0052210 (2012).
28. Narzisi, G. & Mishra, B. Comparing *de novo* genome assembly: the long and short of it. *Plos One* **6**, e19175, https://doi.org/10.1371/journal.pone.0019175 (2011).
29. Phillippy, A. M., Schatz, M. C. & Pop, M. Genome assembly forensics: finding the elusive mis-assembly. *Genome Biology* **9**, R55, https://doi.org/10.1186/gb-2008-9-3-r55 (2008).
30. Cruciani, V. & Mikalsen, S.-O. Evolutionary selection pressure and family relationships among connexin genes. *Biological Chemistry* **388**, 253–264, https://doi.org/10.1515/BC.2007.028 (2007).
31. Eastman, S. D., Chen, T. H.-P., Falk, M. M., Mendelson, T. C. & Iovine, M. K. Phylogenetic analysis of three complete gap junction gene families reveals lineage-specific duplications and highly supported gene classes. *Genomics* **87**, 265–274, https://doi.org/10.1016/j.ygeno.2005.10.005 (2006).
32. Cruciani, V. & Mikalsen, S.-O. The vertebrate connexin family. *Cellular and Molecular Life Sciences* **63**, 1125–1140, https://doi.org/10.1007/s00018-005-5571-8 (2006).

33. Near, T. J. *et al.* Resolution of ray-finned fish phylogeny and timing of diversification. *Proceedings of the National Academy of Sciences* **109**, 13698–13703, https://doi.org/10.1073/pnas.1206625109 (2012).
34. Betancur-R, R. *et al.* The tree of life and a new classification of bony fishes. *PLoS currents* **5**, https://doi.org/10.1371/currents.tol.53ba26640df0ccaee75bb165c8c26288 (2013).
35. Pettersson, M. E. *et al.* A chromosome-level assembly of the Atlantic herring – detection of a supergene and other signals of selection. *bioRxiv*, 668384, https://doi.org/10.1101/668384 (2019).
36. Wences, A. H. & Schatz, M. C. Metassembler: merging and optimizing *de novo* genome assemblies. *Genome Biology* **16**, 207, https://doi.org/10.1186/s13059-015-0764-4 (2015).
37. Cabanettes, F. & Klopp, C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ* **6**, e4958, https://doi.org/10.7717/peerj.4958 (2018).
38. Austin, C. M. *et al. De novo* genome assembly and annotation of Australia's largest freshwater fish, the Murray cod (*Maccullochella peelii*), from Illumina and Nanopore sequencing read. *GigaScience* **6**, 1–6, https://doi.org/10.1093/gigascience/gix063 (2017).
39. Jansen, H. J. *et al.* Rapid *de novo* assembly of the European eel genome from nanopore sequencing reads. *Scientific Reports* **7**, 7213, https://doi.org/10.1038/s41598-017-07650-6 (2017).
40. Michael, T. P. *et al.* High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. *Nature Communications* **9**, 541, https://doi.org/10.1038/s41467-018-03016-2 (2018).
41. Tørresen, O. K. *et al.* An improved genome assembly uncovers prolific tandem repeats in Atlantic cod. *BMC Genomics* **18**, 95, https://doi.org/10.1186/s12864-016-3448-x (2017).
42. Holt, C. *et al.* Improved genome assembly and annotation for the rock pigeon (*Columba livia*). *G3: Genes, Genomes, Genetics* **8**, 1391–1398, https://doi.org/10.1534/g3.117.300443 (2018).
43. Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature News* **533**, 452, https://doi.org/10.1038/533452a (2016).
44. 10x Genomics. CG000100 Rev A Guidelines for de novo assembly of genomes smaller than ~3 Gb using 10x Genomics® Supernova TM V1.2. (10x Genomics, 2017).
45. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120, https://doi.org/10.1093/bioinformatics/btu170 (2014).
46. Chen, S. *et al.* AfterQC: automatic filtering, trimming, error removing and quality control for fastq data. *BMC Bioinformatics* **18**, 80, https://doi.org/10.1186/s12859-017-1469-3 (2017).
47. Leggett, R. M., Clavijo, B. J., Clissold, L., Clark, M. D. & Caccamo, M. NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries. *Bioinformatics* **30**, 566–568, https://doi.org/10.1093/bioinformatics/btt702 (2013).
48. Andrews, S. FastQC: a quality control tool for high throughput sequence data (Available online at, http://www.bioinformatics.babraham.ac.uk/projects/fastqc, 2010).
49. Loman, N. J. & Quinlan, A. R. Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics* **30**, 3399–3401, https://doi.org/10.1093/bioinformatics/btu555 (2014).
50. Waterhouse, R. M. *et al.* BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular biology and evolution* **35**, 543–548, https://doi.org/10.1093/molbev/msx319 (2017).
51. R Core Team. *R: A language and environment for statistical computing* (2015).

Acknowledgements

We are grateful for the staff onboard the local fishing boats 'Sildin' and 'Grani' for supplying samples and Jens Arni Thomassen and Poul Vestergaard from the Faroe Marine Research Institute for helping collect fish data and tissues. Further, we are grateful to Hannes Gislason, Hóraldur Joensen, Debes H. Christiansen, and Janus Vang for their help during the acquisition of funding. Special thanks go to Line H.G. Larsen from Amplexa Genetics for the help regarding laboratory methods and the members of the Flicek Research group for their bioinformatical help. In addition, we thank the following institutions for funding: the Faroese Research Council, the Fisheries Research Fund of the Faroe Islands, Innovationsfonden (Denmark), Statoil Føroyar, the Faroese Pelagic Organisation, and the European Molecular Biology Laboratory.

Author contributions

S.í.K. contributed to the design of the study, conducted the laboratory work, performed the analysis and interpretation of the work as well as writing the manuscript. S.O.M. contributed to the design of the study and writing of the manuscript, as well as supervised the laboratory work and analysis and interpretation of data. E.í.H. and J.A.J. contributed to the acquisition and interpretation of the data. P.F. contributed to the design of the study, writing of the manuscript, and analysis and interpretation of the data. H.A.D. designed the study, acquired funding, contributed to the writing of the manuscript, and supervised the laboratory work and analysis and interpretation of data. All authors contributed to revising the manuscript and approved the final version.

Competing interests

H.A.D. is an employee and co-founder of Amplexa Genetics, a private clinical laboratory with a commercial interest in molecular genetics. S.í.K. was, at the time of the study employed at Amplexa Genetics. S.í.K. and H.A.D. have received funding from the Faroese Pelagic Organisation who have a commercial interest in the investigated species. However, these interests did not influence the design of the study; the collection, analysis, and interpretation of data; or the writing of the manuscript. S.O.M., E.í.H., J.A.J. and P.F. declare that they have no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-54151-9.

Correspondence and requests for materials should be addressed to S.í.K.

Reprints and permissions information is available at www.nature.com/reprints.

3.2. Phylogeny of teleost connexins reveals highly inconsistent intra- and interspecies use of nomenclature and misassemblies in recent teleost chromosome assemblies

Svein-Ole Mikalsen[1, 4], Marni Tausen[1, 2], Sunnvør í Kongsstovu[1,3]

[1]Faculty of Science and Technology, University of the Faroe Islands, Vestara Bryggja 15, FO-100 Tórshavn, The Faroe Islands.

[2]Present affiliation: Bioinformatics Research Centre, Aarhus University, C. F. Møllers Allé 8, 8000 Aarhus C, Denmark.

[3]Amplexa Genetics A/S, Hoyvíksvegur 51, FO-100 Tórshavn, The Faroe Islands.

[4]Corresponding author: sveinom@setur.fo

ORCiD:

Svein-Ole Mikalsen, 0000-0002-7128-4464

Marni Tausen, 0000-0003-0694-9199

Sunnvør í Kongsstovu, 0000-0001-6631-2347

3.2.1. Abstract

Background: Based on an initial collecting of database sequences from the gap junction protein gene family (also called connexin genes) in a few teleosts, the naming of these sequences appeared variable. The reasons could be (i) that the structure in this family is variable across teleosts, or (ii) unfortunate naming. Rather clear rules for the naming of genes in fish and mammals have been outlined by nomenclature committees, including the naming of orthologous and ohnologous genes. We therefore analyzed the connexin gene family in teleosts in more detail. We covered the range of divergence times in teleosts (eel, Atlantic herring, zebrafish, Atlantic cod, three-spined stickleback, Japanese pufferfish and spotted pufferfish; listed from early divergence to late divergence).

Results: The gene family pattern of connexin genes is similar across the analyzed teleosts. However, (i) several nomenclature systems are used, (ii) specific orthologous groups contain genes that are named differently in different species, (iii) several distinct genes have the same name in a species, and (iv) some genes have incorrect names. The latter includes a human connexin pseudogene, claimed as *GJA4P*, but which in reality is *Cx39.2P* (a delta subfamily gene often called *GJD2like*). We point out the ohnologous pairs of genes in teleosts, and we suggest a more consistent nomenclature following the outlined rules from the nomenclature committees. We further show that connexin sequences can indicate some errors in two high-quality chromosome assemblies that became available very recently.

Conclusions: Minimal consistency exists in the present practice of naming teleost connexin genes. A consistent and unified nomenclature would be an advantage for future automatic annotations and would make various types of subsequent genetic analyses easier. Additionally, roughly 5% of the connexin sequences point out misassemblies in the new high-quality chromosome assemblies from herring and cod.

Keywords

connexins; genome duplication; mammals; nomenclature; ohnologs; orthologs; paralogs; phylogenetic trees; teleosts.

3.2.2. Background

Large-scale sequencing techniques developed since the turn of the century have caused a virtual explosion of species with sequenced genomes. A critical part of making all these genomes useful is the process of annotation, of which gene identification and gene naming are indispensable parts [1-3]. Computerized annotation by algorithms and the use of previously identified sequences available in databanks are needed to keep up with the flow of new genomes. However, computerized annotations are only as good as the assumptions behind the algorithms and the available data, including identifications, allow.

The Human Gene Nomenclature Committee states as the first point in its summary guidelines that "each approved gene symbol must be unique" [4]. Some general principles of naming genes in zebrafish (and by extension in other teleosts) are outlined by the Zebrafish Information Network [5]. The Zebrafish Nomenclature Conventions states that "genes should be named after the mammalian ortholog whenever possible" [5]. We here understand orthologs in the same meaning as originally defined by Fitch [6, 7], who divided homologs into two main classes: orthologs and paralogs. In simple terms, orthologs are the same genes in different species. All the other genes in a gene family are paralogs, whether intraspecies or interspecies. Note that in this context, the functional relationship or expression pattern is irrelevant (in contrast to some deviant definitions of orthologs, for example on p. 726 in ref. [8]). Thus, a pseudogene in one species can be an ortholog of a functional gene in another species, even if the pseudogene has no known function or is not expressed.

Giving unique names to unique genes [4] and naming teleost genes according to the mammalian ortholog [5] appear as sound principles. The Zebrafish Information Network details that in the case of duplicated genes resulting from genome duplication, "symbols for the two zebrafish genes should be the same as the approved symbol of the human or mouse ortholog followed by "a" or "b" to indicate that they are duplicated copies" [5]. In the case of tandem gene duplication, the duplicates "with a single mammalian ortholog should have gene symbols appended with a .1, .2, using the same symbol as the mammalian ortholog" [5]. This may not always be easy to establish unequivocally, as it requires much work and there may be a long time between the initial genome assembly and the complete genome being assembled into chromosomes. A good indication of orthology may come from phylogenetic analyses.

Of course, reality is often not simple, as both genome duplications, tandem gene duplications, gene losses, the formation of pseudogenes, retrotranscription and reinsertion, and other genetic events may have occurred since the evolutionary separation of the different species in question. Two genome duplications occurred during the early evolution of vertebrates after the divergence of the urochordates [9-11]. These genome duplications are common to both teleosts and tetrapods. Additionally, another genome duplication occurred in the early evolution of teleosts [12-14].

The pairs of genes created by genome duplication are called ohnologs [15, 16]. As such, ohnologs are a specific subgroup of paralogs [6, 7]. Being on different chromosomes, different genetic events may happen for each member of an ohnologous pair, such as mutations of various kinds, gene losses, tandem gene duplication at one of the sites, etc. It is therefore not necessarily a 1:1 relationship between ohnologs in teleosts (e.g., one of the ohnologs could be lost in one or several species), or between mammalian and teleost orthologs [6, 7]. Furthermore, the synteny (the linear order of genetic elements in DNA) can be muddled. Adding to this evolutionary genetic complexity, there are also technical and bioinformatic caveats, making complete and perfect genome assemblies unlikely. Presently, the published genome assemblies are often estimated to be around 90 % complete [17, 18], being in thousands of scaffolds instead of a few tens of chromosomes. Moreover, numerous kinds of assembly errors [19, 20] can further complicate the annotation process.

It was early observed that certain gene families had unusually large number of members in fish model species [21]. One of these gene families is the gap junction protein gene family, encoding the proteins called connexins (for simplicity, we will generally refer to the genes as connexin genes). This family has approximately twice as many members in teleost species as in other vertebrates [22-24], and as such has retained more than its fair share of genes generated by genome duplication compared with many other gene families, which generally retain 1 to 20% of the duplicated genes (see review by Glasauer and Neuhauss [25]).

Both a size-based (in kiloDalton) nomenclature and a Greek nomenclature have been used in naming the genes in this family (e.g., *connexin43*, abbreviated *cx43*, in the size nomenclature is the same as *gja1* in the Greek nomenclature). A disadvantage with a size-based nomenclature is that the protein size may vary in different species, and thus the relationship with the corresponding genes/proteins in other species may not be immediately clear. The Greek nomenclature divides the group into several subfamilies from alpha to epsilon and with a number that initially stated the chronology of detection. The Human and Mouse Gene Nomenclature Committees have decided to use the Greek gene nomenclature for the connexin genes.

The connexin genes are chordate-specific genes, urochordates being the most primitive organisms having these genes [24, 26], which in the vertebrates have evolved into distinct subfamilies [22, 24, 27, 28]. The connexin proteins are transmembrane molecules that aggregate into hexamers forming a pore through the membrane, often called a hemichannel. Traditionally, it was supposed that hemichannels would not act alone, but rather line up with a corresponding hemichannel from the neighboring cell to form a channel directly from the cytosol in one cell to the cytosol in the other cell, through which small water-soluble molecules and ions can diffuse [29]. In some tissues, such as the heart and uterus, these channels are of utmost importance for passing the electrical impulse from cell to cell, making these organs contract in a synchronized manner [30, 31]. The channels are probably also involved in cellular homeostasis and growth control [32], possibly through interactions with numerous proteins involved in signaling and regulation [33-35]. Additionally, there are now strong indications that hemichannels are functional in their own right [36-38].

The teleosts are the most species-rich group among vertebrates. In connection with the sequencing and assembly of the Atlantic herring genome (S. í Kongsstovu *et al.*, submitted), we collected some teleost connexin sequences, and soon noticed that the naming appeared variable. The two most obvious explanations for the variability were (i) that the structure in this family is variable across the teleosts, or (ii) unfortunate naming. We therefore examined the connexins in teleost species more closely. We updated sequences analyzed in previous work [22-24, 28], and added several other species (Atlantic cod [*Gadus morhua*], Atlantic herring [*Clupea harengus*], and Japanese eel [*Anguilla japonica*]; the latter supported by European and American eel [*Anguilla anguilla* and *Anguilla rostrata*]) [17, 39-42]. This selection of teleosts spans the range of divergence times in this vertebrate group. A genome duplication occurred at the basis of the teleosts ~350 million years ago, and the Elopomorpha (to which eels belong) was the first group to diverge ~300 million years ago [43, 44]. The Clupeiformes (to which herring belongs) and Cypriniformes (to which zebrafish belongs) had a common divergence ~250 million years ago, and soon after (~240 million years ago) split into separate groups. The Acantomorphata diverged ~150 million years ago, and later split into several subgroups, of which the Gadiformes (to which cod belongs) is one [43, 44]. The Perciformes (to which sticklebacks belong) diverged ~100 million years ago [43]. The Tetraodontiformes (pufferfishes) are among the most recently diverged groups, ~70 million years ago [43].

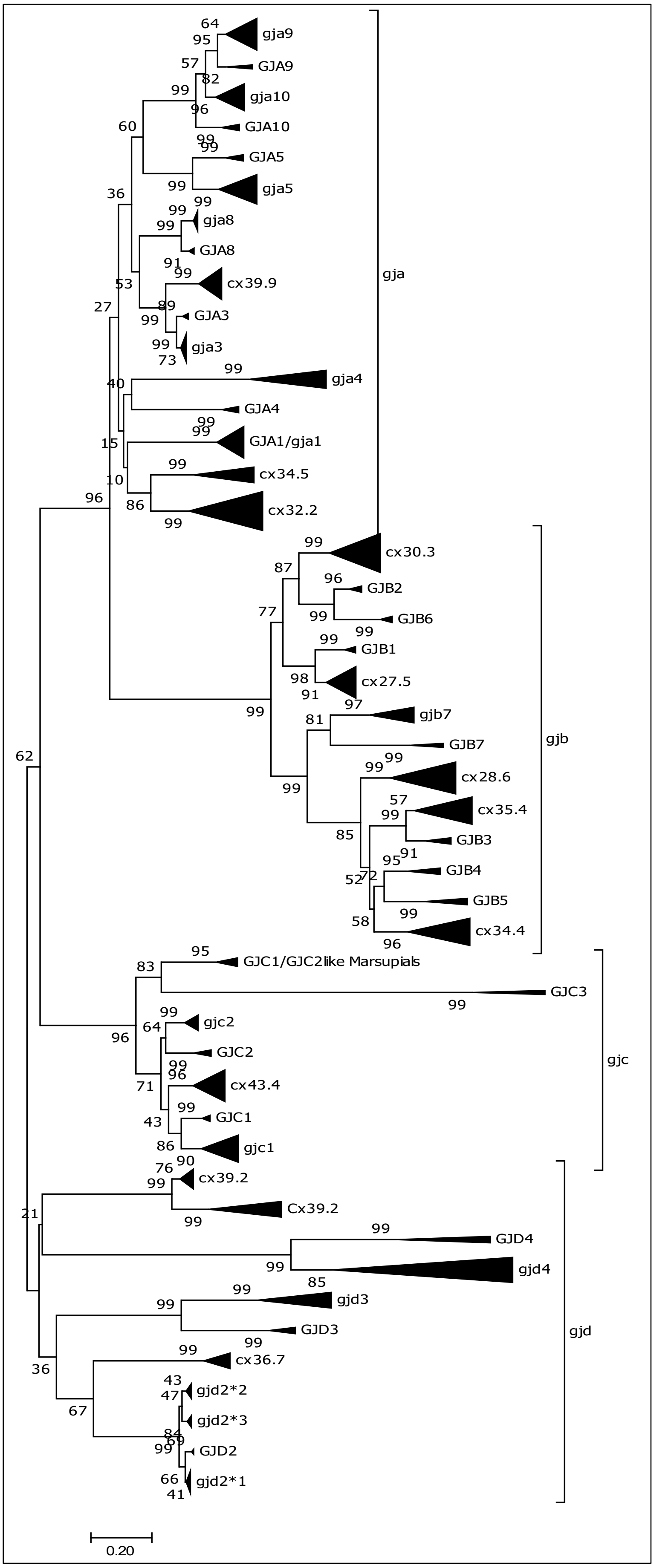
As the genes should be named after the mammalian ortholog whenever possible [5], the connexin sequences from several mammals were included. The sequences were analyzed phylogenetically, using the names indicated in the databases whenever possible. Our results show that a considerable degree of inconsistency exists in the naming of the connexin genes in fish species. There is even a case of inconsistent naming among the human sequences. In our opinion, making the naming in this gene family more congruent and consistent is indeed possible, which will improve the quality and usefulness of future genome annotations.

3.2.3. Results and Discussion

The structure of the teleost gap junction protein gene family

The compressed tree with the connexin subfamilies for teleosts and mammals is shown in Fig. 1. All sequences involved are shown in Suppl. Fig. 1-12. A few of the expanded branches are shown in Figs. 2-6 (Fig. 2, *gjb7*; Fig. 3, *gja4*; Fig. 4, *gjd2*; Fig. 5, the "*gjb4like*" complex; Fig. 6, *cx39.2*), and the remaining branches are shown in Suppl. Fig. 14. In this tree, and in all trees made for the major statistical analyses (Suppl. Table 1), the *GJE1*/*gje1*/*cx23* group was omitted, because the inclusion of the *GJE1* orthologous group caused long-branch attraction [45, 46]. In fact, the long-branch attraction was so intense that it ripped apart both the delta and gamma subfamilies, and caused the highly variable groups of *GJC3* and *GJD4* to locate in the vicinity of the *GJE1* group (compare Fig. 1 and Suppl. Fig. 15). However, we did include a human pseudogene in the *Cx39.2* group (Fig. 6), but not the corresponding pseudogenes from some other mammals. This orthologous group is further discussed below. We also excluded rodent *gja6* (which is the ortholog of the human pseudogene sometimes called *Cx43pX* [28]) and a cod *gjd2* sequence (Gm-NN-*gjd2*1*-G01582). This sequence often split out from its expected *gjd2* group, and we excluded it to make clearer distinctions within the different *gjd2* groups.

Figure 1. Phylogenetic tree for the gap junction protein (connexin) gene family.
The mammalian branches are indicated by upper case letters; teleost branches are indicated by lower case letters. The width of the triangles indicates the number of taxa included in the branch, and the length of the triangles indicates the sequence variation within the branch. The tree was made by the Minimum Evolution method, using amino acids (354 amino acid sequences with 201 positions in the final dataset) and the Dayhoff substitution matrix. The bootstrap values (500 replicates) >50% are shown next to the branches. To avoid disruptive long-branch attraction, some sequences were excluded (see text). This model gives results that are quite close to the majority of results as summed up in Suppl. Table 1, and thus is close to an average tree from all the tests run. The major difference is that the mammalian GJA10 and teleost gja10 have switched places. In the original three, the root of the gjd family splits up in three very close branches, but using the rooting function in the Mega Tree Explorer collected them into one common basal branch. Note the commonly occurring dichotomy with the mammalian sequences in one of the sub-branches and the teleost sequences in the other sub-branch, although some of the teleost groups do not have a mammalian counterpart (and vice versa). The scale bar (lower left) indicates the number of amino acid substitutions per site.



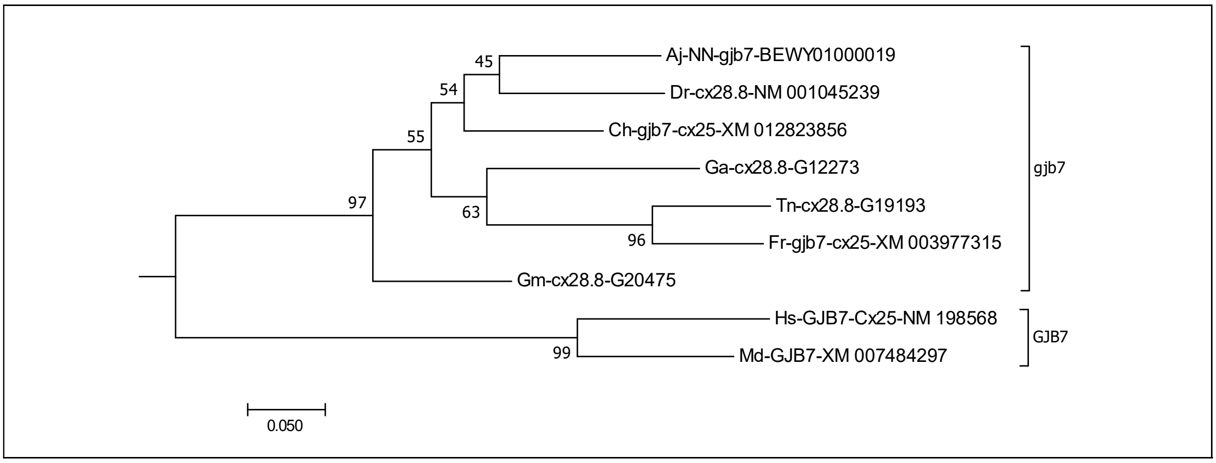


Figure 2. The *GJB7*/*gjb7* branch from the compressed tree shown in Fig. 1. This is an example of a group where all teleost species have only one member, and therefore probably have lost the expected ohnolog partner at a very early stage before the divergence of the different teleosts, similar to most of the other connexins located on the same chromosome (see Table 2).

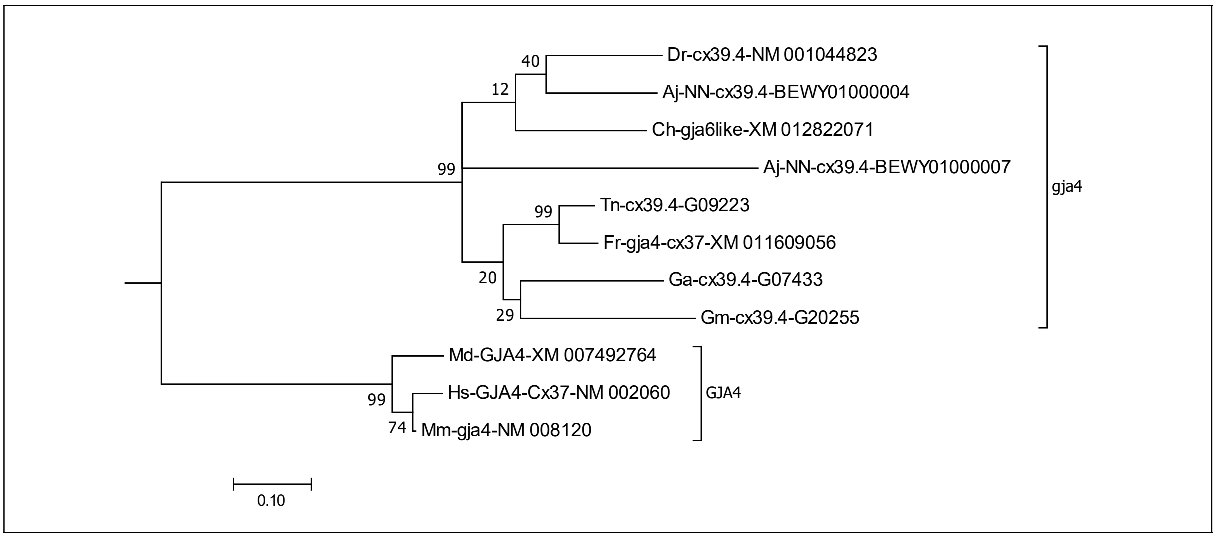


Figure 3. The *GJA4*/*gja4* branch from the compressed tree shown in Fig. 1. This is an example of a group where eel has two members, whereas all the other teleosts have one member. The eel pair is found on two different chromosomes (Table 2), suggesting that one member was lost somewhere in-between the divergence of eels and the other teleosts. Moreover, note that the herring member is wrongly named *gja6like* in GenBank; the correct name would be *gja4*.

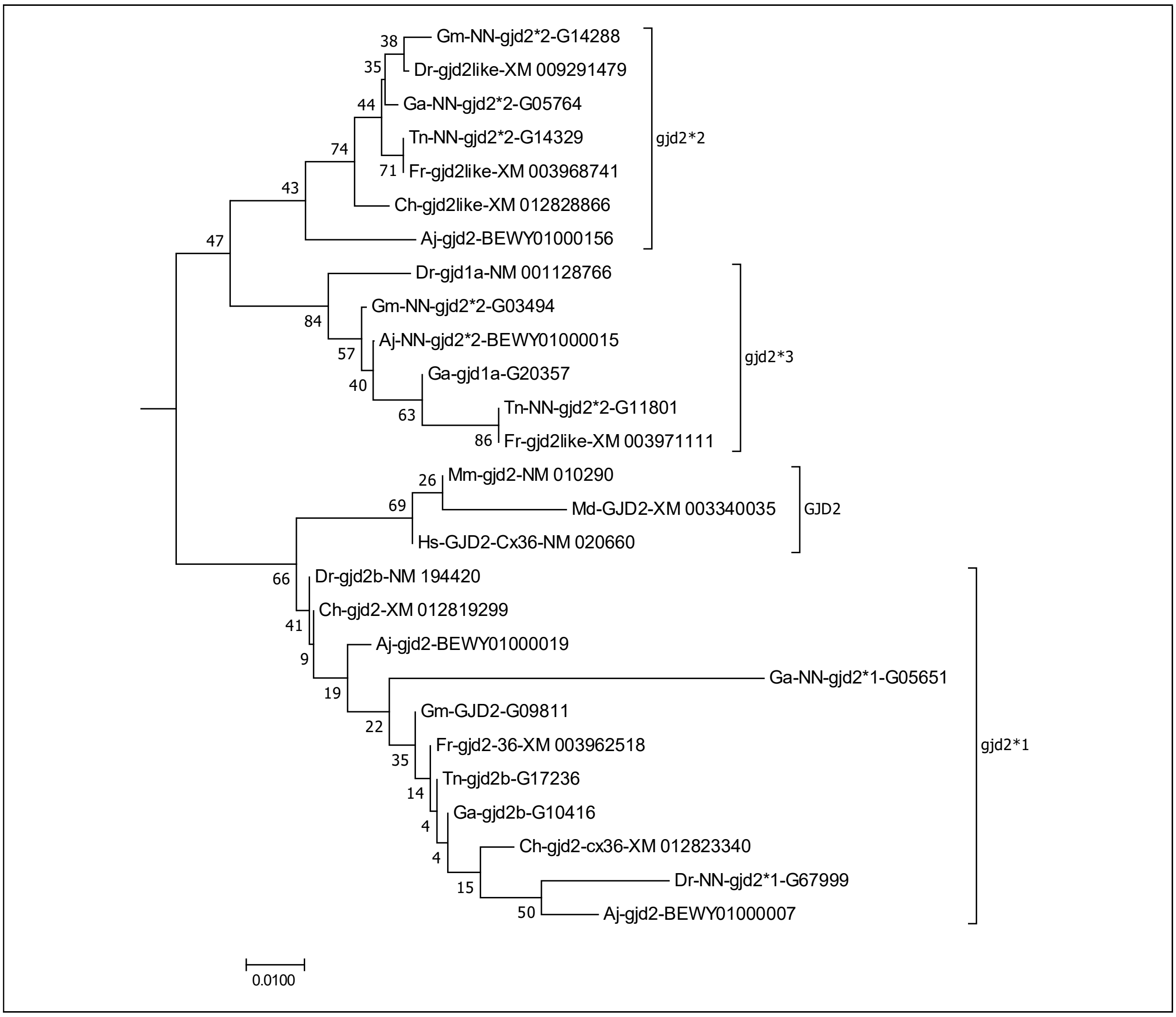


Figure 4. The *gjd2* branch from the compressed tree shown in Fig. 1. This is an example of a group where the structure is considerably more complex in teleosts than in mammals. First, there is one teleost group, here called *gjd2*1*, that in the majority of statistical models locates closest to mammalian *GJD2*. *Gjd2*1* contains two sequences from most fishes, and each members of the pairs are on different chromosomes in all species (Table 2). Secondly, there are two subgroups (here called *gjd2*2* and *gjd2*3*) that are, according to this statistical model, slightly more distantly connected to mammalian *GJD2*. In this statistical model, the *gjd2*2* and *gjd2*3* subgroups have a phylogenetic distribution that is “ohnologically perfect” in that it divides into two sub-subgroups containing one sequence from each species. In all species, the pairs of sequences are found on two different chromosomes (Suppl. Table 7).

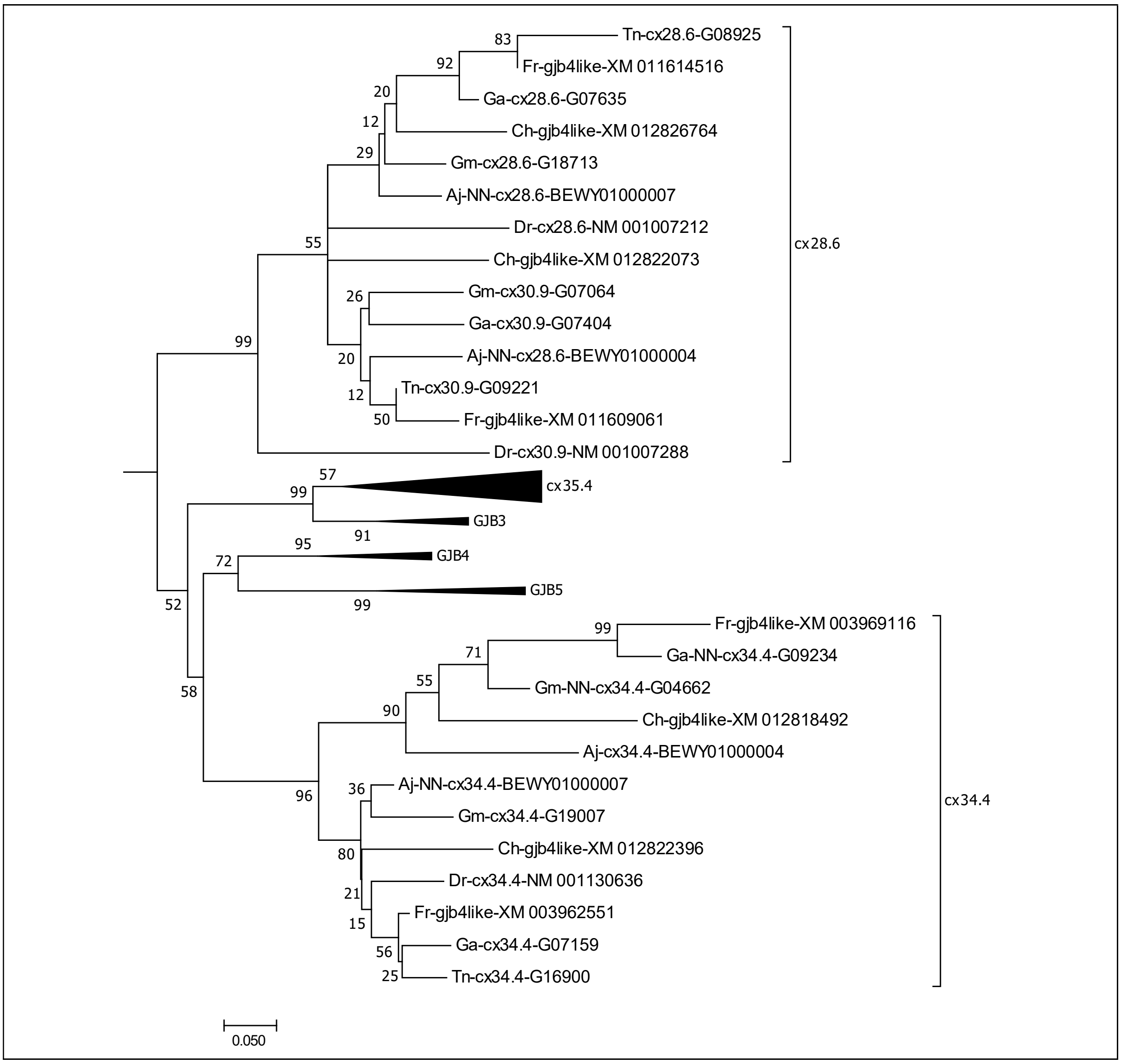


Figure 5. *GJB3*/*GJB4*/*GJB5* related sequences from the compressed tree shown in Fig. 1. This is an example where teleost sequences with the same names are found in clearly distinct branches of the tree. In this case, four Fugu (abbreviated Fr) and four herring (abbreviated Ch) sequences are called *gjb4like*. Two sequences from each species located into each of the two groups here called *cx28.6* and *cx34.4*. Note also that mammalian *GJB4* and *GJB5* were always found as a dichotomous pair, and that *cx34.4* never mixed into the dichotomous *GJB4*/*GJB5* pair (Suppl. Table 1). Similarly, *cx28.6* generally split off at the foot of the collected *GJB3*/*GJB4*/*GJB5*/*cx35.5*/*cx34.4* clade, but in a few cases (with poorer statistics) was positioned closer to *GJB3*/*cx35.4* (Suppl. Table 1). Thus, there is no evidence to support *cx28.6* or *cx34.4* being more closely related to *GJB4* than to *GJB5* as the naming (*gjb4like*) could suggest.

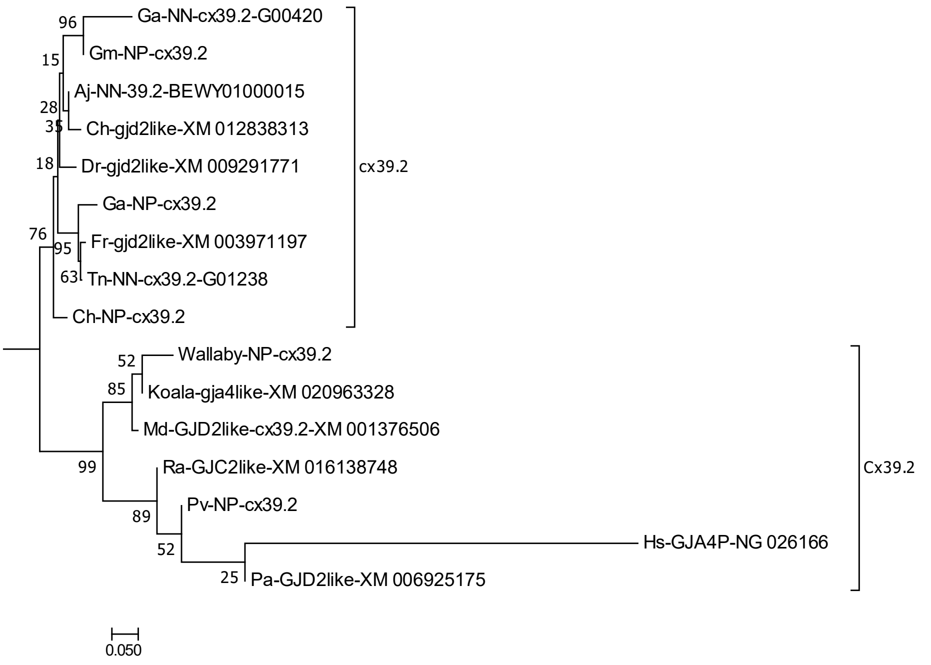


Figure 6. The human pseudogene "*GJA4P*" (NG_02166) always located together with *cx39.2*/*gjd2like* sequences. Note that these "*gjd2like"* sequences must not be confused with paralogous sequences that have the same name in other groups (*cx36.7* and *gjd2*2*).

Overall, it was evident that the structure of the *connexin* subfamilies was similar across all the teleosts. There were examples of species-specific gene duplications or lack of genes, but at the present time we cannot with certainty ascribe all such "anomalies" to biological and genetic reality or to partial genome sequencing and/or erroneous genome assembly. The overall similarity should make it rather simple to extend the gene identifications to other teleost species when their genomes are sequenced, thereby easing their annotation. However, this is dependent on consistency in naming the gene family, which is presently at lack as shown below.

The mixture of nomenclatures

As can be seen in Figs. 2 to 6 (and also in Suppl. Fig. 14 and Suppl. Tables 3-5), there was often little consistency in naming within many of the ortholog or ohnolog groups, as some of the genes were named by the size nomenclature and others are named by the Greek nomenclature. We will here sum up the nomenclature for some of the teleost species.

Zebrafish is undoubtedly the most highly investigated teleost [47], with its genome sequencing starting in 2001, the first genome assemblies available in Ensembl around 2005, with the latest assemblies and annotations from 2017/2018 (Ensembl release 91, CRCz11). Thus, we would expect the gene nomenclature to be of good standard and being consistent with the intentions expressed in the Zebrafish Nomenclature Conventions [5]. In zebrafish, among the 38 unique and predicted genes present in GenBank (Suppl. Table 3 and Suppl. Fig. 5), 25 genes followed the size nomenclature and 13 genes followed the Greek nomenclature. The naming of 37 predicted genes in Ensembl was rather similar to GenBank, with 31 sequences having the same name as in Ensembl (Suppl. Table 3). The differences were that two sequences were not predicted in Ensembl, one sequence was not predicted in GenBank, and three sequences were predicted in Ensembl but were un-named. Only one sequence was clearly named differently, *gja1like* in GenBank and *cx40.8* in Ensembl, although there was one incidence of lower/upper case letters in the Greek nomenclature (*gjd4*/*GJD4*).

Takifugu rubripes, often called Fugu, was the first teleost with its genome published [48], with the last genome assembly from 2011 (in Ensembl) and annotations from 2018 [49]. Before July 2019, there were 42 predicted gap junction protein genes in GenBank (Suppl. Fig. 6 and Suppl. Table 4), three of which followed the size nomenclature, 26 followed the Greek nomenclature, and 13 followed a hybrid nomenclature with both Greek classification and size mentioned. Fugu was recently updated in GenBank (July 2019) and the 13 entries with hybrid nomenclature changed to Greek nomenclature (n many cases also changing accession numbers), but in one case (Fr-*gja3like*-XM_003970457), the prediction was lost in the update. In Ensembl, two of the Fugu genes were named in Greek nomenclature in upper case letters, 14 were named with Greek nomenclature in lower case letters, and 21 were named according to the size nomenclature. Twelve genes could be said to have the same naming in GenBank and Ensembl (not considering upper/lower case letters), using the updated GenBank entries for Fugu (Suppl. Table 4).

For cod sequences in Ensembl (Suppl. Fig. 10, Suppl. Table 5), eight followed Greek nomenclature (six in upper case and two in lower case), 18 followed size nomenclature, 17 were predicted but not named, and one was not predicted (but found by us). The recently available cod chromosome level genome assembly in GenBank [50] and the corresponding gene predictions provided us with the possibility to compare the naming of the new predictions with the Ensembl cod gene predictions (Suppl. Table 5). Only four sequences had been given the same name in Ensembl and GenBank (not considering lower/upper case letters; Suppl. Tables 4 and 6). Additionally, within the new cod chromosome level assembly, there were two genes with no hit and one with a genomic hit, but no gene prediction. The identities between the Ensembl gene sequences and GenBank gene sequences were generally >99.5% (three sequences were <99.5% identity).

In herring (Suppl. Fig. 9), 32 genes followed the Greek nomenclature, four followed the size nomenclature, and eight followed a mixed nomenclature, in addition to two previously non-predicted genes. Only a few of the eel connexins in the GenBank TSA had been named, with several having a hybrid nomenclature not commonly used (such as *CXA5*, *cxb1*, *CXG1*, etc.).

Multiple names for a distinct ortholog within teleosts

There were three common inconsistencies within an orthologous group, two of which are considered in this section, and the third in the next section. The first was that some genes within the group are named according to the Greek nomenclature, and other genes according to the size nomenclature. For example, within the *GJB7* group (also called *connexin25* in mammals), some teleost sequences were named *gjb7* and other sequences were named *cx28.8,* and some combined the Greek and size nomenclature such as *gjb7-cx25* (Fig. 2).

The second inconsistency was that evident orthologs had been given different numbers in the Greek nomenclature. One example was the teleost orthologs for mammalian *GJA4*, also called *connexin37* (Fig. 3). They were called *gja4* in Fugu, *cx39.4* in *Tetraodon*, stickleback and zebrafish, and *gja6like* in Atlantic herring. It should be noted that *GJA6* is a different gene group that was generated by a mammalian-specific gene duplication of *GJA1* (*connexin43*), maybe by retrotransposition. *GJA6* is a pseudogene in humans and some other species (called connexin43-related pseudogene on the X chromosome, *Cx43pX*, in ref. [23, 28]). In other species, including rodents, dog and elephant, *GJA6* appears to be a functional gene [23, 28]. Another example is found within the major *GJD2* group (Fig. 2C). Zebrafish NM_001128766 and stickleback ENSGACG00000020357 (no GenBank entry) were both called *gjd1a*, whereas the orthologs in Fugu were both called *gjd2like* (Fig. 4).

Distinct genes having identical names

The third common inconsistency was that clearly different sequences had the same name. In Fugu (using the predicted GenBank sequences), there were two of each for *Cx32.2like*, *gjb1like*, *gjb2like*, and *gjb3like* genes; three *gja3like* and *gjc1like* genes; four *gjb4like* and *gjd2like* genes (Fig. 4; Suppl. Table 6).

Recently, Atlantic herring (*Clupea harengus*) had its genome sequenced, assembled and annotated [17]. Thus, the prediction and naming of the genes describe much of the current status for automatic annotation. In herring, there were two of each for *gja5like*, *gjd2*, and *gjd3like*; three of C*x32.2*, *gjc1like* and *gjd2like* genes; and four genes called *gja3like* and *gjb4like* (Fig. 4, Suppl. Table 6).

We will use *gjd2like* and *gjd2* as examples. *Gjd2like* was used in several more or less closely related genes in the delta subfamily. More specifically, sequences with this name were found among the *cx36.7*, *cx39.2,* and the central *gjd2* groups. These groups are shortly discussed below.

The central *gjd2* (Fig. 4) is a complex group of sequences that are all closely related to the mammalian *GJD2*. Previously, these genes were named *connexin36* in mammals and *connexin35* or *connexin35.1* [51] in fish. While mammals have one *GJD2* gene, teleosts have up to four (as in zebrafish, Fugu, and stickleback) in this central *gjd2* group. For convenience, we named groups of the teleost genes in the central *gjd2* group as *gjd2*1*, *gjd2*2* and *gjd2*3*, because they sometimes split into three groups, depending on the statistical analysis. Sometimes, one or two sequences split out of the *gjd2*1* group, and ended in-between the other *gjd2*/*GJD2* groups. This happened particularly often with Gm-NN-*gjd2*1*-G01582 (sequence found in Suppl. Fig. 10), which is why we excluded this sequence during the statistical analyses. Generally, the sequences within *gjd2*2* and *gjd2*3* stayed as unified groups, usually as a dichotomous clade (for discussion of ohnologies within these groups, see below).

The mammalian *GJD2* is somewhat promiscuous in terms of which teleost sequence group it most closely adhered to, but most often it was *gjd2*1* or *gjd2*2*. In zebrafish, these genes are among the few places where "a" and "b" have been added to some of the gene names in the databases. In the *gjd2*2/*3* group, one of the zebrafish (and stickleback) genes is called *gjd1a* (but there is no *gjd1b*) and the other *gjd2like*. In the *gjd2*1* group, one of ohnologs in zebrafish, *Tetraodon*, stickleback and cod is called *gjd2b* (but there is no *gjd2a*).

Another group named *gjd2like* (in Fugu and Atlantic herring) was the *cx36.7* group, called *Dr17927* in a previous paper [24]. This group often branched off from the foot of the central *gjd2* complex itself (Fig. 1), but in a few statistical analyses it located closer to *gjd3* or *gjd4* (Suppl. Table 1). As yet, there are no mammalian members in this group, and our previous work [24] suggested that this group was specific to fish.

Another orthologous group often named *gjd2like* has previously, and more uniquely, been called *cx39.2* [28]. This orthologous group divided its location between the delta (most commonly) and gamma subfamilies depending on the model run, but it never located within or at the foot of the central *gjd2* group (in contrast to *cx36.7*). The first mammalian member in the *cx39.2* group was found in opossum [28], but here it is shown that this ortholog is also present in several other mammals, like bats (Fig. 6). A human pseudogene (NG_026166), named "Homo sapiens gap junction alpha 4 pseudogene on chromosome 17" (*GJA4P*) is not a pseudogene related to *GJA4* but rather to the *cx39.2* (*GJD2-like*) group according to the phylogenetic analyses (Fig. 6). Alignments of NG_026166 against *GJA4* and representatives from the *cx39.2*-like group clearly indicated a closer relationship with the latter (Table 1; Suppl. Figs. 13A and 13B; Suppl. Tables 7 and 8). In a comparison at amino acid level between the conserved domains of human GJA4P and GJA4 vs. eel cx39.2 and cx39.4 (Table 1), the identity levels between the GJA4/cx39.4 (human/eel) orthologs were ~55%, the same as for GJA4P/cx39.2 (human/eel), which is clearly higher than GJA4P/GJA4 (human/human; ~38%) and GJA4P/cx39.4 (human/eel; ~34%). Also at nucleotide level, the human *GJA4P* showed higher identities to *cx39.2* orthologs than to *GJA4* orthologs (Suppl. Table 8), e.g., conserved domains of Hs-*GJA4P* was 53.9% identical to opossum *GJA4*-XM_007492764 and 65.3% identical to opossum *cx39.2* (= Md-*GJD2like*-XM_001376506) (Suppl. Table 8). Thus, the alignments were consistent with the phylogenetic results (Figs. 1 and 6), and settled this pseudogene (NG_026166) to be incorrectly named in humans, and is not *GJA4P*, but rather *Cx39.2P*.

Table 1. Conserved domains of human "GJA4P" are more similar to cx39.2 than to GJA4 at amino acid level.

	Hs-GJA4P	Aj-39.2	Hs-GJA4	Aj-39.4-1	Aj-39.4-2
Hs-GJA4P–NG_026166	100.00	54.92	37.82	34.72	33.16
Aj-NN-cx39.2	54.92	100.00	48.70	40.41	44.56
Hs-GJA4-Cx37	37.82	48.70	100.00	53.89	54.40
Aj-NN-gja4-cx39.4-1	34.72	40.41	53.89	100.00	68.21
Aj-NN-gja4-cx39.4-2	33.16	44.56	54.40	68.21	100.00

Human *GJA4P* was aligned as well as possible to all other connexin sequences at nucleotide level before being translated. The alignment is shown in Suppl. Fig. 13B. Note that the identity between eel cx39.2 and human GJA4P is around 55%, which is at the same level as the identity between eel cx39.4 (gja4) and human GJA4. Further note that the identities between eel cx39.2 and eel gja4 are around 40%, which is the same level as the identity between human GJA4P and human GJA4. This is consistent with the results shown in the phylogenetic analyses elsewhere in this paper. Thus, human GJA4P (NG_026166) is incorrectly named, and is in reality a Cx39.2P sequence.

On teleost connexin ohnologies

The phylogenetic analyses provided a strong indication of the presence of several ohnologous pairs in teleosts. However, distinguishing between paralogous pairs that have been created by tandem gene duplication and ohnologous pairs created by genome duplication might difficult, especially if the assembly only exists as contigs or relatively short scaffolds. If a novel teleost genome assembly is being made, it would be valuable to have the answer to this question established in other species, simply because the naming should be different in the two cases. Thus, it is of importance to show whether the ohnologous relationship can be traced across teleosts in a reasonably systematic way. In other words, is the genomic location of a gene and its potential ohnolog in one or two species sufficient to give indications for other species?

As of today, most eukaryotic draft genome assemblies consist of thousands of scaffolds, and even if these scaffolds can be Mb long, they are just a fraction of the size of most eukaryotic chromosomes. For such scaffolds, only connexin genes positioned rather closely are informative. When this analysis started, chromosomal assemblies were not available for herring and cod, but both became available during the summer of 2019 [50, 52].

Table 2. Ohnologies of teleost connexins harbored at the most connexin-rich chromosomes.

Ohnolog A							Connexin	Ohnolog B						
Tn	Fr	Ga	Gm	Dr	Ch	Eel		Eel	Ch	Dr	Gm	Ga	Fr	Tn
-	18	18	21	20	15	19	*gje1*	-	14	-	-	-	-	-
14	1917	18	21	20	15	19	*cx34.5*	-	-	-	-	-	-	-
14	1917	18	21	20	15/15	19	*cx32.2*	-	-	-	-	-	-	-
14	1917	18	21	20/20	15	19	*cx28.9*	-	-	-	-	-	-	-
?	1725	18	21	20	15	19	*gja1*	7	14	17	7	-	-	-
?	16	18	?	20	14	19	*gja10*	-	19	-	5	128	1843	?
14	1688	18	21	20	?	19	*gjb7*	-	-	-	-	-	-	-
10	2	18	21	20	15	19	*gjd2*1*	7	14	17	5	15	-	-
21	7	10	22	17	19	7	*gja9*	sc68	?	16	6	20	12	?
-	2	18	21	17	15	7	*gjd2*1*	19	14	20	5	15	-	10
-	-	18	21	17	19	7	*gja1*	19	14	20	7	-	1725	14
-	2	10	22	17	19	7	*cx35.4*	4	14	-	5	15	12	10
-	2	10	22	17	19	7	*cx34.4*	4	14	-	5	15	12	10
21	10	10	22	19	19	7	*gja4*	4	-	-	-	-	-	-
21	10	10	22	17	19	7	*cx28.6*	4	14	19	5	15	15	10
7	14	4	10	5	8	8	*cx39.9*	15	20	-	7/7	7	15	1
7	14	4	10	5	?	8	*gjb1*	15	20	14	7	7	15	1
2	16	-	-	9	2	8	*gja8*	14	21	1	20	1	-	-
7	16	6	?	9	2	8	*gja5*	14	21	1	-	16	16	17
2/2/2	1/1	1	4	9	2	8	*cx30.3*	14	8/21	-	20	115	8	3
2	1	1	4	9	2	8	*gja3*	14	21	-	20	115	8	3

Using eel as a starting point, the chromosomes/linkage groups/scaffolds with the highest number of connexin genes were identified. The chromosomal location of the corresponding orthologs was identified in the other species (left part of the table). Subsequently, the chromosomal location of the ohnologous genes was identified in the same species (right part of the table). The order of the genes is given by their location on the eel chromosomes. Among the sequences mentioned in this table, there are five obvious examples of tandem gene duplications, indicated by several identical numbers with slashes in-between (e.g., 2/2/2 for *cx30.3* ohnolog A in Tn) and one example of a presumed gene duplication located to different chromosomes (*cx30.3* ohnolog B in Ch). ?, the sequence is unplaced; sc68 and three or four digit numbers indicate scaffold number.

For looking more closely into ohnologous pairs, the Japanese eel genome assembly was used as a starting point, because eel is a member of the early diverging fishes. Table 2 summarizes the situation for the chromosomes (or linkage groups) containing the highest number of connexin genes, and Suppl. Table 8 gives the full overview. Eel linkage group (chromosome) 19 contained eight connexin genes (*gja1*, *cx34.5*, *cx28.9*, *cx32.2*, *gja10*, *gjb7*, *gjd2*1*, *gje1*). The same eight genes were found on zebrafish chromosome 20 and stickleback chromosome 18, and at least seven of them are collected at cod chromosome 21. Thus, there is a strong tendency that linked genes in eel also are linked in the other species. For some unknown evolutionary reason, this chromosome had relatively few examples of ohnologs. The ohnologous chromosome may have gone through some kind of genetic catastrophe. In fact, for the two connexins with the highest number of species showing ohnology, *gja1* and *gjd2*1*, the ohnologs were found on an "unexpected" chromosome (7 in eel, 14 in herring and 17 in zebrafish), because these ohnologs deviated more from the location patterns we found for the other connexins on eel chromosome 7.

Eel chromosome 7 contained five connexin genes, in addition to the ohnologs of *gja1* and *gjd2*1*, namely *gja4*, *gja9*, *cx28.6*, *cx35.4* and *cx34.4*, and four of their ohnologs were placed at chromosome 4 (and chromosome 19 for *gja1* and *gjd2*1*). In stickleback, all five genes were found on chromosome 10 (but chromosome 18 for *gja1* and *gjd2*1*), and three of the ohnologs were found on chromosome 15, the fourth ohnolog (*gja9*) on chromosome 20, and the fifth was missing. In *Tetraodon*, three of the five genes (*gja4*, *gja9*, *cx28.6*) were found on chromosome 21, and two of these had ohnologs, *gja9* on an unplaced scaffold, and *cx28.6* on chromosome 10. *Tetraodon* chromosome 10 also contained the single copies of *cx35.4* and *cx34.4*. In zebrafish, *gja9*, *cx28.6*, *cx35.4*, and *cx34.4* were found on chromosome 17. *Gja4* was present as a single paralog on chromosome 19, which also contained the ohnolog of *cx28.6*. Thus, we see for *gja4*, *gja9*, *cx28.6*, *cx35.4* and *cx34.4* on eel chromosome 7 that there was a strong tendency towards a pattern of consistency in distribution of ohnologous pairs to distinct chromosomes in all the investigated species, while *gja1* and *gjd2*1* tended to deviate.

In general, teleosts had four genes that were very closely related to mammalian *GJD2*. Although one or two of the sequences in the *gjd2*1* group occasionally split out from the remaining genes, the two ohnologs (Table 2) generally stayed together, and there should be no doubt about the proper ohnology. In 14 of 21 statistical analyses *gjd2*1* grouped together with mammalian *GJD2*, and these were considered as the appropriate orthologs. *Gjd2*2* and *gjd2*3* often dichotomously grouped together (in 11 of 21 statistical analyses), but other times split up. We believe that *gjd2*2* and *gjd2*3* most likely are ohnologs, although it could not totally exclude the possibility that they are non-ohnologous paralogs located on different chromosomes.

If the genes that were linked in eel had broken linkages in other species, in many cases two or three of the most closely linked genes have moved to another chromosome than the rest of the group. A more complete overview containing all connexin genes and associated chromosomes is provided in Suppl. Table 8.

Of course, this analysis also showed closely related genes that were not ohnologs. *E.g.*, the genes within the *cx34.5* and *cx32.2* groups (also known as *cx32.7*, *cx32.3* and *cx28.9*) are not ohnologs, because they all are located on the same chromosome (19 in eel, 15 in herring, 20 in zebrafish, 21 in cod, 18 in stickleback, 14 in spotted pufferfish, and scaffold1917 in Fugu).

In summary, over the range of divergence time, large stretches of the chromosomes have been maintained reasonably intact subsequent to the teleost genome duplication. Thus, the corresponding ohnologs are found on other non-random chromosomes. However, both gene losses and tandem duplications might have occurred over the considered evolutionary period, which could complicate the interpretations. Of course, this is even further complicated by the facts that the sequencing itself is probably not able to reach a complete coverage of the genome causing the partial or full absence of a gene, and that the assembly process is not straight-forward.

As an example of practical use of this kind of information, we here briefly apply the knowledge of the outlined patterns of the connexin genes on (i) the first published herring genome assembly [17, 53], which has been used as basis for gene predictions (XM accession numbers in GenBank); (ii) the new herring chromosome level assembly [52, 54]; and (iii) a herring genome assembly made by the present authors (S. í Kongsstovu *et al.*, submitted) [55]. Although the herring gene predictions were superior when compared with most other fishes (in the sense that the predictions tended to follow the expected gene patterns), there were still some features worth noting.

- First, there was one easily found connexin (*cx39.2*) that was not predicted in the annotation from the first herring genome assembly.

- Second, several connexin genes showed identical or near identical duplicates in the first herring genome assembly. The *gjb3like*-XM_ XM012822385 (one of the ohnologs in the *cx35.4* group) was identical to XM_012822374 and XM_012822365, found at three locations on scaffold NW_012217989. The *gjb3like*-XM_012818491 (the second ohnolog in the *cx35.4* group) was identical to XM_012818489; found at two locations on scaffold NM_012210726. The *gjb4like*-XM_012818492 (one of the ohnologs in the *cx34.4* group) was nearly identical to XM_012819490, and both were found on scaffold NW_012219726. The *gjd3like*-XM_012837668 was nearly identical to XM_012837669, and both were found on scaffold NW_012223269. Although such copies are not entirely biologically implausible, they are not probable, and are more likely caused by assembly errors. Indeed, in the initial states of our own assembly most of them were not present in duplicate sequences, only becoming so in the last step where our assembly was fused with the published herring genome (S. í Kongsstovu *et al.*, submitted). In the recently (summer 2019) released herring chromosome level assembly [52, 54] most of these duplicates have collapsed into a single copy of the sequence.

- Third, three connexin genes have "disappeared" from the new herring chromosome level assembly. These are *gja9like*-XM_012824682, *gjb1like*-XM_012819602 and *gjb7like*-XM_012823856. The corresponding orthologs are found in the other teleost species, and - even more importantly - hits were found in the two other herring genome assemblies. We have verified the presence of these genes in our early assemblies (S. í Kongsstovu *et al.*, submitted). This strongly indicates misassemblies in the new chromosome level assembly. More specifically, the lack of *gjb7* indicates a misassembly on chromosome 14 or 15, and, indeed, an alignment of the relevant scaffold and chromosome alignments show breaks and inversion around the expected position of *gjb7* at chromosome 15 (Fig. 7). The apparent lack of the *gjb1* ohnolog indicates a misassembly on chromosome 8, where we indeed found breaks and inversions (not shown). We expected that the lack of the *gja9* ohnolog to indicate a misassembly on chromosome 14, but we found the relevant scaffold to align with chromosome 11, where again breaks and inversions were found (not shown).

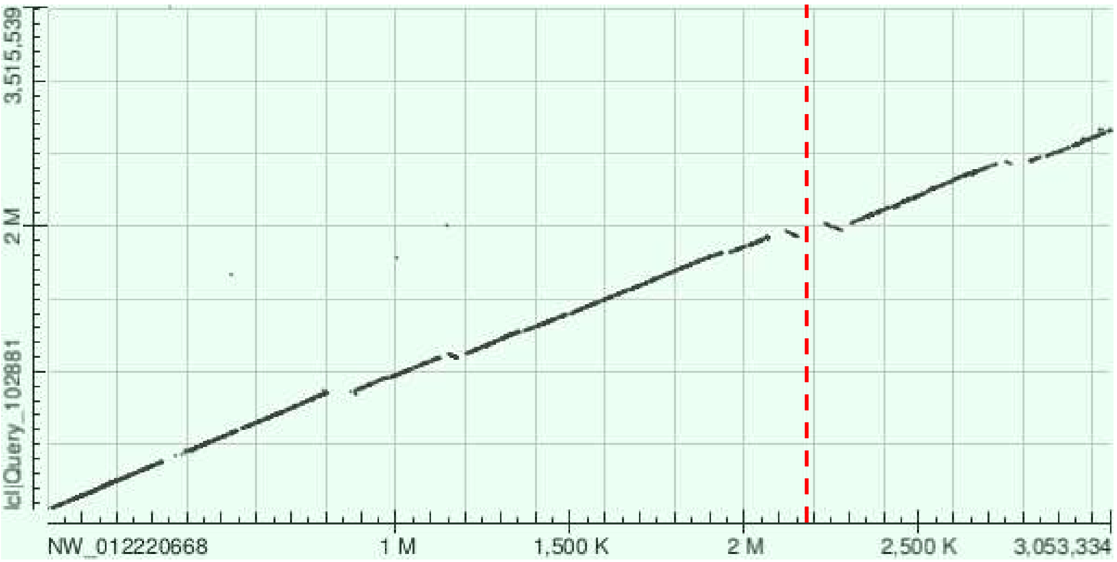


Figure 7. Problem in herring assembly of chromosome 15 at assumed position of *gjb7*. Scaffold NW_012220668 from the draft herring genome assembly contains *gjb7* in position 2189757-2188978 (*i.e.*, on the reverse strand). This scaffold was aligned with herring chromosome 15 assembly LR535871 position 0 to 3,500,000 using the alignment option in Blast. The position of *gjb7* on NW_012220668 is indicated by the red dotted line. There are apparent inversions and breaks in the area where *gjb7* was expected in chromosome 15. The word size in the alignment was 256.

Regarding the third point above, also the recent chromosome level assembly in cod [50] showed a "no hit" for the Gm-*gja10* ohnolog Gm-*cx52.6*-G05425 and for Gm-*gja5*-G04028 (Suppl. Table 4 and 8). The lack of *gja5* suggested a problem in the assembly of cod chromosome 20 around position 1,000,000 (Suppl. Fig. 16A). Gm-*cx52.6* is located on a small and unplaced contig (not even containing the full-length sequence of the gene), which was unusable for dot plot alignment at a chromosomal scale. By using suitable scaffolds containing the *cx52.6* ortholog from herring and stickleback, we believe there is a problem in assembly of cod chromosome 21 around position 2,700,000 (Suppl. Fig. 16B and C). Also other alignments with the corresponding zebrafish sequence pointed to the same location.

Our present analyses share some common grounds with Core Eukaryotic Gene Mapping Approach (CEGMA) [56] and Benchmarking Universal Single Copy Orthologs (BUSCO) [57, 58], in that selected genes are investigated for their presence in a genome to verify the completeness of a newly assembled genome. They differ in that it is a multimember family of genes, consisting of several subfamilies, as well as that the genes have two conserved domains that have some reciprocal similarities. In the context of genome assembly and gene annotation, the connexins are a randomly selected gene family. It is curious how, even in very recent high-quality genome assemblies, such as those of Atlantic herring and cod, these genes can indicate certain potential misassemblies. This situation can possibly be extended to other gene families and single genes, as the number of missing BUSCOs in the herring genome in the herring genome increased from 2.9% (131/4584) in the draft herring genome assembly [17] to 8.1% (374/4584) in the chromosome level assembly [54] according to our analysis (S. í Kongsstovu *et al.*, submitted).

We believe that improved gene predictions and annotations are possible through the proper incorporation of knowledge into the algorithms. Furthermore, it would certainly help if the genes were labeled with unique names, as is one of the underlying logics in the instructions from the Human Gene Nomenclature Committee and the Zebrafish Gene Nomenclature Conventions. A more consistent nomenclature suggestion is described in the following section.

A more consistent nomenclature suggestion

For most of the genes in the teleost connexin family, it is easy to suggest names that follow the nomenclature guidelines. Suppl. Fig. 17 presents a suggestion. Here we maintained the Greek nomenclature naming and numbering of those genes that have well established names in human and mouse, and the corresponding orthologs in teleosts. We fully avoided the "-like" names, as they often are used for several distinct genes and thus do not indicate a concrete orthologous group, and in this way can be misleading.

The subfamily number (*gjd1/2/3/4*, etc.) for the groups where new names are suggested does not consider the chronological order of detection, but rather the numbers that are available. For example, *cx39.9* is closely related to *gja3*, and is in fact often called *gja3like*. As *gja1*, *gja3*, *gja4, gja5,* and *gja6* already are occupied, while *gja2* is not, we suggest calling the present *cx39.9/gja3like* for *gja2*. The genes in the *cx34.5*, *cx28.9* and *cx32.2* groups are called *gja11*, *gja12* and *gja13*, respectively. We skip *gja7*, as this name has historically been used for *Cx45* (= *GJC1*).

In the beta subfamily, there is a particular problem in that the mammalian *GJB2* and *GJB6* are always located in a dichotomous manner, and similarly for *GJB4* and *GJB5*. There were no indications that *cx30.3* located closer to either of *GJB2* or *GJB6*, and similarly, *cx34.4* did not locate closer to either of *GJB4* or *GJB5*. It might be that *cx30.3* is a precursor gene for both *GJB2* and *GJB6*, and *cx34.4* is a precursor gene for *GJB4* and *GJB5*, as we have suggested earlier [23, 24]. Thus, several possibilities could exist for naming these genes, such as *gjb8* (following the present pattern in the Greek nomenclature), *pre-gjb2/6* (indicating the potential of being a precursor for the two mammalian genes), or *gjb26* (a variant of the previous, but with the potential danger that this could be mistaken for *cx26*).

Statistically, a strong link exists between *cx35.4/gjb3like* and *GJB3*. We therefore suggest that *cx35.4* should be called *gjb3*, despite the lack of the hallmark in the mammalian GJB3 protein, namely the amino acid sequence CX_5CX_5C in the second extracellular loop, where all other connexins (except the GJE1 proteins) have the sequence CX_4CX_5C.

In the gamma subfamily, there are two groups concerned with renaming. The first one is in marsupials, where the majority of statistical analyses (Suppl. Table 1) support *GJC1like/GJC2like* genes probably being the orthologs of eutherian *GJC3*, as originally suggested [28]. The second group is *cx43.4/44.2/gjc1like*, which we suggest is renamed *gjc4*. In the delta subfamily, the major problems concern the *gjd2* complex. As briefly discussed above, we consider *gjd2*2* and *gjd2*3* probable ohnologs, and suggest that they are named *gjd1*, fitting with a zebrafish and a stickleback sequence within this group already named *gjd1*. The ohnolog pairs within *gjd2*1* are probably orthologs with mammalian GJD2, and consequently we suggest they are named *gjd2*. The teleost *cx36.7/gjd2lik*e group never dichotomized with any of the mammalian genes and most often branched off from the root of the *gjd2* complex. We suggest this group should be called *gjd5*. The last group is the little-studied *cx39.2* group, which in mammals has a variety of names in database gene predictions, such as *GJC2like*, *GJD2like* and *GJA4like*. The mammalian genes robustly dichotomize with the corresponding teleost genes, which in the databases usually are called *gjd2like*. We suggest that this clade is called *GJD6* in mammals (thus, the human pseudogene NG_026166 should be called *GJD6P*) and *gjd6* in teleosts.

Conclusions

The practice of naming connexin genes in teleosts exhibits many inconsistencies. Commonly, distinct genes are assigned the same name, and there are examples of clearly incorrect names, even in mammals, including that of a human pseudogene (NG_026166). By using many different phylogenetic models and substitution matrices, we were able to define teleost sequences that had a dichotomous relationship with the corresponding mammalian sequences, and thereby point out the sequences that should have the same name as their mammalian orthologous counterpart. Conversely, if there was no mammalian counterpart they should have a unique name. It was further settled which of the teleost sequences that existed in ohnologous pairs, and thereby should have their names followed by "a" or "b". To quite some extent, it is possible to predict on which chromosome a teleost connexin should be located. We investigated two very recent high-quality chromosome assemblies (herring and cod), finding that roughly 5% of the expected connexin sequences were absent (two in cod and three in herring). We found likely misassemblies or gaps at the expected positions for the missing connexins in the chromosome assemblies.

3.2.5. Methods

Collection of sequences

We used only the coding part of the genes, in particular the conserved parts (explained below). Previously collected sequences [24, 28] were checked against the present and updated versions of the genomes to include potential revisions of the gene sequences. Additionally, previously undetected sequences were included. If the experimentally confirmed or predicted sequences were available in GenBank, their accession numbers were also collected (to ensure the unique naming of the sequences). Depending on species and gene in question, we have used the NCBI Reference Sequences whenever possible. Otherwise, gene/RNA names or numbers were collected from Ensembl. All sequences, with GenBank accession numbers or Ensembl gene numbers if relevant, are provided in Supplement Figs. 1 – 12.

Among teleosts, we have collected sequences from zebrafish (*Danio rerio*, abbreviated Dr), stickleback (*Gasterosteus aculeatus,* Ga) [59], Japanese pufferfish (*Takifugu rubripes,* Fr; called Fugu in the text) [48, 60], green spotted pufferfish (*Tetraodon nigroviridis,* Tn) [61], Atlantic herring (*Clupea harengus,* Ch) [17, 53], Atlantic cod (*Gadus morhua*, Gm) [39, 62] and European, American or Japanese eel (*Anguilla anguilla*, Aa; *Anguilla rostrata*, Ar; or *Anguilla japonica*, Aj). For eel, we have chosen to refer to an improved *Anguilla japonica* assembly [63, 64] because it has by far the longest scaffolds, aided by other genome shotgun assemblies of *A. japonica* [65], *A. anguilla* [41] and *A. rostrata* [66], as well as transcriptome shotgun assemblies (TSA) from *A. anguilla* [67-69] and *A. japonica* [70].

As a comparison for the fish sequences, and to follow the Zebrafish Nomenclature Conventions [5], we collected sequences from humans (*Homo sapiens,* Hs), mouse (*Mus musculus,* Mm), and opossum (*Monodelphis domestica,* Md), and supplemented them with certain single sequences from platypus (*Ornithorhynchus anatinus*, Oa), koala (*Phascolarctos cinereus*), Tasmanian devil (*Sarcophilus harrisii*, Sh), wallaby (*Notamacropus eugenii*), large flying fox (*Pteropus vampyrus,* Pv), black flying fox (*Pteropus alecto*, Pa), Egyptian rousette (*Rousettus aegyptiacus*, Ra), aardvark (*Orycteropus afer afer,* Afer), manatee (*Trichechus manatus*, Tm), African elephant (*Loxodonta africana*, La) and armadillo (*Dasypus novemcinctus*, Dn). All sequences are given in the Supplemental Information, where also the relevant database can be inferred according to the name/identity we have given the sequence.

Suggested deviations from the predicted sequences are indicated in the Supplemental Information. If the predicted sequences did not contain potential start and stop codons, we analyzed the genomes to extend the sequences to those codons, following the pattern established by connexins orthologs in other species. If the predicted sequences contained introns, we investigated whether moving the exon-intron borders improved the similarity between sequences and the established sequence patterns, even by including the whole intron as a part of the exon. In a few cases, we also suggested other types of modifications, following the patterns established for these sequences in other species. Furthermore, any unpredicted sequences (*i.e.*, those not predicted in Ensembl or GenBank) we found during the present searches, were included.

Several pseudogenes exist in the gap junction gene family, also in humans [28]. With a single exception, obvious pseudogenes are not included in the present analyses. The one exception is a novel human pseudogene (GenBank NG_026166; claimed as *GJA4* pseudogene) that we did not detect in our previous analyses [23, 24, 28]. Additionally, orthologs to NG_026166 were extracted from the genomes of several mammalian species (Suppl. Fig. 12).

Naming terminology

To ensure uniqueness of every name used in the present work, we added the GenBank accession number or an abbreviated form of the Ensembl gene number to the names for which predictions were available. The disadvantage of using the Ensembl gene number is that it is unstable, and future updates may cause changes in the number.

Specific gene names were generally abbreviated as indicated by the database, or the abbreviations can be inferred from the database name. *E.g.*, for XM_003965660, the full name ("definition") is "Takifugu rubripes gap junction protein, alpha 9, 59 kDa (gja9), mRNA". In this case, the name is given with both the Greek and size nomenclature, and the name is abbreviated in lower case in parentheses. Thus, we have here used the gene name Fr-*gja9-cx59*-XM_003965660. For XM_021466745, the full name is "Danio rerio connexin 55.5 (cx55.5), transcript variant X1, mRNA". We here abbreviated the name to Dr-*cx55.5*-XM_021466745. For XM_011619942, the full name is "Takifugu rubripes gap junction alpha-10 protein-like (LOC1010664818), mRNA", and it was abbreviated Fr-*gja10like*- XM_011619942. Where several transcript variants are experimentally shown or predicted, we only used transcript variant X1.

If the gene was predicted in the Ensembl database, but no name was available, we used a relevant gene name to indicate the correct group of sequences. For example, the *Tetraodon gjb2/6*-like sequence ENSTNIG00000010340 (with the corresponding transcript prediction ENSTNIT00000013438) had no name or description. We abbreviated the gene Tn-NN-*cx30.3*-G10340 (where NN = No Name). This is an example of a gene for which our transcript prediction differed from the database, as indicated in the Supplemental Information.

If the gene was not predicted in a species, but found in our Blast searches, it was suitably named but with the prefix NP (Not Predicted). One example is Tn-NP-*cx30.3*. Thus, *Tetraodon* has a total of four genes in the *Cx30.3* group, two that have been predicted and are named in Ensembl, one that has been predicted but not named, and one that has not been predicted by the database (but by us).

To be able to follow certain very closely related groups of sequences in an easy manner, previously un-named (or unpredicted) sequences in the *cx30.3* and *gjd2* groups were named with the postfixes *1/*2/*3 for the purposes of the present manuscript.

Phylogenetic analyses

The phylogenetic analyses were performed in MEGA7 [71] or MEGA-X [72] using the conserved domains essentially as described in Cruciani and Mikalsen [24] because of the distant evolutionary relationship between mammals and fish. Here, we extended the previously defined conserved domains by 15 nucleotides in 3'-direction for the first conserved domain (*i.e.*, into the sequence corresponding the intracellular loop), and by 15 nucleotides in both 5'- and 3'-direction for the second conserved domain. All sequences and the limits of the sequences used in the phylogenetic analyses are presented in the Suppl. Fig 1-12, where previously defined conserved sequences [24] are marked in yellow, and the 15 nucleotide extensions are marked in gray.

The main questions for the phylogenetic analyses were related and also partly overlapping, and were as follows: (i) The connection between the naming of the teleost sequences (naming taken from the main databases GenBank and Ensembl) and their position in a specific orthologous group, *i.e.*, do teleost orthologs have the same name? (ii) The (orthologous) relationships between the teleost sequences and the corresponding mammalian sequences. Is there a (reasonably) stable structure in the connexin gene family across the teleosts, *i.e.*, do teleost connexins distribute into orthologous groups in a manner more or less similar to the mammalian sequences? (iii) The ohnologies among the teleost sequences. Note that our present questions do not concern the relatedness within the whole tree (*i.e.*, the complete evolutionary history of the connexin gene family).

Because the methods for phylogenetic inferences have different strengths and weaknesses with regard to the degree of relatedness of the sequences, the differences in evolutionary rates in different branches, how highly divergent sequences are behaving, etc., we used several methods for phylogenetic inferences, such as distance methods (Neighbor Joining and Minimum Evolution), Maximum Likelihood and Maximum Parsimony. All these methods are included in the MEGA phylogenetic software. If all, or most, of the statistical comparisons supported a specific dichotomous relationship, we deemed the results more trustable. Each method was used both at amino acid and nucleotide levels (the latter using only positions 1 and 2 in the codon) and with different substitution models, and in many cases with both bootstrap and interior branch statistics. In total, 21 statistical analyses were performed, and they are summarized in Suppl. Table 1, with the corresponding parameter settings in Suppl. Table 2.

3.2.6. Declarations

Ethics approval and consent to participate

All sequences were obtained from pre-existing databases and genome assemblies. No animals, or samples from animals, were handled in the context of the present work, and no approvals were required for this work.

Availability of data and materials

The datasets supporting the conclusions of this article are included within the article and its additional files. All sequences, with all required information, are in the Additional File 1. Please note that the data have been handled manually, and human error and inconsistencies could have occurred. If significant errors are detected, we would be grateful to receive a notification.

Competing interests

The authors declare that they have no competing interests.

Funding

SíK was supported by grants from the Faroese Research Council, the Fisheries Research Fund of the Faroe Islands, Statoil Føroyar, the Faroese Pelagic Organisation and the Danish Innovation Fund.

Authors' contributions

SOM: Conceived and designed the study, collected database sequences, analyzed data, and wrote the manuscript. MT: Contributed to the design of the study, collected database sequences, analyzed data, commented on the manuscript, and approved the final manuscript. SíK: Contributed to the design of the study, sequenced and assembled the herring genome, analyzed the herring sequences and performed comparative genomics against the other published herring genome assemblies, commented on the manuscript, and approved the final manuscript.

Additional files

Additional file 1.

File format: pdf. Contains Supplementary Figs. 1 – 17.

<u>Description of data:</u>

Suppl. Fig. 1. Human (*Homo sapiens*) connexins.

Suppl. Fig. 2. Mouse (*Mus musculus*) connexins.

Suppl. Fig. 3. Opossum (*Monodelphis domestica*) connexins.

Suppl. Fig. 4. *GJC1like* and *GJA9* connexin sequences from other marsupials and platypus.

Suppl. Fig. 5. Zebrafish (*Danio rerio*) connexins.

Suppl. Fig. 6. Japanese pufferfish (Fugu; *Takifugu rubripes*) connexins.

Suppl. Fig. 7. Green spotted pufferfish (*Tetraodon nigroviridis*) connexins.

Suppl. Fig. 8. Three-spined stickleback (*Gasterosteus aculeatus*) connexins.

Suppl. Fig. 9. Atlantic herring (*Clupea harengus*) connexins.

Suppl. Fig. 10. Atlantic cod (*Gadus morhua*) connexins.

Suppl. Fig. 11. Japanese eel (*Anguilla japonica*) connexins.

Suppl. Fig. 12. *Connexin39.2* ("*gjd2like*") from mammals.

Suppl. Fig. 13. Comparisons of human "*GJA4P*" against *connexin39.2* and *GJA4*. **A.** Alignment of conserved domains in human "*GJA4P*" (NG_026166) against *connexin39.2* ("*gjd2like*") in various species at protein level. **B.** Alignment of conserved domains in human "GJA4P" (NG_026166) against GJA4 (connexin37) from human and eel at protein level.

Suppl. Fig. 14. Expanded branches from the phylogenetic tree shown in Fig. 1. **A.** Expanded view of the mammalian and teleost *GJA1* branch. **B.** Expanded view of mammalian and teleost *GJA3* branch, and the associated teleost *cx39.9*. **C.** Expanded view of the mammalian and teleost *GJA4* branch. **D.** Expanded view of the mammalian and teleost *GJA5* branch. **E.** Expanded view of the mammalian and teleost *GJA9* and *GJA10* branches. **F.** Expanded view of the teleost *cx34.5* and *cx32.2* branches. **G.** Expanded view of the mammalian and teleost *GJB1* branch. **H.** Expanded view of mammalian and teleost *GJB2* and *GJB6* branch, and teleost

cx30.3 branches. **I.** Expanded view of the mammalian *GJB3* and teleost *cx35.4* branches. **J.** Expanded view of the mammalian and teleost *GJB7* branch. **L.**
Expanded view of the teleost *cx28.6* group, and its relationship with *GJB3*/*GJB4*/*GJB5*. **M.** Expanded view of eutherian *GJC3* and marsupial *GJC1like* and *GJC2like* branches. **N.** Expanded view of mammalian and teleost *GJC1* and teleost *cx43.4* branches. **O.** Expanded view of mammalian and teleost *GJC2*, and its relationship with *GJC1* and *cx43.4*. **P.** Expanded view of mammalian and teleost *Cx39.2* branch. **Q.** Expanded view over the central *GJD2* complex. **R.** Expanded view of mammalian and teleost *GJD3* branch. **S.** Expanded view of mammalian and teleost *GJD4* branch. **T.** Expanded view of teleost *cx36.7* branch.

Suppl. Fig. 15. Compressed phylogenetic tree illustrating long-branch attraction between *gjc3*, *gjd4* and *gje1* groups.

Suppl. Fig. 16. Searching for positions of connexins lacking in chromosome assemblies. **A.** Problem in cod assembly of chromosome 20 at assumed position of *gja5*. **B.** Alignments with sequences from herring and sticleback point to the same area on cod chromosome 21, indicated expected position of *gja10-cx52.6*. **C.** Alignments of herring and stickleback scaffolds containing *cx52.6*.

Suppl. Fig. 17. A homogeneous and consistent nomenclature for gap junction protein genes.

Additional file 2

File format: pdf. Contains Supplementary Tables 1 – 9.

Description of data:

Suppl. Table 1. Statistical support for clade grouping.

Suppl. Table 2. Parameter overview for statistical analyses of phylogenetic trees.

Suppl. Table 3. Comparison between zebrafish connexin sequences from Ensembl and GenBank.

Suppl. Table 4. Comparison between Fugu connexin sequences from Ensembl and GenBank

Suppl. Table 5. Comparison between cod connexin sequences from the Ensembl and GenBank assemblies.

Suppl. Table 6. Naming of connexin genes in Ensembl and GenBank.

Suppl. Table 7. Percentages of amino acid identities between conserved domains in mammalian Cx39.2, including human "GJA4P"-NG_026166, and eel cx39.2 (one of the "gjd2like" sequences.

Suppl. Table 8. Human *GJA4P* is more similar to *GJD2like* (*connexin39.2*) than GJA4 at nucleotide level.

Suppl. Table 9. Ohnology among teleost connexins.

3.2.7. References

1. Gaasterland T, Oprea M. Whole-genome analysis: annotations and updates. Curr Opin Struct Biol. 2001; 11:377-381.
2. Stein L. Genome annotation: from sequence to biology. Nat Rev Genet. 2001; 2:493-503.
3. Rouze P, Pavy N, Rombauts S. Genome annotation: which tools do we have for it? Curr Opin Plant Biol. 1999; 2:90-95.
4. Human Gene Name Committee [https://www.genenames.org/about/guidelines/]. Last accessed: 11. Feb. 2019.
5. Zebrafish Information Network [http://zfin.org/]. Last accessed: 25. March 2019
6. Fitch WM. Distinguishing homologous from analogous proteins. Syst Zool. 1970; 19:99-113.
7. Fitch WM. Homology. A personal view on some of the problems. Trends Genet. 2000; 16:227-231.
8. Hartl DL, Cochrane BJ: Genetics. Analysis of Genes and Genomes, 9th Ed. Burlington, MA: Jones and Bartlett Learning; 2018.
9. Putnam NH, Butts T, Ferrier DE, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu JK *et al*. The amphioxus genome and the evolution of the chordate karyotype. Nature. 2008; 453:1064-1071.
10. Smith JJ, Kuraku S, Holt C, Sauka-Spengler T, Jiang N, Campbell MS, Yandell MD, Manousaki T, Meyer A, Bloom OE *et al*. Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution. Nat Genet. 2013; 45:415-421, 421e411-412.
11. Ohno S: Evolution by Gene Duplication. Berlin: Springer Verlag; 1970.
12. Taylor JS, Braasch I, Frickey T, Meyer A, Van de Peer Y. Genome duplication, a trait shared by 22000 species of ray-finned fish. Genome Res. 2003; 13:382-390.
13. Meyer A, Van de Peer Y. From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). Bioessays. 2005; 27:937-945.
14. Vandepoele K, De Vos W, Taylor JS, Meyer A, Van de Peer Y. Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. Proc Natl Acad Sci U S A. 2004; 101:1638-1643.
15. Wolfe K. Robustness--it's not where you think it is. Nat Genet. 2000; 25:3-4.
16. Postlethwait JH. The zebrafish genome in context: ohnologs gone missing. J Exp Zool B Mol Dev Evol. 2007; 308:563-577.
17. Martinez Barrio A, Lamichhaney S, Fan G, Rafati N, Pettersson M, Zhang H, Dainat J, Ekman D, Hoppner M, Jern P *et al*. The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. eLife. 2016; 5:e12081.

18. Cai H, Li Q, Fang X, Li J, Curtis NE, Altenburger A, Shibata T, Feng M, Maeda T, Schwartz JA *et al*. A draft genome assembly of the solar-powered sea slug *Elysia chlorotica*. Sci Data. 2019; 6:190022.

19. Vezzi F, Narzisi G, Mishra B. Reevaluating assembly evaluations with feature response curves: GAGE and assemblathons. PLoS One. 2012; 7:e52210.

20. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013; 29:1072-1075.

21. Amores A, Force A, Yan YL, Joly L, Amemiya C, Fritz A, Ho RK, Langeland J, Prince V, Wang YL *et al*. Zebrafish *hox* clusters and vertebrate genome evolution. Science. 1998; 282:1711-1714.

22. Eastman SD, Chen TH, Falk MM, Mendelson TC, Iovine MK. Phylogenetic analysis of three complete gap junction gene families reveals lineage-specific duplications and highly supported gene classes. Genomics. 2006; 87:265-274.

23. Cruciani V, Mikalsen SO. The vertebrate connexin family. Cell Mol Life Sci. 2006; 63:1125-1140.

24. Cruciani V, Mikalsen SO. Evolutionary selection pressure and family relationships among connexin genes. Biol Chem. 2007; 388:253-264.

25. Glasauer SMK, Neuhauss SCF. Whole-genome duplication in teleost fishes and its evolutionary consequences. Mol Genet Genomics. 2014; 289:1045-1060.

26. Sasakura Y, Shoguchi E, Takatori N, Wada S, Meinertzhagen IA, Satou Y, Satoh N. A genomewide survey of developmentally relevant genes in *Ciona intestinalis*. X. Genes for cell junctions and extracellular matrix. Dev Genes Evol. 2003; 213:303-313.

27. Bennett MV, Zheng X, Sogin ML. The connexins and their family tree. Soc Gen Physiol Ser. 1994; 49:223-233.

28. Cruciani V, Mikalsen SO. The connexin gene family in mammals. Biol Chem. 2005; 386:325-332.

29. Harris AL. Emerging issues of connexin channels: biophysics fills the gap. Q Rev Biophys. 2001; 34:325-472.

30. Procida K, Jorgensen L, Schmitt N, Delmar M, Taffet SM, Holstein-Rathlou NH, Nielsen MS, Braunstein TH. Phosphorylation of connexin43 on serine 306 regulates electrical coupling. Heart Rhythm. 2009; 6:1632-1638.

31. Kidder GM, Winterhager E. Physiological roles of connexins in labour and lactation. Reproduction. 2015; 150:R129-136.

32. Mesnil M. Connexins and cancer. Biol Cell. 2002; 94:493-500.

33. Genet N, Bhatt N, Bourdieu A, Hirschi KK. Multifaceted roles of connexin 43 in stem cell niches. Curr Stem Cell Rep. 2018; 4:1-12.

34. Sorgen PL, Trease AJ, Spagnol G, Delmar M, Nielsen MS. Protein-protein interactions with connexin 43: regulation and function. Int J Mol Sci. 2018; 19:1428.

35. Sundset R, Ytrehus K, Mikalsen SO. Connexin, connection, conductance: Towards understanding induction of arrhythmias? Heart Rhythm. 2009; 6:1639-1640.

36. Bennett MV, Contreras JE, Bukauskas FF, Saez JC. New roles for astrocytes: gap junction hemichannels have something to communicate. Trends Neurosci. 2003; 26:610-617.

37. Wang N, De Bock M, Decrock E, Bol M, Gadicherla A, Vinken M, Rogiers V, Bukauskas FF, Bultynck G, Leybaert L. Paracrine signaling through plasma membrane hemichannels. Biochim Biophys Acta. 2013; 1828:35-50.

38. Orellana JA, Saez JC, Bennett MV, Berman JW, Morgello S, Eugenin EA. HIV increases the release of dickkopf-1 protein from human astrocytes by a Cx43 hemichannel-dependent mechanism. J Neurochem. 2014; 128:752-763.

39. 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; 477:207-210.

40. Henkel CV, Dirks RP, de Wijze DL, Minegishi Y, Aoyama J, Jansen HJ, Turner B, Knudsen B, Bundgaard M, Hvam KL *et al*. First draft genome sequence of the Japanese eel, *Anguilla japonica*. Gene. 2012; 511:195-201.

41. Jansen HJ, Liem M, Jong-Raadsen SA, Dufour S, Weltzien FA, Swinkels W, Koelewijn A, Palstra AP, Pelster B, Spaink HP *et al*. Rapid *de novo* assembly of the European eel genome from nanopore sequencing reads. Sci Rep. 2017; 7:7213.

42. Igarashi Y, Zhang H, Tan E, Sekino M, Yoshitake K, Kinoshita S, Mitsuyama S, Yoshinaga T, Chow S, Kurogi H *et al*. Whole-genome sequencing of 84 Japanese eels reveals evidence against panmixia and support for sympatric speciation. Genes (Basel). 2018; 9:474.

43. Near TJ, Eytan RI, Dornburg A, Kuhn KL, Moore JA, Davis MP, Wainwright PC, Friedman M, Smith WL. Resolution of ray-finned fish phylogeny and timing of diversification. Proc Natl Acad Sci U S A. 2012; 109:13698-13703.

44. Betancur RR, Broughton RE, Wiley EO, Carpenter K, Lopez JA, Li C, Holcroft NI, Arcila D, Sanciangco M, Cureton Ii JC *et al*. The tree of life and a new classification of bony fishes. PLoS Curr. 2013; 5:10.1371/currents.tol.1353ba26640df26640ccaee26675bb26165c26648c26288.

45. Baldauf SL. Phylogeny for the faint of heart: a tutorial. Trends Genet. 2003; 19:345-351.

46. Bergsten J. A review of long-branch attraction. Cladistics. 2005; 21:163-193.

47. Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, Collins JE, Humphray S, McLaren K, Matthews L *et al*. The zebrafish reference genome sequence and its relationship to the human genome. Nature. 2013; 496:498-503.

48. Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A *et al*. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science. 2002; 297:1301-1310.

49. Ensembl: Fugu (*Takifugu rubripes*) genome database [http://www.ensembl.org/Takifugu_rubripes/Info/Annotation]. Last accessed: 15. March 2019.

50. *Gadus morhua* (Atlantic cod) chromosome level assembly, GFC_902167405 [https://www.ncbi.nlm.nih.gov/assembly/GCF_902167405.1/]. Last accessed: 20. Sept. 2019.

51. O'Brien J, al-Ubaidi MR, Ripps H. Connexin 35: a gap-junctional protein expressed preferentially in the skate retina. Mol Biol Cell. 1996; 7:233-243.

52. *Clupea harengus* (Atlantic herring) chromosome level assembly GCA_900700415 [https://www.ncbi.nlm.nih.gov/assembly/GCA_900700415.1]. Last accessed: 20 sept 2019.

53. *Clupea harengus* (Atlantic herring) genome assembly GCA_000966335.1 [https://www.ncbi.nlm.nih.gov/assembly/GCF_000966335.1]. Last accessed: 20. Sept 2019.

54. 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 - detection of a supergene and other signals of selection. https://wwwbiorxivorg/content/101101/668384v1. 2019.

55. *Clupea harengus* (Atlantic herring) genome assembly GCA_900323705 [https://www.ncbi.nlm.nih.gov/assembly/GCA_900323705.1]. Last accessed: 20 Sept. 2019.

56. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007; 23:1061-1067.

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

58. Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol. 2017; 35:543-548.

59. Ensembl: Three-spined stickleback (*Gasterosteus aculeatus*) genome database [https://www.ensembl.org/Gasterosteus_aculeatus/Info/Index]. Last accessed: 1st Sept 2019.

60. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol. 2007; 8:R143.

61. Ensembl: Tetraodon (*Tetradon nigroviridis*) genome database [https://www.ensembl.org/Tetraodon_nigroviridis/Info/Index]. Last accessed: 1st Sept 2019.

62. Ensembl: Atlantic cod (*Gadus morhua*) genome database [https://www.ensembl.org/Gadus_morhua/Info/Index]. Last accessed: 1st Sept. 2019.

63. Nomura K, Fujiwara A, Iwasaki Y, Nishiki I, Matsuura A, Ozaki A, Sudo R, Tanaka H. Genetic parameters and quantitative trait loci analysis associated with body size and timing at metamorphosis into glass eels in captive-bred Japanese eels (*Anguilla japonica*). PLoS One. 2018; 13:e0201784.

64. *Anguilla japonica* (Japanese eel) genome assembly GCA_0035977225 [https://www.ncbi.nlm.nih.gov/assembly/GCA_003597225.1]. Last accessed: 24th Oct, 2019.

65. Nakamura Y, Yasuike M, Mekuchi M, Iwasaki Y, Ojima N, Fujiwara A, Chow S, Saitoh K. Rhodopsin gene copies in Japanese eel originated in a teleost-specific genome duplication. Zoological Lett. 2017; 3:18.

66. Pavey SA, Laporte M, Normandeau E, Gaudin J, Letourneau L, Boisvert S, Corbeil J, Audet C, Bernatchez L. Draft genome of the American eel (*Anguilla rostrata*). Mol Ecol Resour. 2017; 17:806-811.

67. Bracamonte SE. *Anguilla anguilla* spleen and head kidney transcriptome [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA419718]. Last accessed: 25. May 2019.

68. Pasquier J, Cabau C, Nguyen T, Jouanno E, Severac D, Braasch I, Journot L, Pontarotti P, Klopp C, Postlethwait JH *et al*. Gene evolution and gene expression after whole genome duplication in fish: the PhyloFish database. BMC Genomics. 2016; 17:368.

69. Perrier F, Bertucci A, Pierron F, Feurtet-Mazel A, Simon O, Klopp C, Candaudap F, Pokrovsky O, Etcheverria B, Mornet S *et al*. Transcriptomic profiling responses in liver and brain tissues of European eel *Anguilla anguilla* after a gold nanoparticle trophic exposure. [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA432560]. Last accessed: 25. May 2019.

70. Tse WK, Sun J, Zhang H, Law AY, Yeung BH, Chow SC, Qiu JW, Wong CK. Transcriptomic and iTRAQ proteomic approaches reveal novel short-term hyperosmotic stress responsive proteins in the gill of the Japanese eel (*Anguilla japonica*). J Proteomics. 2013; 89:81-94.

71. Kumar S, Stecher G, Tamura K. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol. 2016; 33:1870-1874.

72. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol. 2018; 35:1547-1549.

3.3. Identification of male heterogametic sex determining regions on the Atlantic herring *Clupea harengus* genome

Authors

í Kongsstovu, Sunnvør[1,2,4]; Dahl, Hans Atli[1]; Gislason, Hannes[2]; í Homrum, Eydna[3]; Jacobsen, Jan Arge[3]; Flicek, Paul[4]; Mikalsen, Svein-Ole[2]

Affiliations

[1] Amplexa Genetics A/S, Hoyvíksvegur 51, FO-100 Tórshavn, Faroe Islands.

[2] University of the Faroe Islands, Dept. of Science and Technology, Vestara Bryggja 15, FO-100 Tórshavn, Faroe Islands.

[3] Faroe Marine Research Institute, Nóatún 1, FO-100 Tórshavn, Faroe Islands.

[4] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.

Corresponding author: Sunnvør í Kongsstovu; Hoyvíksvegur 51, FO-100 Tórshavn, Faroe Islands; skik@amplexa.com

Funding Information: The study was funded by the Faroese Research Council, the Fisheries Research Fund of the Faroe Islands, Innovationsfonden (Denmark), Statoil Føroyar, the Faroese Pelagic Organisation and the European Molecular Biology Laboratory.

3.3.1. Abstract

The sex determination system of the commercially important fish, Atlantic herring *Clupea harengus* L. was investigated. Low coverage whole genome sequencing in 48 females and 55 males and a genome wide association study revealed six short genomic regions to be associated with sex. Two scaffolds with two regions showed higher level of significance than the other regions. The genotyping data of the SNPs associated with sex showed that 98.9% of the available female genotypes were homozygous for the reference alleles, while 70.4% of the available male genotypes were heterozygous. This is close to the theoretical expectation of homo/heterozygous distribution at low sequencing coverage when the males are factually heterozygous. This suggests a male heterogametic sex determination system in *C. harengus,* consistent with other species within the Clupeiformes group. The results may also suggest that *C. harengus* sex determination could be polygenic. There were 20 protein coding genes on the significant regions but none of these genes were previously reported master sex regulation genes, or obviously related to sex determination. However, many of these genes are expressed in testis or ovary in other species, but the exact genes controlling sex determination in *C. harengus* could not be identified.

Key Words

Atlantic herring, *Clupea harengus*, GWAS, sex determination.

3.3.2. Introduction

The evolution of sexual reproduction has resulted in several sex determination systems, with both gonochorous organisms (the stable separation of sexes in different individuals), stable hermaphrodites, and organisms that change sex dependent on age, environmental and/or social cues (Devlin and Nagahama, 2002; Shen and Wang, 2014). Each of the different systems has evolved independently several times through evolutionary history, and even within each system there might be several different mechanisms for determining the sex of an organism (Ashman *et al.*, 2014). The best-known system is the XY sex chromosomes found in mammals, where the females have two X chromosomes while the males have an X and a Y chromosome. Thus, the XY system is a male heterogametic system. The sex-determining region Y (*SRY)* gene is located on the Y chromosome and signals to the body to develop into a male rather than a female, which is the default (Kashimada and Koopman, 2010). The ZW system is a similar system where the females are heterogametic. This system is found in birds and some amphibians (Bull, 1983; Yoshimoto and Ito, 2011). These two systems are simple but do not represent the complexity of sex determination systems in the animal kingdom. Systems with only one sex chromosome also exist; for example, the X0 system where males have only one sex chromosome, and the Z0 system where females have only one sex chromosome (Clinton, 1998; Bachtrog *et al.*, 2014). Sex determination systems can also be more complex with multiple chromosomes or genes affecting the sex (Bachtrog *et al.*, 2014; Roberts *et al.*, 2016). There are even systems where age and size (Allsop and West, 2003), societal factors (Fricke, 1979; Buston, 2003), or environmental factors such as temperature (Pieau, 1996) play important roles in sex determination. In some organisms, both genetic and environmental factors are involved in determining the sex, for example in the Nile tilapia *Oreochromis niloticus* (Linnaeus 1758) (Baroiller *et al.*, 2009) and Atlantic silverside *Menidia menidia* (Linnaeus 1766) (Lagomarsino and Conover, 1993).

Teleost fish display a variety of sex determination systems (reviewed in (Devlin and Nagahama, 2002; Brykov, 2014)) and the plasticity of the teleost genomes makes it possible for new systems to evolve relatively quickly. This makes teleost fish good candidates for studying the evolution of sex determination. Although there are sex determination systems in fish that are influenced by non-genetic factors (see references above), genetic sex determination seems to be more common. The male heterogametic system (from now on referred to as the XY system) has been established in fish, for example bighead carp *Hypophthalmichthys nobilis* (Richardson 1845) and silver carp *Hypophthalmichthys molitrix* (Valenciennes 1844) (Liu *et al.*, 2018), as has the female heterogametic system (from now on referred to as the ZW system) in half-smooth tongue sole *Cynoglossus semilaevis* Günther 1873 (Chen *et al.*, 2014). The cichlid fishes of Lake Malawi have families with the XY system and others with the ZW system, but interestingly the species *Metriaclima pyrsonotus* (Stauffer, Bowers, Kellogg & McKaye 1997) has both these systems that show strong epistatic interactions between them (Ser *et al.*, 2010). There are also several polygenic systems found in fish, for example in the European sea bass *Dicentrarchus labrax* (Linnaeus 1758) (Palaiokostas *et al.*, 2015) and the cichlid fish *Astatotilapia burtoni* (Günther 1894) (Roberts *et al.*, 2016). There are even individuals from the same species that have different sex determination systems, e.g. some zebrafish *Danio rerio* (Hamilton 1822) laboratory strains have lost the sex determining region that is present in wild type *D. rerio* and has therefore evolved a new polygenic system that is still not fully understood (Wilson *et al.*, 2014).

In some organisms (e.g. mammals and birds) sex chromosomes have evolved that contain master sex regulation (MSR) genes that control the sex of the organism, like the previously mentioned *SRY*. Most species of fish do not have specific heteromorphic chromosomes that control sex but have regions on autosomes that are associated with sex determination. These sex regions sometimes contain MSR genes or candidate MSR genes, like the Y chromosome-specific anti-Müllerian hormone (*amhy)* gene in Patagonian pejerrey *Odontesthes hatcheri* (Eigenmann 1909) (Hattori *et al.*, 2012) or the sexually dimorphic gene on the Y chromosome (*sdy)* in rainbow trout *Oncorhynchus mykiss* (Walbaum 1792) (Yano *et al.*, 2012). However, sometimes there are no obvious causal genes found on these regions that have been associated with sex e.g. in the mandarin fish *Siniperca chuatsi* (Basilewsky 1855) (Sun *et al.*, 2017).

In the Clupeidae family, few species have been studied regarding their sex determination systems. In the Tree of Sex Consortium database (Ashman *et al.*, 2014) only six Clupeiformes species are mentioned; four of these are a part of the Clupeidae family. Two of them are hemaphrodites (the toli shad *Tenualosa toli* (Valenciennes 1847) and the longtail shad *T. macrura* (Bleeker 1852)) while the Argentine menhaden *Brevoortia pectinata* (Jenyns 1842) and the Brazilian menhaden *B. aurea* (Spix & Agassiz 1829) are both gonochoristic. *B. pectinata* is homomorphic and *B. aurea* is male heterogametic with X_1X_2Y sex chromosomes (Brum, 1992). In addition, the Gulf menhaden *Brevoortia patronus* Goode 1878, the yellowfin menhaden *Brevoortia smithi* Hildebrand 1941 and the Atlantic menhaden *Brevoortia tyrannus* (Latrobe 1802) are also gonochoristic and homomorphic (Doucette Jr and Fitzsimons, 1988) but their sex determination systems are not known.

The sex determination system of the commercially important Atlantic herring *Clupea harengus* L. has not yet been described. Increasing the knowledge of sex determination at this branch of the tree of life would shed more light upon the evolution of sex determination in teleost fish. We therefore undertook this study to find regions on the *C. harengus* genome that are associated with sex determination.

3.3.3. Materials and Methods

Ethics

The *C. harengus* samples were received from stock assessment cruises and from commercial catches. No fish were caught, or handled while alive, for the purpose of this project. All fish were dead when they were selected for this project. Concerning the specific ethics questions listed in Instructions to Authors by Journal of Fish Biology, the status for the present project is as follows:

Were fishes collected as part of faunal surveys?

> No. We do not consider "faunal surveys" equivalent with stock assessment cruises. Stock assessment cruises are needed for setting quotas and make national, bilateral and international agreements for the commercial exploitation of fish stocks. Around half of the fish were collected on stock assessment cruises and the rest from commercial catches.

Were fishes killed during or at the end of your experiment (*e.g.,* for tissue sampling)?

> No. Fish were already dead when they were enrolled in this project.

Were surgical procedures performed?

> No.

Did the experimental conditions severely distress any fishes involved in your experiments?

> No.

Did any procedures (*e.g.,* predation studies, toxicity testing) cause lasting harm to sentient fishes?

> No.

Did any procedure involve sentient, un-anaesthetised animals that were subjected chemical agents that induce neuromuscular blockade, such as muscle relaxants?

> No.

Samples and DNA extraction

Kidney samples were taken from the 103 fish, obtained from stock assessment surveys and from commercial catches, and the sex was determined by visual inspection of the gonads. Revealing 48 females and 55 males. DNA was extracted from the kidney tissue from these fish, using an AS1000 Maxwell 16 instrument (Promega, Wisconsin, United States) and the

Maxwell 16 Tissue DNA purification kit (Promega). DNA concentration was measured using a Qubit 3.0 fluorometer (ThermoFisher Scientific).

Sequencing

The isolated DNA from each individual was fragmented to roughly 300 bp using a Covaris M220 (Covaris, Chicago, United States) and the libraries prepared using the KAPA TLP Library Preparation Kit Illumina platforms (KAPABiosystems, Massachusetts, United States) and quantified using the KAPA Library Quantification Kit (KAPABiosystems), following the manufacturer's instructions. After quantification the libraries were pooled to equal proportion and sequenced on a NextSeq500 (Illumina, California, United States) using the High Output v2 kit (Illumina).

Data processing and variant calling

The sequencing data were trimmed using Trimmomatic v0.36 to remove adapters and low-quality bases (Q score < 20) (Bolger *et al.*, 2014). AfterQC v0.4.0 was used to remove the polyG reads (Sun *et al.*, 2017). FastQC v0.11.5 was used to assess the quality of all the sequencing data (Andrews, 2010). The data were then aligned to the draft *C. harengus* genome (GCF_000966335.1_ASM96633v1) using BWA-MEM v0.7.15 (Li, 2013) and SAMtools v1.3 was used for sorting, converting (between SAM and BAM file format) and removing PCR duplicates from the alignment files (Li *et al.*, 2009). Single nucleotide polymorphisms (SNPs) were called using FreeBayes v1.1.0 (Garrison and Marth, 2012). Low quality SNPs were filtered out. These include SNPs not in Hardy-Weinberg equilibrium, SNPs with a quality score lower than 20 (QUAL < 20), SNPs where fewer than 30 samples had data (NS < 30) and SNPs with coverage lower than 2 (DP < 2).

Association analysis

A genome wide association study (GWAS) was performed, using Plink v1.07 (Purcell *et al.*, 2007), to test if any of the SNPs identified were associated with sex. The Bonferroni correction for multiple testing assumes that each test is independent. This assumption is not always true for a GWAS, because of linked SNPs. Thus, the Bonferroni correction can be considered too conservative. For human GWAS the significance threshold is commonly accepted at -

Log$_{10}$(p)≤7.3 for common variants (Fadista *et al.*, 2016). However, no studies have been published yet that show that this value is also appropriate for herring. Therefore, we choose to use the conservative Bonferroni correction and accept significance at -Log$_{10}$(p)≤8.7. The R packages qqman (Turner, 2014) and ABHgenotypeR (Furuta *et al.*, 2017) were used for visualization of the results.

Statistical analysis

The experimental genotyping data versus coverage was compared to theoretical models of only homozygous genotypes and only heterozygous genotypes. In the model for only homozygous genotypes, we assumed that the probability of having x identical reads given homozygous genotypes was $P(x|hom) = 1$. In the model for only heterozygous genotypes, we assumed that the probability of having x identical reads given heterozygous genotypes was $P(x|het) = 1/2^{x-1}$ (Chenuil, 2012). To test if the observed proportions of homozygotes at each coverage step were significantly different from the theoretically expected proportions, exact binominal tests were carried out. No tests were carried out for coverage higher than 21, because of low number of samples with such high coverage.

To test if there was a significant difference between the proportion of high coverage (>5x) homozygous genotypes in males at the different sex regions a Chi-squared test was used. Bonferroni corrections were made to correct for multiple testing (0.05/6) and significance accepted at 0.008.

Search for causal genes

Possible orthologs for the genes on the significant regions were found via OrthoDB (Kriventseva *et al.*, 2019). If nothing was found, a blast search of the gene sequence was performed to identify possible orthologs. Function of the orthologs were investigated in the literature and in the UniProt database. The mRNAs of the genes or orthologs were investigated to see if the SNPs and indels identified in the study were located in exons or introns.

The sequences of known sex determination or differentiation genes in fish were aligned against the *C. harengus* genome to investigate if any of these genes were present but not predicted for the *C. harengus* genome.

3.3.4. Results

Identification of sex regions on the C. harengus genome

SNPs were found via low-coverage whole genome sequencing, and a genome wide association study (GWAS) was carried out to identify the regions on the genome associated with sex, similar to Purcell *et al.* (2018). Whole genome sequencing of 103 *C. harengus* (48 females and 55 males) resulted in 386x coverage of the *C. harengus* genome (176x coverage of the female genome and 210x coverage of the male genome) and after SNP calling and filtering 50.089.222 SNPs were identified. To find genomic regions associated with sex, a GWAS was performed and resulted in 604 SNPs significantly associated with sex. Potentially spurious findings were filtered out based on their relatively poor p-values and no other significant p-values in close proximity (Reed *et al.*, 2015). The remaining 584 associated SNPs aggregated in 6 regions (Table 1 and Figure 1a) and are listed in Supplement file 1. Sex Region (SR)1.1 and SR1.2, and SR4.1 and SR4.2 are closely located and are only visible as separate peaks when zoomed in on their respective scaffolds (Figure 1b and 1c).

For the significant SNPs, 98.9% of the available female genotypes were homozygous for the reference alleles, while 70.4% of the available male genotypes were heterozygous (Figure 2 and Table 2). A closer look at the SNP on SR1.1 with the most significant p-value (NW_012219506.1:975264, p = 2.486e-17), showed the general genotype trend. When genotypes with coverage 1 were included, 32 of the 52 males had heterozygote genotypes for the mentioned SNP, with 5 males homozygous for the reference allele and 15 for the alternative allele. All 43 females were homozygous for the reference allele. After removal of genotypes with coverage of 1, 41 females had genotyping data and all of them were homozygous for the reference allele. Forty-five males had genotyping data for this SNP, 2 were homozygous for the reference allele, 11 were homozygous for the alternative allele, while 32 were heterozygous. All the significant SNPs showed a similar pattern (Table 2). In addition, the average coverage ± standard deviation of the male SNPs that were heterozygous was higher than, the coverage for the homozygous reference and alternative allele (Table 3). These results suggest a XY sex determination system for *C. harengus*.

Table 1. Regions of the Atlantic herring *Clupea harengus* genome associated with sex, identified in a GWAS.

Significant region	Scaffold	Position	No. Significant SNPs
SR1.1	NW_012219506.1	919996 - 1039292	514
SR1.2	NW_012219506.1	2305958 - 2321039	42
SR2	NW_012219703.1	28023 - 28094	3
SR3	NW_012221357.1	979071 - 979160	4
S4.1	NW_012223947.1	3178113 - 3182376	6
S4.2	NW_012223947.1	7965877 - 7968244	15

Table 2. Genotype count for the 584 SNPs associated with sex in Atlantic herring *Clupea harengus*. Genotypes with coverage higher than 1x are included.

Genotype	Females	Males	Total
Homozygous (reference + alternative)	20348 (20286 + 62)	7442 (3761 + 3681)	27790
Heterozygous	168	17699	17867
Total	20516	25141	45657

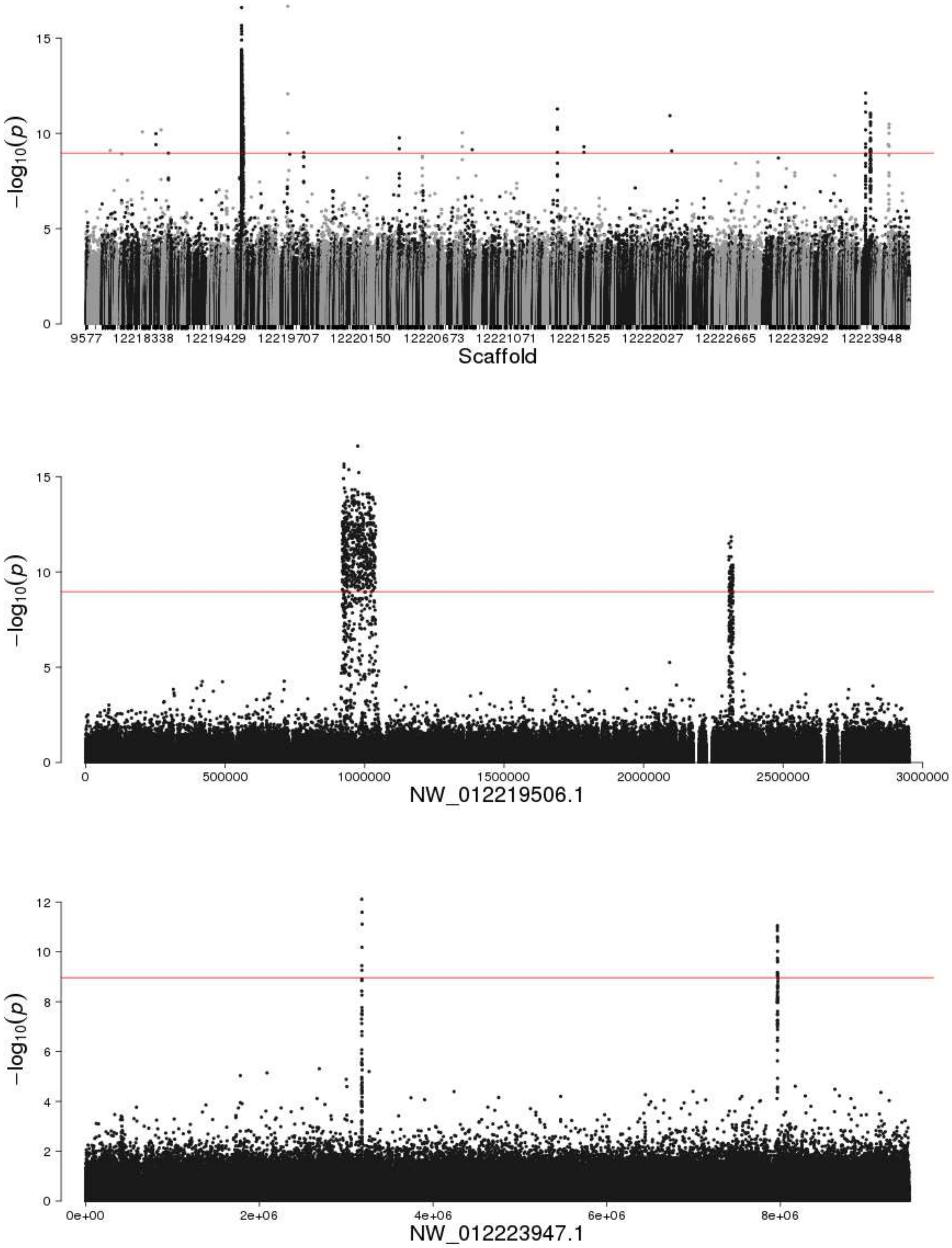


Figure 1. Manhattan plot showing -log of the p-values from the GWAS investigating sex determination regions on the Atlantic herring *Clupea harengus* genome. Red line indicates the genome wide significance ($-\text{Log}_{10}(p) \leq 8.7$). First panel (a) shows the p-values across the whole *C. harengus* genome, while the second (b) and third (c) panels show scaffolds NW_012219506.1 and NW_012223947.1, respectively.

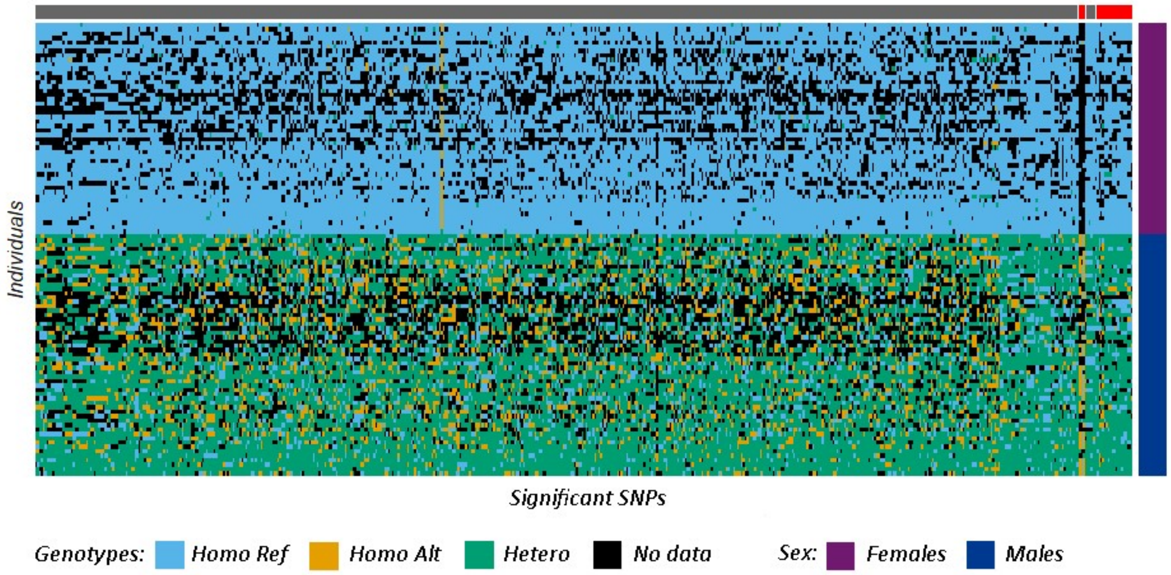


Figure 2. Genotypes for the SNPs significantly associated with sex in Atlantic herring *Clupea harengus*. The alternating grey and red line at the top shows SR1, SR2, SR3 and SR4, in this order from the left.

Table 3. Average coverage for the individual SNPs associated with sex in Atlantic herring *Clupea harengus*. S.D. = standard deviation, n = number of samples.

	Females			**Males**		
Genotype	**Average**	**S.D.**	**n**	**Average**	**S.D.**	**n**
Homozygous reference allele	4.98	3.41	20286	3.51	1.35	3761
Homozygous alternative allele	4.47	0.25	62	3.18	1.16	3681
Heterozygous	4.91	0.45	168	6.00	3.90	17699

The erroneous call at low sequence coverage of homozygotes from factual heterozygotes has been pointed out before (Chenuil, 2012). Thus, the true rate of heterozygotes in our data is higher than our result of 70.4% but this could not be detected due to low sequencing coverage, resulting in male genotypes possibly being wrongly called as homozygous. To investigate this further, the observed proportions of homozygous female and male genotypes versus coverage were compared to the corresponding theoretically expected probabilities $P(x|hom) = 1$ and $P(x|het) = 1/2^{x-1}$ (Chenuil, 2012) (Figure 3). Even though both the observed male and female proportions of homozygotes versus coverage were significantly different from the theoretically expected proportions (Table S1 in Supplementary File 2), the observed male proportions followed the same trend as the theoretically expected proportions for only heterozygotes, and the observed female proportions followed the same trend as the theoretically expected proportions for only homozygotes (Figure 3). This indicates that perhaps not all, but the majority, of the significant SNPs must be heterozygous for males to develop. Nevertheless, the results support the suggestion of a XY sex determination system for *C. harengus*.

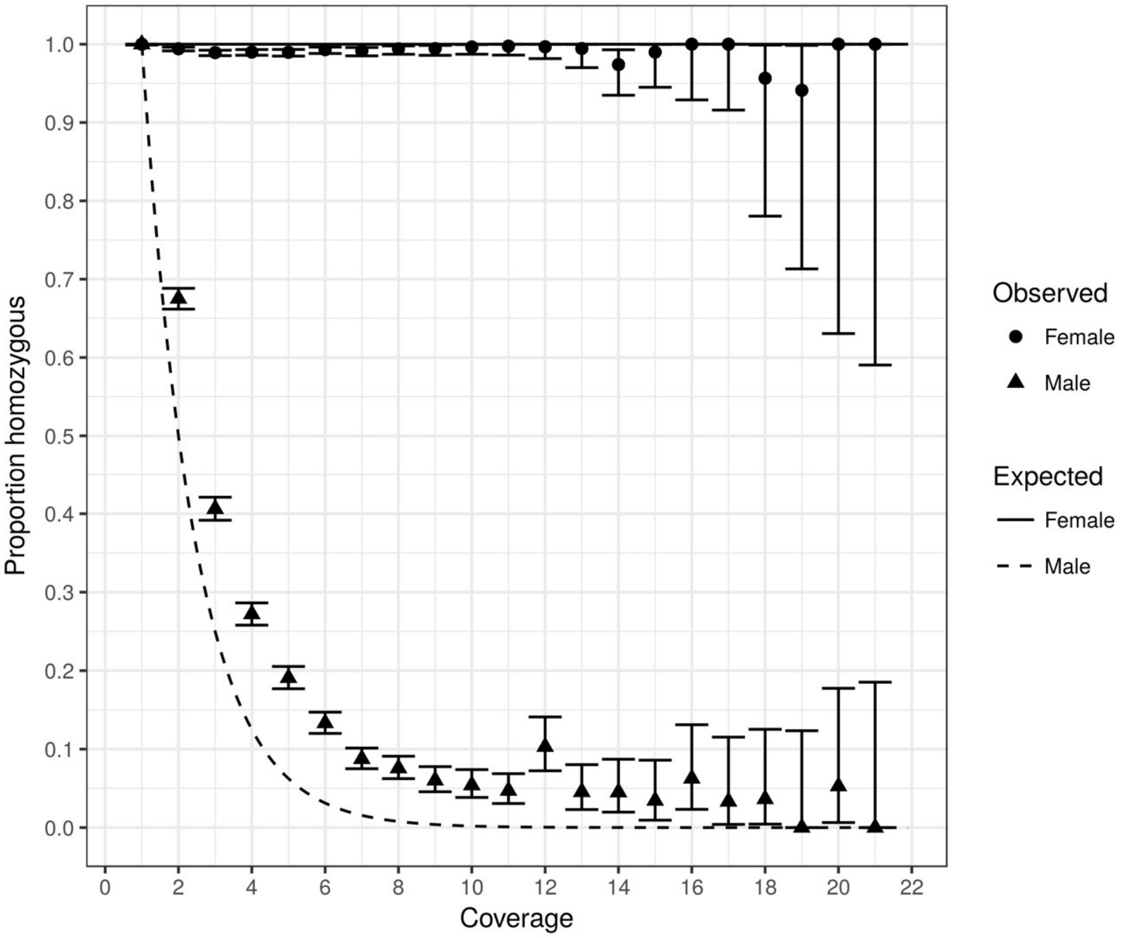


Figure 3. The experimentally observed proportions of homozygous female and male genotypes of the SNPs significantly associated with sex in Atlantic herring *Clupea harengus* versus coverage and the corresponding theoretically expected probabilities. Triangle (▲), male data; circles (●), female data; solid line, expected distribution if all genotypes were homozygous ($P(x|hom) = 1$); dashed line, expected distribution if all genotypes were heterozygous ($P(x|het) = 1/2^{x-1}$). Error bars correspond to 95% confidence intervals from the binomial test.

We see four possible explanations for the small, but significant, deviations from the theoretical expected proportions in Figure 3. (i) The physiological sex has been wrongly registered. We think this explanation is unlikely. Figure 2 would then have indicated this with horizontal lines of the deviating zygosity. (ii) Random variations due to a limited number of individuals tested. We cannot fully exclude this possibility, although we have investigated 55 males. (iii) Some of the SNPs are not important for the male sex determination, and do not have to be present; they are mere passenger variations. (iv) A small proportion of the males have an alternative sex determination mechanism. As an approach to investigate the third possibility, the proportions of homozygous calls with coverage higher than 5 in males were sorted according to the different regions (Table 4). The proportions of homozygotes were significantly lower in regions SR1.1 and SR4.2, than in SR1.2 and SR4.1 (Table 4), suggesting that SR1.2 and SR4.1 could be of less importance for sex determination.

Table 4. The proportions of homozygous calls for significant SNPs in Atlantic herring *Clupea harengus* males, sorted according to the different regions associated with sex. Genotypes with coverage lower than 5 have been filtered out. A Chi-squared test was used to test if the proportions were significantly different and the 95% confidence interval (95% CI) was calculated assuming normal distribution. Column 6 and 7 show the p-values when regions were compared to SR1.1 and SR4.2, respectively.

Sex region	Total genotypes with data	Homozygou s genotypes	Proportio n	95% CI	SR1.1 p-value	SR4.2 p-value
SR1.1	7220	578	0.0801	0.0063	-	0.0756
SR1.2	1157	129	0.1115	0.0181	0.0004	0.0011
SR2	11	11	1.0000	†	†	†
SR3	119	13	0.1092	0.0560	0.3218	0.0573
SR4.1	161	24	0.1491	0.0550	0.0025	0.0004
SR4.2	405	22	0.0543	0.0221	0.0756	-

† Sample size for SR2 is too low to assume normal distribution.

Search for possible sex determination genes

As several regions were associated with sex, more than one gene could be involved in sex determination. We investigated the protein-coding genes in these regions. Three of the regions (SR2, SR3 and SR4.1) did not contain any predicted protein-coding genes in *C. harengus*. The three remaining regions contained 20 genes. These genes are listed in Table 5. None of these genes have previously been shown to be MSR genes in other organisms or shown to be a part of the sex determination pathway. However, to investigate further, possible orthologs for these genes were found and their reported functions investigated. None of the 20 genes were obvious candidates for being MSR genes, but 6 of the genes showed some potential linkage with sex determination or sex related functions. These genes were the cation channel sperm associated 3 (*catsper3),* IQ motif containing D (*iqcd)*, protein KIAA2022-like *(loc105890446*), Merlin-like (*loc105890474*), ribose-phosphate pyrophosphokinase 1-like (*loc105890483*) and two pore segment channel 1 (*tpcn1)*. The human *Homo sapiens* orthologs of *catsper3* (*CATSPER3;* GeneID: 347732) and *iqcd* (*DRC1;* GeneID: 92749) are both expressed in the testis and *CATSPER3* plays a role in the fitness of sperm cells (Bastian *et al.*, 2008; Strünker *et al.*, 2011). *Catsper3* has also been shown to be important for fertilisation of *C. harengus* eggs (Yanagimachi *et al.*, 2017a). The *D. rerio* ortholog of *loc105890474*, neurofibromin 2b (*nf2b;* GeneID: 405887) has been shown to be highly expressed in testis, while the *D. rerio* ortholog of *tpcn1,* (*tpcn1;* GeneID: 567534) has highest expression levels in mature ovarian follicle and testis (Bastian *et al.*, 2008). The *H. sapiens* ortholog of *loc105890446*, neurite extension and migration factor (*NEXMIF;* GeneID: 340533) is X-linked and may therefore play a role in sex-related processes (Cantagrel *et al.*, 2004). The *H. sapiens* ortholog of *loc105890483*, phosphoribosyl pyrophosphate synthetase 1-like 1 *(PRPS1L1*; GeneID: 221823), is specifically expressed in the testis and encodes a protein that is highly homologous to the two subunits of phosphoribosylpyrophosphate synthetase encoded by the X-linked genes, *PRPS1* and *PRPS2* (Taira *et al.*, 1990). The *D. rerio* ortholog of the same gene, phosphoribosyl pyrophosphate synthetase 1B (*prps1b;* Gene ID: 560827) is expressed in 29 organs, with highest expression level in mature ovarian follicle (Bastian *et al.*, 2008).

To examine whether the SNPs associated with sex could have functional consequences, the locations of the SNPs were investigated in more detail. Of these 584 SNPs, roughly half were located outside protein-coding genes (Table 6). Among these, 74 SNPs were located within 1 kbp upstream of genes, a region where regulatory sequences of the genes often (but not always) reside. Amid these 74 SNPs 10 were located within 100 bp upstream of genes, where the TATA box and B recognition element are located in eukaryotes and archaea. These SNPs could potentially affect the expression of genes.

The other half of the SNPs were located in protein-coding genes. However, the overwhelming majority among these SNPs (222) were located in introns (Table 6). Among the 73 SNPs located in exons, 30 SNPs caused nonsynonymous substitutions (Table 6). Among the 20 genes in the sex regions, 10 genes had significant SNPs that caused nonconservative nonsynonymous substitutions in exons. These substitutions were in *tpcn1, iqcd,* T-box transcription factor TBX5-like (*loc105890445*), *loc105890446, loc105890447,* claudin-4-like (*loc105890498*), claudin-4-like (*loc105890449*), *loc105890474, catsper3,* and bone morphogenetic protein receptor type-1B-like (*loc105911882).* The nonsynonymous SNPs and their corresponding amino acid substitutions are listed in Table 7.

Sex specific insertions and deletions (indels) in the exons of the 20 genes were also investigated. There were 4 deletions and 1 insertion that were only present in males. None were found unique to females. These indels are listed in Table 8, along with the genotypes for the male fish. Indels 1, 4 and 5 caused frameshifts and would therefore most likely have a strong effect on the subsequent protein function. Indels 2 and 3 did not cause frameshifts and would give a loss of 1 and 13 amino acids, respectively. These deletions may or may not influence the protein function.

Table 5. Genes on the regions associated with sex in Atlantic herring *Clupea harengus*. There were no predicted genes on SR2, SR3 and SR4.1.

SR1.1 - NW_012219506.1: 919996 - 1039292

Gene	Product	GeneID:	Location
abhd11	Abhydrolase domain containing 11	105890485	C 999675 – 1005968
cldn3	Claudin 3	105890497	C 1011020 – 1011667
iqcd	IQ motif containing D	105890443	942869 – 945013
loc105890445	T-box transcription factor TBX5-like	105890445	C 952223 – 956360
loc105890446	Protein KIAA2022-like	105890446	C 970114 – 973638
loc105890447	Uncharacterized LOC105890447†	105890447	994211 – 996255
loc105890448	Claudin-4-like	105890448	1025734 – 1026366
loc105890449	Claudin-4-like	105890449	1028925 – 1029878
loc105890474	Merlin-like	105890474	1035468 – 1050358
loc105890483	Ribose-phosphate pyrophosphokinase 1-like	105890483	C 958340 – 966172
loc105890490	Protein NipSnap homolog 2-like	105890490	984248 – 992195
loc105890498	Claudin-4-like	105890498	1017053 – 1017704
loc105890500	Claudin-4-like a	105890500	1020757 – 1021389
loc105890503	Claudin-4-like b	105890503	1023975 – 1024544
loc105890510	Uncharacterized LOC105890510†	105890510	996295 – 997511
tpcn1	Two pore segment channel 1	105890473	C 927449 – 941538
mettl27	Methyltransferase like 27	105890450	C 1031693 – 1034546

SR1.2 - NW_012219506.1: 2305958 - 2321039

Gene	Product	GeneID	Location
loc105890513	Pterin-4-alpha-carbinolamine dehydratase 2-like	105890513	2290066 – 2319170
catsper3	Cation channel, sperm associated 3	105890515	2311908 – 2313564

SR4.2 - NW_012223947.1:7965797 - 7971594

Gene	Product	GeneID	Location
loc105911882	Bone morphogenetic protein receptor type-1B-like	105911882	7959993 – 7968805

† Both *loc105890447* and *loc105890510* show similarities to bicaudal D-related protein-like orthologs. See Supplement file 3.

Table 6. Location of the SNPs associated with sex in Atlantic herring *Clupea harengus*.

Location of SNPs associated with sex	Number of SNPs
In protein-coding genes	295
In introns	222
In exons	73
In predicted untranslated regions (UTR)	13
Synonymous substitutions	30
Nonsynonymous substitutions	30
Conservative substitutions	4
Nonconservative substitutions†	26
Not in genes	289
1 kbp upstream of gene start	74
100 bp upstream of gene start	10
Total	584

†Detail of these nonconservative nonsynonymous substitutions are listed in Table 7.

The indel genotypes followed the same pattern as the SNPs, i.e. females were homozygous for the reference allele (indel not present) and the majority of males were heterozygous for the indels. No single indel was present in all male fish, making it less likely that only one of these indels results in the male phenotype, although because of the low sequencing depth some indels could be undetected, similar to the SNPs mentioned above. None of the amino acid substitutions in the exons nor the indels gave a clear suggestion of a single gene that could be the sex determination gene.

None of the known sex determination or differentiation genes in fish were found on or close to the sex regions identified in this study. This could suggest that *C. harengus* has an unknown sex determination mechanism.

Table 7. Nonconservative nonsynonymous substitutions in genes on the Atlantic herring *Clupea harengus* genome caused by SNPs significantly associated with sex.

Gene	Chromosome	Position	Substitution
tpcn1	NW_012219506.1	928664	Ala -> His
	NW_012219506.1	929843	Ser -> Ala
	NW_012219506.1	936941	Pro -> Thr
	NW_012219506.1	937019	Ala -> Ser
iqcd	NW_012219506.1	942896	Gln -> Lys
	NW_012219506.1	944399	Ser -> Phe
loc105890445	NW_012219506.1	952391	His -> Asn
	NW_012219506.1	953918	Ser -> Ala
	NW_012219506.1	954778	Ala -> Ser
loc105890446	NW_012219506.1	971050	His -> Asn
	NW_012219506.1	971891	Lys -> Thr
	NW_012219506.1	971908	Ala -> Ser
	NW_012219506.1	971918	Cys -> Phe
	NW_012219506.1	972354	Gln -> His
	NW_012219506.1	972965	Ala -> Asp
	NW_012219506.1	973180	Stop -> Gly
loc105890447	NW_012219506.1	994532	Arg -> Gly
loc105890498	NW_012219506.1	1017621	Pro -> Thr
loc105890449	NW_012219506.1	1029733	Gly -> Glu
	NW_012219506.1	1029837	Ala -> Thr
loc105890474	NW_012219506.1	1039292	Arg -> Ser
catsper3	NW_012219506.1	2312043	Ser -> Arg
	NW_012219506.1	2312062	Ser -> Ala
loc105911882	NW_012223947.1	7965877	Tyr -> Phe
	NW_012223947.1	7965966	Lys -> Glu
	NW_012223947.1	7968244	Met/start -> Ile

Table 8. Sex specific deletions and insertions in the exonic regions of the genes on the sex regions of the Atlantic herring *Clupea harengus* genome. There were 55 males but not all individuals have data for all variations because of the low sequencing coverage.

Indel no.	Gene	Position	Deletion size	Insertion size	Male genotypes (H [†]/R [‡]/A [§])
1	*abhd11*	NW_012219506.1:1000250	10	-	19/22/4
2	*loc105890448*	NW_012219506.1:1026313	3	-	26/11/9
3	*catsper3*	NW_012219506.1:2311993	39	-	16/35/0
4	*catsper3*	NW_012219506.1:2312476	1	-	33/20/1
5	*loc105890490*	NW_012219506.1:992162	-	8	27/13/10

[†] H = Heterozygous

[‡] R = Homozygous reference allele

[§] A = Homozygous alternative allele

3.3.5. Discussion

We identified six regions on the *C. harengus* genome that were associated with sex; four of these (on two scaffolds) showed higher association than the others (SR1.1, SR1.2, SR4.1 and SR4.2). The data strongly indicates that females are homozygous, while the males are heterozygous for the SNPs in these sex-associated regions. This is consistent with a XY sex determination system. There are 20 protein-coding genes in these significant regions but no obvious MSR genes. However, some of these genes could potentially be affecting sex determination or development, as they are associated with sex organs or sex functions as shortly referred to in the Results section. Neither the investigation of the amino acid substitutions caused by SNPs nor the investigation of indels pointed to a simple monogenic sex determination system in *C. harengus*.

Low sequencing coverage

The SNPs were identified by low coverage whole-genome sequencing (on average 3 to 4x over the whole genome). This potentially results in some caveats regarding the genotypes. First, it is more likely to miss genotypic data for some of the SNPs in some of the individuals, simply because the area has not been sequenced. Second, sequencing errors are more likely to be implemented as variations and could result in falsely called low-frequency alleles. This is not a problem in the present situation as we are dealing with high-frequency alleles. Third, and more seriously, if, by chance, only one of the alleles from a heterozygous individual is sequenced, the genotype would always be called as a homozygous. With an average coverage of 3x, the probability of sequencing only one of the two alleles, is on average 12.5% (and 12.5% for the other allele). Thus, statistically we will achieve a 75% detection rate in a group consisting of 100% heterozygotes (Chenuil, 2012). Our data are rather close to this theoretical expectation with about 15.0% (3761/25141) of male genotypes called as homozygous reference allele, another 14.6% (3681/25141) as homozygous alternative allele and 70.4% (17699/25141) as heterozygotes. Our results also show that the male homozygous genotypes have on average lower coverage than the heterozygous genotypes, making it more probable that they are miscalled (Table 3). Furthermore, the observed male proportions of homozygotes versus coverage followed the same trend as the theoretically expected proportions, if all genotypes were truly heterozygous (Figure 3). We are therefore tempted to claim that all the significant SNPs, and in particular those that are biologically significant, are heterozygotes in males.

One way to verify this would be to repeat the experiment with higher coverage. Meynert *et al.* (2014) showed experimentally that 9-13x coverage was needed to correctly call 95% of heterozygous genotypes. Chenuil (2012, Figure 1) showed that with a coverage of 5x (where all reads show the same allele) and a heterozygous rate of 0.5 the homozygous genotype would be correct 95% of the times. Therefore, a read depth of more than 5 would be appropriate to increase sensitivity of correct genotype to above 95% for both homo- and heterozygotes.

In our study, the number of individuals sequenced partly makes up for the weakness caused by low coverage and shows that 98.9% of the female genotypes are homozygous while at least 70.4% of the males are heterozygous. When genotyping an ideal male heterogametic sex determining system with sex linked SNP markers, we would expect females to be homozygous, while males would be heterozygous at these markers. The very high proportion of homozygous females (98.9%) strongly support this hypothesis, while the measured proportion of heterozygous males is limited by the much lower heterozygous sensitivity of the method at low coverage.

Sex regions on the C. harengus genome

The association between sex and SR1.1 was stronger than for the other regions, and it was also the region that possessed the highest number of protein-coding genes (Table 5). As sex chromosomes evolve, they tend to become less stable and accumulate genes that are sex specific/beneficial, and eventually recombination between the homologous chromosomes stops and they become heteromorphic over time (reviewed in (Charlesworth *et al.*, 2005)). However, not all species develop heteromorphic sex chromosomes (Wright *et al.*, 2016), for example the tiger pufferfish *Takifugu rubripes* (Temminck & Schlegel 1850) where only one SNP is associated with the phenotypic sex (Kamiya *et al.*, 2012). Our results show that larger regions are associated with sex in *C. harengus* but if these are early heteromorphic sex chromosomes in development or not, we cannot tell. The *C. harengus* genome assembly is highly fragmented (6,915 scaffolds), so it is possible that all the sex regions identified in this study are on the same chromosome, but the assembly is not yet of high enough quality to show this. We think this is likely, because the random segregation of chromosomes during meiosis would otherwise ensure that the different sex regions sometimes would end up in different gametes, and thereby distribute among both males and females in the offspring (assuming only two sexes in herring).

SR2 also shows very strong association with sex. As can be seen at far right in Figure 2, the females tended to have no data here (tendency to black vertical line), while the males are homozygous for the alternative allele (tendency to brown vertical line). This suggests a deletion at this location, but no deletion was called by FreeBayes. When manually looking at the alignment of sequencing reads to this location of the draft genome, there were fewer reads from the female fish, but there was not a clear difference between the females and the males (data not shown). The genome assembly seems to be of poor quality at this location making the interpretation difficult. Therefore, no conclusion about SR2 was made.

Search for potential genes involved in sex determination

As mentioned in the Results, the genes on the sex regions are not known MSR genes or known to be part of the sex determination pathway. Interestingly, among the 6 genes that we have pointed out as most likely candidates for being involved in sex-specific processes, only 1 (ribose-phosphate pyrophoshokinase 1-like; *loc105890483*) did not contain a significant SNP that caused a nonconservative nonsynonymous substitution (Table 7). This neither confirms or refutes the importance for any of these 6 genes, but it is interesting that in the gene *iqcd* there was a substitution at amino acid position 315, from serine (polar side chain) to phenylalanine (bulky nonpolar side chain). This position is in a coiled coil domain of the Iqcd protein in *D. rerio*, and this substitution could disrupt the folding of the coil and affect the activity of the protein. Catsper3 has a nonconservative substitution at amino acid position 52 (serine to alanine) which is located in a transmembrane domain (compared to *H. sapiens* CATSPER3), again possibly causing alterations to the 3D structure and functions of the protein. This is interesting because Catsper3 is important for fertilisation of *C. harengus* eggs (Yanagimachi *et al.*, 2017b) and in mice (Jin *et al.*, 2007), but we have no information for the present change in question. There were also significant SNPs present in probable promoter regions (within 1 kb upstream) of 17 of the 20 genes in the sex regions. These SNPs could alter the expression of these genes and thereby affect the sex determination.

It is worth noting that the *C. harengus* genome assembly is quite recent (Martinez Barrio *et al.*, 2016) and there could still be several protein coding genes that are not yet predicted, partly because of potential suboptimal prediction algorithms and partly because they could be within non-sequenced regions. It is also likely that there are many nonidentified non-coding genes, e.g. lncRNAs or miRNAs. In addition to genes, there are also many regulatory elements that are not necessarily close to the genes they regulate. *C. harengus* is not a model organism so there are limited studies with this species, but the ENCODE project has inferred many functions for non-coding parts of the *H. sapiens* genome (ENCODE Project Consortium, 2012). It is very likely that similar non-coding elements exist in the *C. harengus* genome, and some of the SNPs found in this study to be associated with sex could affect a non-coding element that has not been identified yet.

Evolution of sex determination within the Clupeiformes order

Teleost fishes have very diverse sex determinations systems (Bachtrog *et al.*, 2014). The XY sex determination system for *C. harengus*, suggested in this study, fits well with the other Clupeiformes mentioned in the Introduction.

Work by Pennell *et al*. (2018) indicated that in fish, transitions from gonochoroism to hermaphroditism occur at higher rates than the reverse, and transitions from female to male heterogamety occur at higher rates than the reverse. They also found similar transition rates between homomorphic and heteromorphic sex chromosomes in both fish and amphibians. This could suggest that the common ancestor for Clupeidae and Engraulidae had a Z0 or ZW sex determination system, which is still present in *Coilia nasus*. The common ancestor for Clupeidae then lost the Z chromosome and adapted to a XY system, which has been found in *Brevoortia* and now also *Clupea*. These sex chromosomes are early in their evolution and are still homomorphic, as seen in *Brevoortia* spp and *C. harengus*. A single known exception is *B. aurea*, a species that has heteromorphic sex chromosomes with two X and one Y chromosome (Brum, 1992). As *Tenulosa* split from *Brevoortia* and *Clupea* they evolved to be hermaphrodites. This is of course at present just a speculative hypothesis.

Conclusion and future

We identified regions on the *C. harengus* genome that were associated with sex and the genotypes of the SNPs associated with sex indicated a XY sex determination system for *C. harengus*, which is consistent with the other Clupeiformes species. Nonetheless, the exact genes for sex determination were not identified. None of the known sex determination genes in fish were found on or close to the sex regions, indicating that *C. harengus* could have a previously unregistered unknown sex determination mechanism. New experiments where these sex regions are sequenced at a higher coverage for both males and females should be carried out to reproduce and better delineate the sex determination regions. This would also better characterize the potential existence of homozygous SNPs in a small proportion of the males in these regions.

Acknowledgements

We thank the staff onboard local fishing boats 'Gunnvør', 'Sildin' and 'Grani', G. J. Óskarsson from the Marine and Freshwater Research Institute Iceland, and P. Macdonald, L. Henderson and M. Hamilton from the NAFC Marine Centre Shetland for providing samples. In addition, we thank J. A. Thomassen and P. Vestergaard from the Faroe Marine Research Institute for helping collect fish data and tissue and L. H. G. Larsen from Amplexa Genetics for help regarding laboratory methods. We are also grateful to H. Joensen, D. H. Christiansen and J. Vang for help during acquisition of funding. A special thanks to the European Molecular Biology Laboratory for access to computer resources.

Supporting Information

Supplement File 1. List of SNP associated with sex in Atlantic herring.

Supplementary File 2. Test results from the comparison of the observed proportions of homozygous female and male genotypes versus coverage to the corresponding theoretically expected probabilities

Supplement File 3. Alignment of *loc105890447* and *loc105890510* to bicaudal D-related protein-like orthologs.

3.3.6. Contributions

SíK: Carried out the laboratory work, the analysis and interpretation of the data, as well as writing the manuscript. SOM: Contributed to the design of the study, writing of the manuscript and supervised the laboratory work, analysis and interpretation of data. EíH and JAJ: Contributed to the acquisition and interpretation of the data. HG: Contributed to statistical analysis, interpretation of the data, acquired funding and contributed to writing of the manuscript. PF: Contributed to the design of the study, writing of the manuscript, analysis and interpretation of the data. HAD: Designed the study, acquired funding and contributed to writing of the manuscript and supervised the laboratory work, analysis and interpretation of data. All authors contributed to revision of the manuscript and approved the final version.

Significance Statement

Sex determination systems within teleost fish are diverse, making teleost fish good candidates for studying the evolution of sex determination. To investigate the sex determination system of *Clupea harengus*, we performed a genome wide association study and identified six regions on the genome associated with sex. Closer inspection indicated a male heterogametic system that could be polygenic, fitting well with the other Clupeiformes species. However, the exact genes controlling sex determination were not identified.

3.3.7. References

Allsop, D. J. & West, S. A. (2003). Constant relative age and size at sex change for sequentially hermaphroditic fish. *Journal of Evolutionary Biology* **16**, 921-929.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Ashman, T.-L., Bachtrog, D., Blackmon, H., Goldberg, E. E., Hahn, M. W., Kirkpatrick, M., Kitano, J., Mank, J. E., Mayrose, I. & Ming, R. (2014). Tree of Sex: A database of sexual systems. *Scientific Data* **1**, 140015.

Bachtrog, D., Mank, J. E., Peichel, C. L., Kirkpatrick, M., Otto, S. P., Ashman, T.-L., Hahn, M. W., Kitano, J., Mayrose, I. & Ming, R. (2014). Sex determination: why so many ways of doing it? *PLoS Biology* **12**, e1001899.

Baroiller, J.-F., D'Cotta, H., Bezault, E., Wessels, S. & Hoerstgen-Schwark, G. (2009). Tilapia sex determination: where temperature and genetics meet. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology* **153**, 30-38.

Bastian, F., Parmentier, G., Roux, J., Moretti, S., Laudet, V. & Robinson-Rechavi, M. (2008). Bgee: integrating and comparing heterogeneous transcriptome data among species. In *International Workshop on Data Integration in the Life Sciences*, pp. 124-131: Springer.

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

Brum, M. (1992). Multiple sex chromosomes in South Atlantic fish, *Brevoortia aurea*, Clupeidae. *Brazilian Journal of Genetics* **15**, 547-553.

Brykov, V. A. (2014). Mechanisms of sex determination in fish: evolutionary and practical aspects. *Russian Journal of Marine Biology* **40**, 407-417.

Bull, J. J. (1983). *Evolution of sex determining mechanisms*: The Benjamin/Cummings Publishing Company, Inc.

Buston, P. (2003). Social hierarchies: size and growth modification in clownfish. *Nature* **424**, 145.

Cantagrel, V., Lossi, A., Boulanger, S., Depetris, D., Mattei, M., Gecz, J., Schwartz, C., Van Maldergem, L. & Villard, L. (2004). Disruption of a new X linked gene highly expressed in brain in a family with two mentally retarded males. *Journal of Medical Genetics* **41**, 736-742.

Charlesworth, D., Charlesworth, B. & Marais, G. (2005). Steps in the evolution of heteromorphic sex chromosomes. *Heredity* **95**, 118-128.

Chen, S., Zhang, G., Shao, C., Huang, Q., Liu, G., Zhang, P., Song, W., An, N., Chalopin, D. & Volff, J.-N. (2014). Whole-genome sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle. *Nature Genetics* **46**, 253-260.

Chenuil, A. (2012). How to infer reliable diploid genotypes from NGS or traditional sequence data: from basic probability to experimental optimization. *Journal of Evolutionary Biology* **25**, 949-960.

Clinton, M. (1998). Sex determination and gonadal development: a bird's eye view. *Journal of Experimental Zoology* **281**, 457-465.

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

Doucette Jr, A. J. & Fitzsimons, J. M. (1988). Karyology of elopiform and clupeiform fishes. *Copeia*, 124-130.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74.

Fadista, J., Manning, A. K., Florez, J. C. & Groop, L. (2016). The (in) famous GWAS P-value threshold revisited and updated for low-frequency variants. *European Journal of Human Genetics* **24**, 1202-1205.

Fricke, H. W. (1979). Mating system, resource defence and sex change in the anemonefish *Amphiprion akallopisos*. *Zeitschrift für Tierpsychologie* **50**, 313-326.

Furuta, T., Ashikari, M., Jena, K. K., Doi, K. & Reuscher, S. (2017). Adapting genotyping-by-sequencing for rice F2 populations. *G3: Genes, Genomes, Genetics* **7**, 881-893.

Garrison, E. & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*.

Hattori, R. S., Murai, Y., Oura, M., Masuda, S., Majhi, S. K., Sakamoto, T., Fernandino, J. I., Somoza, G. M., Yokota, M. & Strüssmann, C. A. (2012). A Y-linked anti-Müllerian hormone duplication takes over a critical role in sex determination. *Proceedings of the National Academy of Sciences* **109**, 2955-2959.

Jin, J., Jin, N., Zheng, H., Ro, S., Tafolla, D., Sanders, K. M. & Yan, W. (2007). Catsper3 and Catsper4 are essential for sperm hyperactivated motility and male fertility in the mouse. *Biology of Reproduction* **77**, 37-44.

Kamiya, T., Kai, W., Tasumi, S., Oka, A., Matsunaga, T., Mizuno, N., Fujita, M., Suetake, H., Suzuki, S. & Hosoya, S. (2012). A trans-species missense SNP in Amhr2 is associated with sex determination in the tiger pufferfish, *Takifugu rubripes* (fugu). *PLoS Genetics* **8**, e1002798.

Kashimada, K. & Koopman, P. (2010). Sry: the master switch in mammalian sex determination. *Development* **137**, 3921-3930.

Kriventseva, E. V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F. A. & Zdobnov, E. M. (2019). OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research* **47**, D807-D811.

Lagomarsino, I. V. & Conover, D. O. (1993). Variation in environmental and genotypic sex-determining mechanisms across a latitudinal gradient in the fish, *Menidia menidia*. *Evolution* **47**, 487-494.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. In *arXiv preprint arXiv:1303.3997*.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079.

Liu, H., Pang, M., Yu, X., Zhou, Y., Tong, J. & Fu, B. (2018). Sex-specific markers developed by next-generation sequencing confirmed an XX/XY sex determination system in bighead carp (*Hypophthalmichthys nobilis*) and silver carp (*Hypophthalmichthys molitrix*). *DNA Research* **25**, 257-264.

Martinez Barrio, A., Lamichhaney, S., Fan, G., Rafati, N., Pettersson, M., Zhang, H., Dainat, J., Ekman, D., Höppner, M., Jern, P., Martin, M., Nystedt, B., Liu, X., Chen, W., Liang, X., Shi, C., Fu, Y., Ma, K., Zhan, X., Feng, C., Gustafson, U., Rubin, C. J., Sällman Almén, M., Blass, M., Casini, M., Folkvord, A., Laikre, L., Ryman, N., Ming-Yuen Lee, S., Xu, X. & Andersson, L. (2016). The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. *eLife* **5**, e.12081.

Palaiokostas, C., Bekaert, M., Taggart, J. B., Gharbi, K., McAndrew, B. J., Chatain, B., Penman, D. J. & Vandeputte, M. (2015). A new SNP-based vision of the genetics of sex determination in European sea bass (*Dicentrarchus labrax*). *Genetics Selection Evolution* **47**, 68.

Pennell, M. W., Mank, J. E. & Peichel, C. L. (2018). Transitions in sex determination and sex chromosomes across vertebrate species. *Molecular Ecology* **27**, 3950-3963.

Pieau, C. (1996). Temperature variation and sex determination in reptiles. *BioEssays* **18**, 19-26.

Purcell, C. M., Seetharam, A. S., Snodgrass, O., Ortega-García, S., Hyde, J. R. & Severin, A. J. (2018). Insights into teleost sex determination from the *Seriola dorsalis* genome assembly. *BMC Genomics* **19**, 31.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. & Daly, M. J. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81**, 559-575.

Reed, E., Nunez, S., Kulp, D., Qian, J., Reilly, M. P. & Foulkes, A. S. (2015). A guide to genome-wide association analysis and post-analytic interrogation. *Statistics in Medicine* **34**, 3769-3792.

Roberts, N. B., Juntti, S. A., Coyle, K. P., Dumont, B. L., Stanley, M. K., Ryan, A. Q., Fernald, R. D. & Roberts, R. B. (2016). Polygenic sex determination in the cichlid fish *Astatotilapia burtoni*. *BMC Genomics* **17**, 835.

Ser, J. R., Roberts, R. B. & Kocher, T. D. (2010). Multiple interacting loci control sex determination in lake Malawi cichlid fish. *Evolution: International Journal of Organic Evolution* **64**, 486-501.

Shen, Z.-G. & Wang, H.-P. (2014). Molecular players involved in temperature-dependent sex determination and sex differentiation in Teleost fish. *Genetics Selection Evolution* **46**, 26.

Strünker, T., Goodwin, N., Brenker, C., Kashikar, N. D., Weyand, I., Seifert, R. & Kaupp, U. B. (2011). The CatSper channel mediates progesterone-induced Ca^{2+} influx in human sperm. *Nature* **471**, 382-386.

Sun, C., Niu, Y., Ye, X., Dong, J., Hu, W., Zeng, Q., Chen, Z., Tian, Y., Zhang, J. & Lu, M. (2017). Construction of a high-density linkage map and mapping of sex determination and growth-related loci in the mandarin fish (*Siniperca chuatsi*). *BMC Genomics* **18**, 446.

Taira, M., Iizasa, T., Shimada, H., Kudoh, J., Shimizu, N. & Tatibana, M. (1990). A human testis-specific mRNA for phosphoribosylpyrophosphate synthetase that initiates from a non-AUG codon. *Journal of Biological Chemistry* **265**, 16491-16497.

Turner, S. D. (2014). qqman: an R package for visualizing GWAS results using QQ and manhattan plots. *BioRxiv*, 005165.

Wilson, C. A., High, S. K., McCluskey, B. M., Amores, A., Yan, Y.-l., Titus, T. A., Anderson, J. L., Batzel, P., Carvan, M. J. & Schartl, M. (2014). Wild sex in zebrafish: loss of the natural sex determinant in domesticated strains. *Genetics* **198**, 1291-1308.

Wright, A. E., Dean, R., Zimmer, F. & Mank, J. E. (2016). How to make a sex chromosome. *Nature Communications* **7**, 12087.

Yanagimachi, R., Harumi, T., Matsubara, H., Yan, W., Yuan, S., Hirohashi, N., Iida, T., Yamaha, E., Arai, K., Matsubara, T., Andoh, T., Vines, C. & Cherr, G. N. (2017a). Chemical and physical guidance of fish spermatozoa into the egg through the micropyle. *Biology of Reproduction* **96**, 780-799.

Yanagimachi, R., Harumi, T., Matsubara, H., Yan, W., Yuan, S., Hirohashi, N., Iida, T., Yamaha, E., Arai, K., Matsubara, T., Andoh, T., Vines, C. & Cherr, G. N. (2017b). Chemical and physical guidance of fish spermatozoa into the egg through the micropyle. *Biology of Reproduction* **96**, 780-799.

Yano, A., Guyomard, R., Nicol, B., Jouanno, E., Quillet, E., Klopp, C., Cabau, C., Bouchez, O., Fostier, A. & Guiguen, Y. (2012). An immune-related gene evolved into the master sex-determining gene in rainbow trout, Oncorhynchus mykiss. *Current Biology* **22**, 1423-1428.

Yoshimoto, S. & Ito, M. (2011). A ZZ/ZW-type sex determination in Xenopus laevis. *The FEBS journal* **278**, 1020-1026.

3.4. Atlantic herring (*Clupea harengus*) population structure in the Northeast Atlantic Ocean

Authors

Sunnvør í Kongsstovu[1,2,5], Svein-Ole Mikalsen[2], Eydna í Homrum[3], Jan Arge Jacobsen[3], Thomas D. Als[4], Hannes Gislason[2], Paul Flicek[5] and Hans Atli Dahl[6]

[1] Amplexa Genetics A/S, Hoyvíksvegur 51, FO-100 Tórshavn, Faroe Islands.

[2] University of the Faroe Islands, Dept. of Science and Technology, Vestara Bryggja 15, FO-100 Tórshavn, Faroe Islands.

[3] Faroe Marine Research Institute, Nóatún 1, FO-100 Tórshavn, Faroe Islands.

[4] Aarhus University, Department of Biomedicine, Bartholins Allé 6, 8000 Aarhus C, Denmark.

[5] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK.

[6] Amplexa Genetics A/S, Tolderlundsvej 3B, 2. DK-5000 Odense C, Denmark.

3.4.1 Abstract

Atlantic herring *Clupea harengus* L has a vast geographical distribution and migrates between common feeding grounds, wintering areas, and traditional spawning grounds, thereby leading to spatiotemporal population structures. Herring population structure is complex with a few very large migratory stocks and many small local populations, each with their own spawning ground and time. During parts of the year, several herring populations frequently mix, resulting in mixed fisheries. Understanding the population structure is important for maintaining sustainable herring fisheries. Studies have shown that herring in the Baltic Sea and North Sea are genetically distinct from those in the Northeast Atlantic. Further population structure within the Northeast Atlantic has been identified but the studies have generally not been able to distinguish between all putative populations or stocks.

The aim of this study was to investigate the genetic population structure of herring in the Faroese and surrounding waters, and to develop genetic markers that could distinguish between four known herring stocks (Norwegian spring-spawning herring [NSSH], Icelandic summer-spawning herring [ISSH], North Sea autumn-spawning herring [NSAH], and Faroese autumn-spawning herring [FASH]). Herring from the four stocks were sequenced at low coverage, and single nucleotide polymorphisms (SNPs) were called and used for population structure analysis and individual assignment.

The results showed that all four stocks appeared to be genetically differentiated, but cluster-analysis only identified three clusters. A list of ancestry-informative SNPs enabling distinction between stocks performed well on the original data (90% assignment accuracy). However, when additional samples were genotyped, ISSH and FASH could not be clearly distinguished, but if samples from these two stocks were pooled, the overall assignment accuracy was 89%.

3.4.2 Introduction

The fishing industry and aquaculture are economical mainstays of the Faroe Islands. One of the major target species for fisheries is herring. In 2017, 108,244 tonnes of herring were caught by the Faroese fishing fleet, yielding 7.5% of the total value of exports [1]. Herring fisheries are also crucial for the rest of the world, both economically and as a nutritional resource. Herring is also an important part of the ecology in the Atlantic Ocean.

Atlantic herring (*Clupea harengus*) is a school-forming, pelagic fish that migrates between common feeding grounds, wintering areas, and traditional spawning grounds. Atlantic herring have a vast geographical distribution on both sides of the Atlantic Ocean. Here, our focus is on the east side of the Atlantic, where herring can be found from Svalbard south to the northern Bay of Biscay, and from South Greenland to Novaya Zemlya in Russia, including the Baltic Sea [2].

Numerous herring populations exist, each with their own spawning ground and time. Herring population structure is complex with a few very large migratory stocks and many small local stocks [3]. Both in the Baltic sea and around the British Isles there are herring components of several populations [4, 5], but here we focus on the population structure of herring in the Northeast Atlantic. Herring fisheries are managed in stocks, which are putative populations. Some stocks have indeed been shown to be distinct biological populations, whereas others have not been distinguished or studied. The Norwegian spring-spawning herring (NSSH) is the largest stock in the Northeast Atlantic. It spawns on the coast of Norway and feeds in the open ocean between Norway, the Faroe Islands, and Iceland. In Norwegian waters, there are also the Norwegian local spring-spawning herring (NLSSH) [6], mainly spawning in local fjords, and the Norwegian autumn spawning herring (NASH) [7]. The Icelandic summer-spawning herring (ISSH) can be found around Iceland as well as a small stock of Icelandic spring-spawning herring (ISPH), but the latter has not fully recovered from its collapse in the 1960s [8]. In addition, the NSSH stock migrates to Icelandic waters during parts of the year [8]. Historically, the NSSH spawned in Faroese waters but the stock collapsed in the 60s; this spawning site was abandoned and has not been re-established [9, 10]. Today, only the local Faroese autumn spawning herring (FASH), also called the fjord herring, spawns in Faroese waters. This stock is sporadic, and the exact spawning locations are unknown. Nevertheless, this fjord herring has been observed occasionally since at least the 1780s [10].

During parts of the year, NSSH, ISSH, and FASH all mix to some degree. In addition, the North Sea autumn-spawning herring (NSAH) stock can mix with the other stocks. This means that herring fisheries for these target stocks result in some degree of mixed catches, and distinguishing between the different stocks can be problematic. Morphological, physiological, and biological characteristics are examined to assign individuals to a stock, but these characteristics can be interpreted differently by different people [11].

Herring is an important national and international resource and keeping the fisheries sustainable is of major importance. Smith *et al.* [12] showed how neglecting to account for population structure in fisheries management can result in overexploitation and the loss of genetic diversity. Knowledge of population structure is necessary to ensure that fisheries target the intended population, as well as to set realistic regulations for fisheries management. Moreover, further understanding of population structure is crucial for understanding the distributional range and migration behaviour of the species. Population structure is not only important in maintaining sustainable fisheries but also in the fight against illegal, unreported, and unregulated fishing, as well as in the forensic identification of fish and fish products throughout the food processing chain.

Several methods for distinguishing between herring populations have been studied; for example, using phenotypic traits such as vertebrate count, the outline of otoliths, and otolith microchemistry [5, 13, 14]. These methods have been able to distinguish between populations to a varying degree, but a downside of these characteristics is that they are affected by the environment and serve merely as proxies for genetic differences. Different genetic methods based on microsatellites and/or single nucleotide polymorphisms (SNPs) have been conducted to investigate the population structure of Atlantic herring. McPherson *et al.* [15] demonstrated a significant difference between Atlantic herring in the Northeast and Northwest Atlantic, as well as among spawning groups in the Northwest Atlantic. Other studies have been able to show that both the Baltic herring and North Sea herring are genetically distinct from the herring in the Northeast Atlantic [16, 17]. Distinct populations have also been found within the Baltic Sea [18-20]. Similarly, Pampoulie *et al.* [11] showed that the local fjord herring (*i.e.,* NLSSH) is distinct from NSSH. The few studies that have included the FASH stock have not been able to distinguish it from the other Northeast herring [11, 21, 22]; only one study [23] has shown a difference between ISSH and NSSH, whereas others have not been able to replicate this [11].

The aim of this study was to find genetic markers that could distinguish between four herring stocks from the Northeast Atlantic to be used for individual assignment. A second aim was to investigate the population structure of herring in Faroese waters, specifically to investigate if the FASH stock is genetically distinct from the other stocks. This was done using low coverage sequencing of samples from the four stocks, and the identification of SNPs genetically differentiated between the stocks. Furthermore, population structure was analysed using cluster analysis.

3.4.3. Materials and methods

Sample collection

Herring samples were collected on a research cruise during the summer of 2015, conducted by the Faroe Marine Research Institute (FAMRI), obtained from fishing boats or kindly provided by NAFC Marine Centre in Scalloway or the Marine Research Institute in Iceland.

The length, weight, sex, and maturity stage of the fish were recorded, and the otoliths extracted. The fish age and spawning type were inferred from the otoliths; a hyaline otolith nucleus indicated autumn spawners and an opaque nucleus indicated spring spawners [24]. The maturity stage and spawning type were used, together with the location and time of catch, to identify which stock the individual fish belonged to. These are the methods used by FAMRI to assign herring from fisheries to stocks. Table 1 shows the time of catch and maturity stage distribution of the samples, and Figure 1 shows the sampling sites.

Table 1. Overview of the Atlantic herring samples used in this study. The stock was inferred from maturity stage, date, and location of the sample. NSSH = Norwegian spring-spawning herring, NSAH = North Sea autumn-spawning herring, FASH = Faroese autumn spawning herring, and ISSH = Icelandic summer-spawning herring. Samples marked with RC in column 12 were caught on a research cruise, whereas samples marked F were caught by fishing boats.

Sample	**Date**	**Sample**	**Maturity stage**								**Caught on RC or F**	**Stock**
		size	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**		
15520045	10-Jul-15	3			2		1				RC	NSSH
15520047	11-Jul-15	9			9						RC	NSSH
15520051	11-Jul-15	8			5	3					RC	NSSH
15520055	12-Jul-15	7			6		1				RC	NSSH
15520059	13-Jul-15	5			4	1					RC	NSSH
15520063	13-Jul-15	2			3						RC	NSSH
20155056	03-Dec-15	14			9	5					F	NSSH
20175044	25-Oct-17	41				41					F	NSSH
20160459	02-Jun-16	2				1	1				F	FASH
20175014	15-Feb-17	16	1	8	7						F	FASH
20175015	15-Feb-17	9	3	4	2						F	FASH
20145054	07-Dec-14	7	2	5							F	FASH
20155014	28-Aug-15	9					2	6		1	F	FASH
20175036	02-Oct-17	16		6	4	3	3				F	FASH
20175037	17-Aug-17	12			1	2	6	3			F	FASH
20175038	18-Sep-17	13		1	3	3	5	1			F	FASH
20175060	15-Oct-17	4						4			F	FASH
20185027	25-Jul-18	60		6	16	17	14	2		5	F	NSAH
20165079	01-Nov-16	17		4		1		5	7		F	NSAH
20175020	17-Feb-17	22	12	5	5						RC	ISSH
20175021	17-Feb-17	22	7	1	13					1	RC	ISSH
20175022	22-Feb-17	23			5	1				17	RC	ISSH
20175023	22-Feb-17	23			9	5	2			7	RC	ISSH

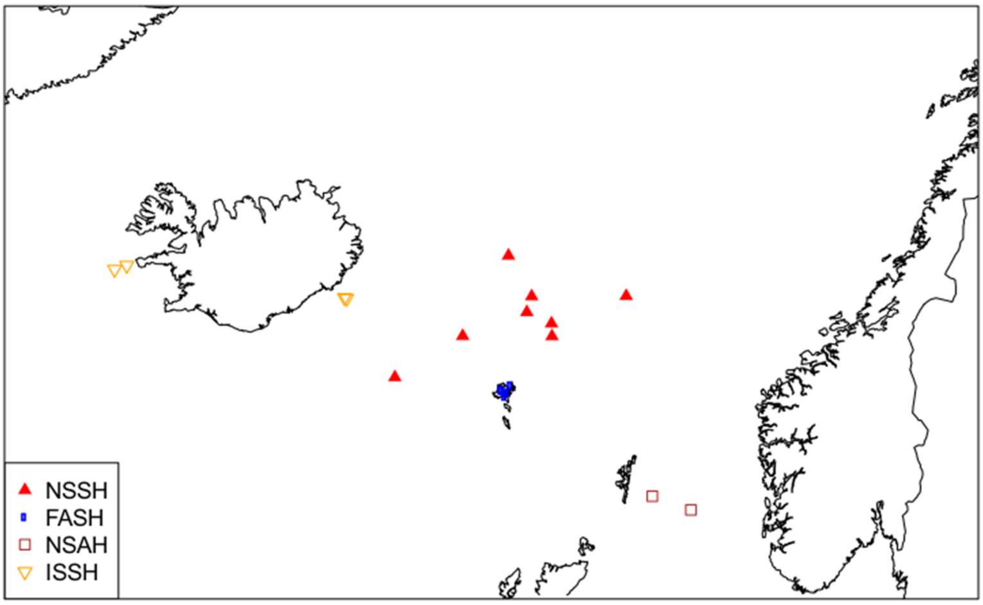


Figure 1. Sampling sites of the Atlantic herring used in this study. NSSH = Norwegian spring-spawning herring (▲), NSAH = North Sea autumn-spawning herring (□), FASH = Faroese autumn spawning herring (▪), and ISSH = Icelandic summer-spawning herring (▼).

Baseline samples

DNA extraction and sequencing

The DNA was extracted from tissue samples from 103 herring (29 NSSH, 30 ISSH, 17 NSAH, and 27 FASH) using an AS1000 Maxwell 16 instrument (Promega, Wisconsin, United States) and the Maxwell 16 Tissue DNA Purification Kit (Promega). DNA concentration was measured using a Qubit 3.0 fluorometer (ThermoFisher Scientific).

The DNA was fragmented to roughly 300 bp using a Covaris M220 (Covaris, Chicago, United States) and the libraries were prepared using the KAPA TLP Library Preparation Kit Illumina Platforms (KAPABiosystems, Massachusetts, United States), following the manufacturers' instructions. The library from each individual herring was indexed using indexed adapters (Pentabase, Odense, Denmark) and quantified using the KAPA Library Quantification Kit (KAPABiosystems), following the manufacturers' instructions. After quantification, the libraries were pooled to equal proportions and sequenced on a NextSeq500 (Illumina, California, United States) using the High Output v2 Kit (Illumina).

Sequencing data preprocessing

The sequencing data were trimmed to remove adapters and low-quality bases (Q score < 20) using Trimmomatic v0.36 [25]. AfterQC v0.4.0 was used to remove the polyG reads [26], and FastQC v0.11.5 was used to assess the quality of all the sequencing data [27]. The data were then aligned to the draft herring genome (GCF_000966335.1_ASM96633v1) using BWA-MEM v0.7.15 [28]. Furthermore, SAMtools v1.3 was used for sorting, converting, and removing PCR duplicates from the alignment files [29].

Population structure – Genotype call method

Population structure in the baseline samples was investigated using two methods. The first was using genotype calls as follows. SNPs were called using FreeBayes v1.1.0 with pooled data (sequencing reads from individuals from the same stock were pooled) and individual data [30]. Low-quality SNPs (QUAL < 20) were filtered out, and for the individual data, SNPs where fewer than 30 individuals had data (NS < 30) were also filtered out; and for the pooled data, SNPs with NS < 4 were filtered out.

SNP-wise Weir and Cockerham [31] F_{ST} were calculated for pairwise combinations of stocks using VCFtools v0.1.15 [32]. For every pairwise comparison, the 100 SNPs with the highest F_{ST} were selected. The duplicates were removed and the genotypes for these SNPs were extracted from the individual sequencing data. Pink v1.07 [33] was then used to calculate the linkage disequilibrium (LD) and to prune the SNPs based on the LD. The 154 remaining SNPs were used for further analysis.

The most likely number of populations (K) was calculated using STRUCTURE v2.3.4 [34], with an admixture model with correlated allele frequencies but without information on sample location. STRUCTURE ran 10 independent runs for K = 1–8 with 100,000 burn-ins and 100,000 iterations. Subsequently, Clumpak [35] was used to estimate the optimal number of K using the Evanno method [36]. Pairwise putative population difference and global F_{ST} values were calculated using the test_diff() and Fst() functions in the Genepop R package [37].

Population structure – Genotype likelihood method

Because of the low sequencing coverage, the genotype calls were prone to error. Therefore, population structure in the baseline samples was also investigated using NGSadmix in the ANGSD software package, which uses the genotype likelihood and works well with low sequencing coverage [38, 39]. Genotype likelihoods were called from the BAM files using ANGSD, with the p-value cut-off set at 10^-6. SNPs where more than 90 individuals had missing genotype likelihood as well as those with a minor allele frequency lower than 0.05 were filtered out.

Population structure was investigated using the resulting genotype likelihoods and NGSadmix for K = 1–8. In addition, a principle component analysis (PCA) was performed using the same genotype likelihoods and PCAngsd [40].

Test samples

Genotyping

Tissue samples from 240 herring (60 from each stock) were sent to LGC Genomics (Berlin, Germany) for genotyping, and 500 SNPs were genotyped using their SeqSNP service. These 500 SNPs were selected based on their discriminatory power between the stocks; 154 SNPs were selected based on the pooled data; and the rest were selected based on the individual-level data (using the same method).

The DNA extracted from these tissue samples was highly fragmented (fragment sizes below 1 kb) because of the nature of the samples (see the Discussion). The LGC SeqSNP method is best suited for fragments of 10 kb and higher. Nevertheless, these samples were typical samples from fishery landings—the type of samples in need of individual assignment in the event of mixed fisheries. For this reason, we chose to perform the genotyping experiment with the low-quality DNA.

Assignment

The baseline samples were subjected to a population assignment test using Monte-Carlo cross-validation, using the R package assignPOP [41] and the 154 SNPs selected earlier (above).

Sixty iterations were run with all loci and 50, 70, and 90% of the individuals were used for training.

Additionally, the test samples were subjected to a population assignment test using assignPOP (with the same parameters as described above) and the 500 genotyped SNPs, to test their discriminatory power. The test samples were also assigned to a population using the assign.X function in assignPOP with default parameters and the baseline samples as reference samples.

3.4.4. Results

Sequencing, SNP calling, and genotyping

Sequencing of the 103 herring genomes resulted in a total coverage of 386x and an average coverage ± standard deviation (SD) of 3.75x ± 1.79 at the individual level. Table 2 presents the coverage for each population. Using the individual data, 47,687,055 SNPs were called but because of the low average coverage, the sequencing reads from the individuals were pooled at the population level. Using this pooled data, 22,513,479 SNPs were called, and the list of selected SNPs was further narrowed based on F_{ST} (see the Methods) and LD-pruned, leading to a selection of 154 SNPs with the highest discriminatory power.

Due to suboptimal DNA quality, only 377 of 500 SNPs and 238 of 240 individual herring sent for genotyping at LGC Genomics passed the quality control (see the Methods).

Table 2. Sequencing coverage of the four Atlantic herring stocks. NSSH = Norwegian spring-spawning herring, NSAH = North Sea autumn-spawning herring, FASH = Faroese autumn spawning herring, ISSH = Icelandic summer-spawning herring, SD = standard deviation.

Stock	No. of individuals	Coverage (x)	Mean coverage ± SD (x)
NSSH	29	109	3.6 ± 1.0
ISSH	30	70	2.3 ± 0.8
NSAH	17	118	6.9 ± 1.1
FASH	27	102	3.6 ± 1.3

Population structure – Genotype call method

Significant genetic differences existed between all four putative populations based on both the 154 SNPs from the baseline samples and the 377 SNPs from the test samples (Table 3). In the STRUCTURE analysis, the mean likelihood plateaued at K = 3 (Figure 2a). However, the Evanno method showed that K = 2 was the most likely number of clusters (Figure 3a). For K = 2, NSSH formed one cluster while the other three stocks formed the second cluster (Figure 4). To search for substructures in the second cluster, the NSSH sample was removed and the analysis was run again. This time the most likely number of K was also 2 using the Evanno method (Figure 3a) and 4 using only STRUCTURE (Figure 2c). The NSAH sample formed one cluster and ISSH and FASH the other, apart from five FASH individuals that clustered with NSAH (Figures 4). This indicated that there was substructure in the second cluster that was detected in the first analysis. The five FASH individuals that clustered with NSAH were likely migrants from the NSAH stock that were caught in the Faroese fjords. A third analysis with only the ISSH and FASH samples was performed, and the most likely K was 4 with both STRUCTURE and the Evanno method (Figures 2e and 3a), but the STRUCTURE results were difficult to interpret. However, no meaningful spatial or temporal pattern was observed, which was interpreted as no substructure (Supplementary Figure S3).

The same analyses were repeated without the five suspected migrant individuals. This provided similar results for the most likely number of K (Figure 2b, d, and f and Figure 3b). However, the most likely number of K when only the FASH and ISSH samples were analysed was now 2 using the Evanno method (Figure 3b), which indicated substructure.

Barplots for all Ks from the six STRUCTURE analyses can be seen in Supplementary Figures S1–S6.

Table 3. Pairwise stock difference and F_{ST}. This was calculated using the test_diff() and Fst() functions in the Genepop R package. The 154 SNPs are from the sequencing experiment (baseline samples) and the 377 SNPs are from the genotyping experiments (test samples). NSSH = Norwegian spring-spawning herring, NSAH = North Sea autumn-spawning herring, FASH = Faroese autumn spawning herring, ISSH = Icelandic summer-spawning herring, df = degrees of freedom.

	154 SNPs				377 SNPs			
Stock pair	**Chi^2**	**df**	**p-value**	**F_{ST}**	**Chi^2**	**df**	**p-value**	**F_{ST}**
FASH/ISSH	960.62	306	< 0.00001	0.1352	1393.13	798	< 0.00001	0.0146
FASH/NSSH	550.40	308	< 0.00001	0.3910	1283.91	796	< 0.00001	0.1658
ISSH/NSSH	452.20	304	< 0.00001	0.4238	910.69	798	< 0.005	0.1850
FASH/NSAH	895.63	304	< 0.00001	0.2754	1220.92	798	< 0.00001	0.1288
ISSH/NSAH	842.47	302	< 0.00001	0.3801	1139.04	804	< 0.00001	0.2111
NSSH/NSAH	756.33	302	< 0.00001	0.3905	891.22	798	< 0.05	0.2389

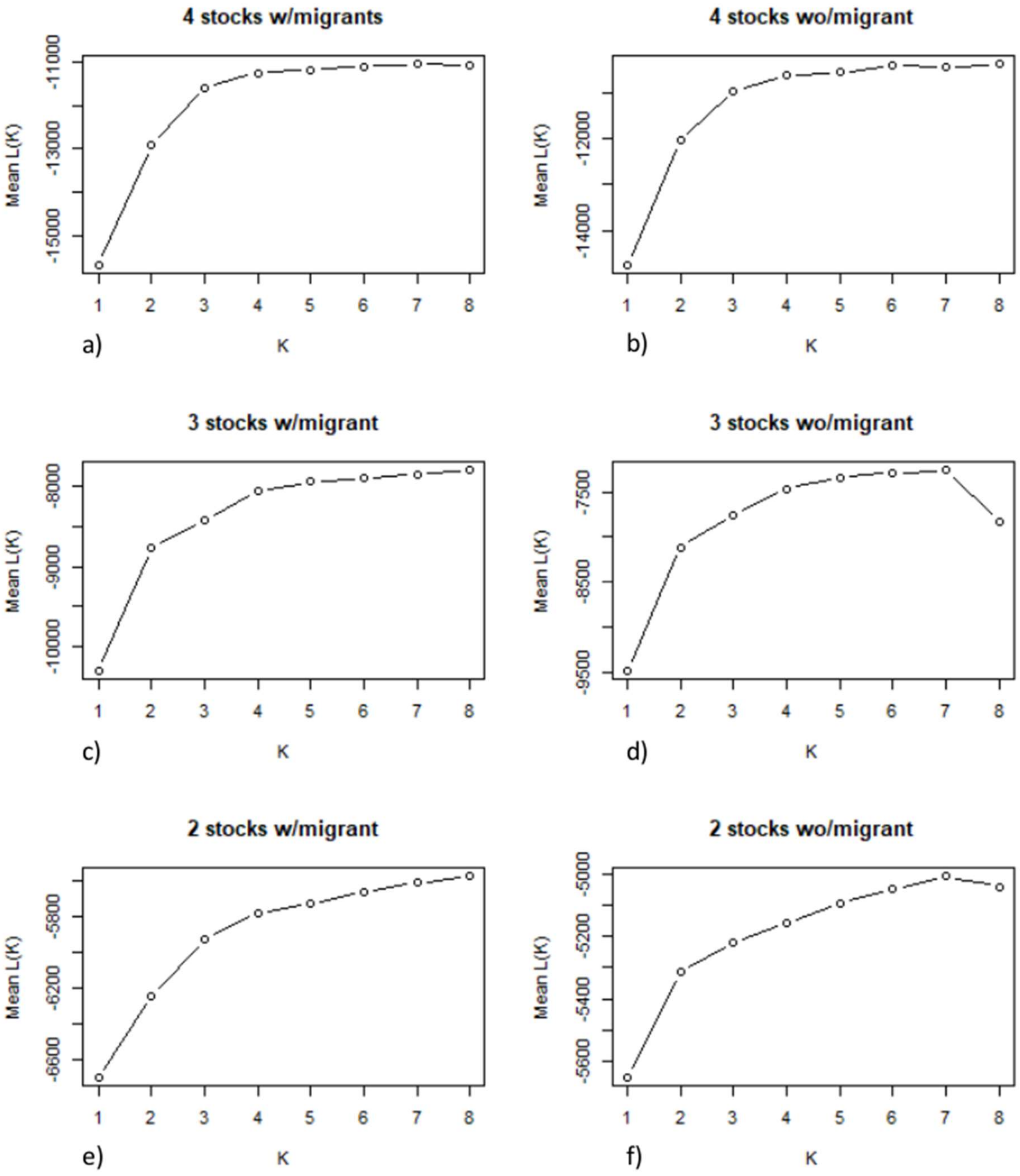


Figure 2. Mean likelihood of K given data from STRUCTURE analyses. The stocks used in a) and b) were Norwegian spring-spawning herring (NSSH), North Sea autumn-spawning herring (NSAH), Faroese autumn spawning herring (FASH), and Icelandic summer-spawning herring (ISSH). The stocks used in c) and d) were NSAH, FASH, and ISSH. The stocks used in e) and f) were FASH and ISSH. All analyses were run with an admixture model with correlated allele frequencies, but without information on sample location for 10 independent runs for K = 1–8 with 100,000 burn-ins and 100,000 iterations. In b), d), and f), five suspected migrant herring were removed (see text).

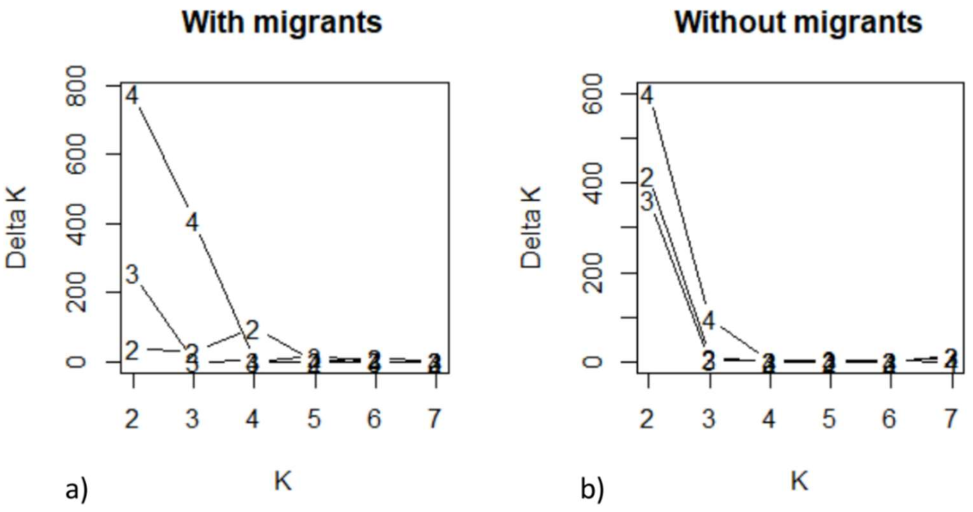


Figure 3. Delta K from the Evanno method of determining the most likely number of K, using STRUCTURE results. The data points show the number of stocks used in the analysis: 4 = Norwegian spring-spawning herring (NSSH), North Sea autumn-spawning herring (NSAH), Faroese autumn spawning herring (FASH), and Icelandic summer-spawning herring (ISSH). 3 = NSAH, FASH, and ISSH. 2 = FASH and ISSH. In b), five suspected migrant herring were removed (see text).

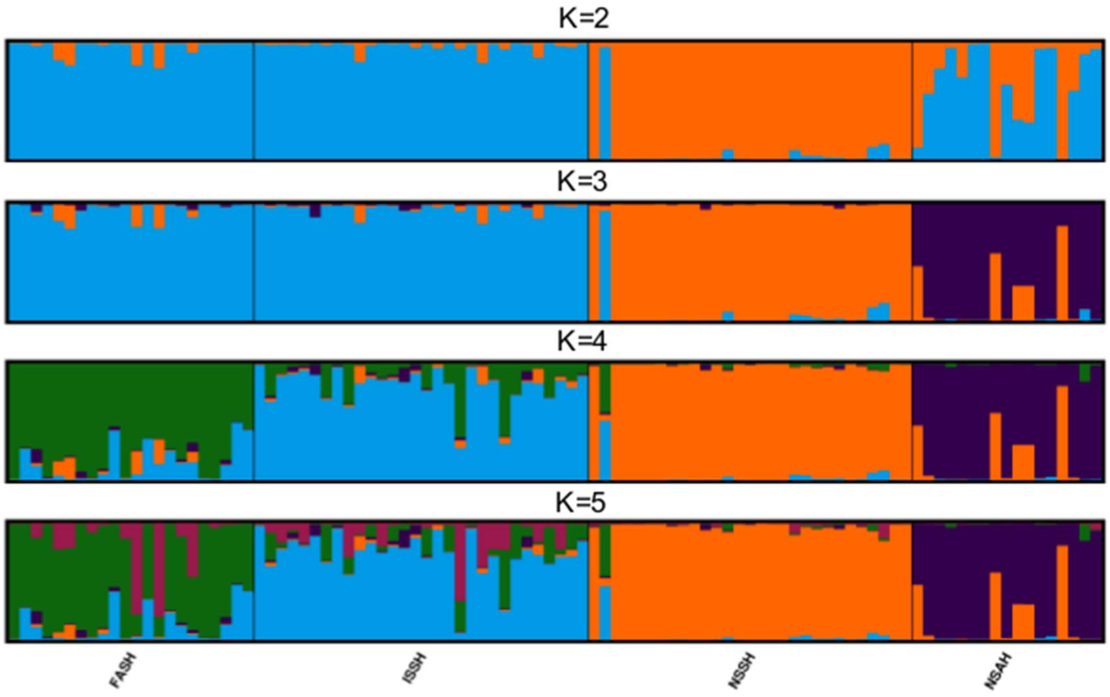


Figure 4. Barplots showing the population structure of Atlantic herring based on the genotype of 154 SNPs for K = 2–5. NSSH = Norwegian spring-spawning herring, NSAH = North Sea autumn-spawning herring, FASH = Faroese autumn spawning herring, and ISSH = Icelandic summer-spawning herring. Five migrant herring have been removed (see text.)

Population structure – Genotype likelihood method

The NGSadmix software and genotype likelihoods were used to investigate the admixture for K = 1–8, and 4,672,907 SNPs were used for this analysis. The same analysis was also run without the abovementioned migrant herring (Figure 5). The barplots from NGSadmix and STRUCTURE looked similar, but notably, the NSAH sample stood out in the NGSadmix K = 2 (Figure 5), whereas NSSH stood out in STRUCTURE K = 2 (Figure 4). At K = 4, FASH and ISSH seemed to have a dissimilar admixture in the STRUCTURE results, whereas the admixture looked highly similar in the NGSadmix results. The results for K = 6–8 (without migrants) can be seen in Supplementary Figure S7 and K = 2–8 for the same analysis, including the migrants can be seen in Supplementary Figure S8.

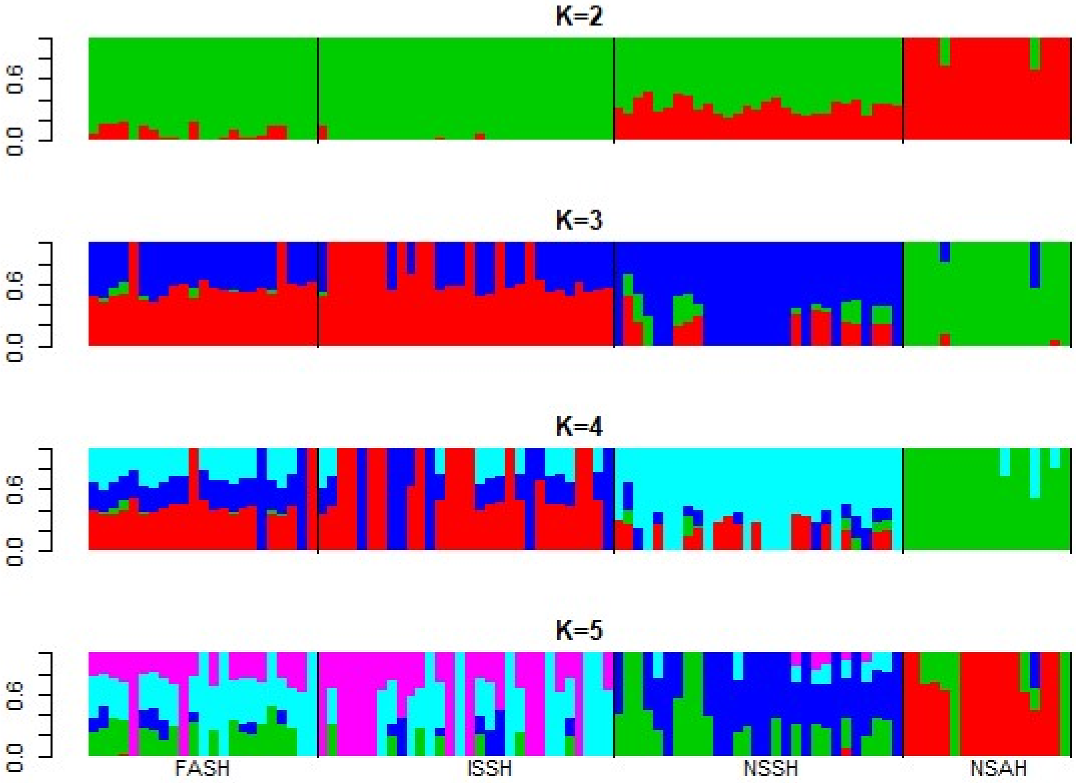


Figure 5. Barplots showing the population structure of Atlantic herring based on the genotype likelihood of 4.6 M SNPs for K = 2–5. NSSH = Norwegian spring-spawning herring, NSAH = North Sea autumn-spawning herring, FASH = Faroese autumn spawning herring, and ISSH = Icelandic summer-spawning herring. Five migrant herring have been removed (see text).

The PCA showed similar results as STRUCTURE and NGSadmix with three clusters (Figure 6). The NSSH samples clustered together, with the exception of one sample that clustered together with ISSH and FASH. The ISSH samples clustered closely together. Again, the FASH samples clustered together with the ISSH samples, except for five individuals that clustered with the NSAH samples. These were the aforementioned five suspected NASH migrants. The FASH individuals were mostly on the left side of the FASH–ISSH cluster, whereas the ISSH individuals were mostly on the right side of the cluster; but substantial overlap occurred in the middle. The NSAH samples did not cluster with samples from the other stocks (except for the five aforementioned FASH individulas), but they did not form a close cluster as the other stock samples did. They were more spread out and could be interpreted as several clusters, indicating substructure. The FASH–ISSH cluster was more closely inspected in a new PCA with only the FASH and ISSH samples and excluding the five suspected migrant herring. However, the results from this analysis provided similar results to the first PCA (Supplementary Figure S9)

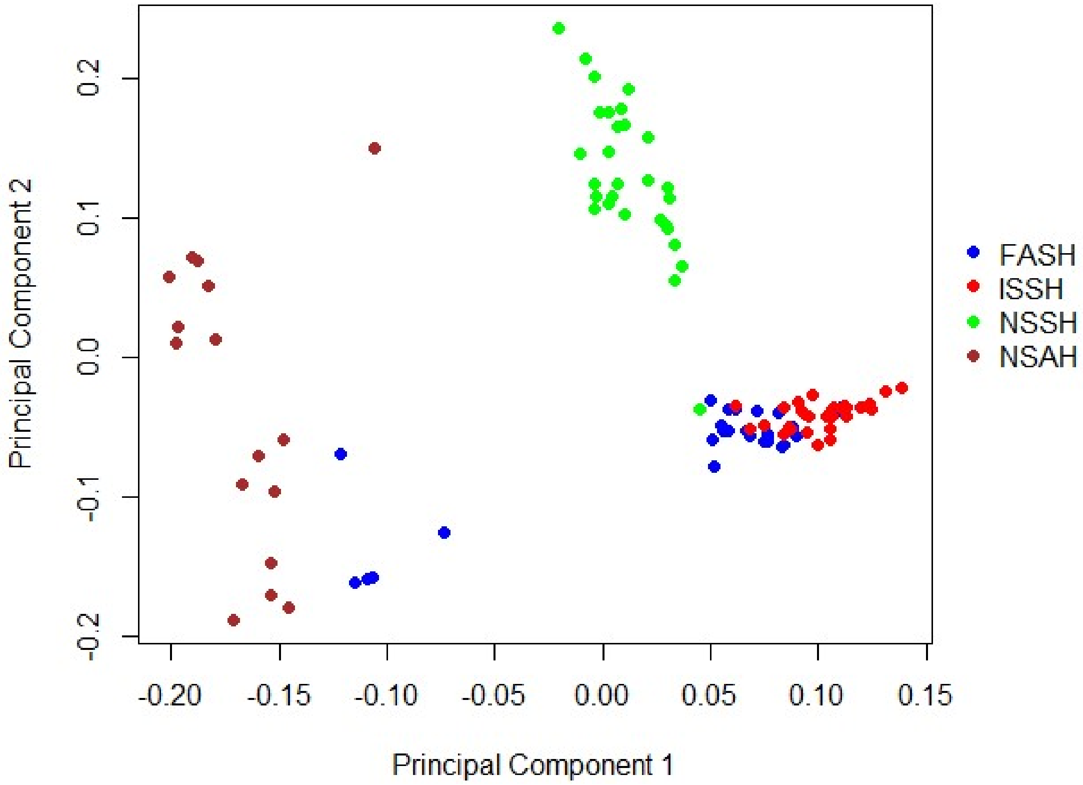


Figure 6. Principal component analysis with genotype likelihoods from the four stocks FASH, ISSH, NSSH, and NSAH.

Assignment

To test whether the 154 SNPs could be used for assigning individuals to a putative population, the R package assignPOP was employed to run 90 tests using Monte-Carlo cross-validation, with 0.5, 0.7, and 0.9 proportions of individuals as training sets to assign the rest of the individuals. The overall assignment accuracy was approximately 90%, with FASH having the poorest accuracy and ISSH assignment being the most accurate (Table 4). As expected, a higher proportion of individuals used in the training set gave higher accuracy, although there were some outliers (Supplementary Figure S10). Using the baseline samples for assignment is not optimal but it indicated that the 154 SNPs could be used to assign individuals to these putative populations.

Table 4. Mean assignment ± standard deviation of Atlantic herring across 90 tests from the Monte-Carlo cross-validation. Based on 154 SNPs. NSSH = Norwegian spring-spawning herring, NSAH = North Sea autumn-spawning herring, FASH = Faroese autumn spawning herring, and ISSH = Icelandic summer-spawning herring.

	FASH	ISSH	NSAH	NSSH
FASH	0.89 ± 0.13	0.04 ± 0.09	0.07 ± 0.10	0.00 ± 0.01
ISSH	0.03 ± 0.07	0.97 ± 0.07	0.00 ± 0.00	0.00 ± 0.00
NSAH	0.02 ± 0.05	0.00 ± 0.00	0.92 ± 0.13	0.07 ± 0.13
NSSH	0.02 ± 0.05	0.02 ± 0.08	0.00 ± 0.00	0.96 ± 0.09

To test the assignment properly, the test samples that were genotyped for 377 SNPs were assigned to a putative population using the assign.X function in the assignPOP package and the baseline samples as the reference for the putative populations. The overall assignment accuracy was 65% but the results were vastly different between the stocks (Figure 7). NSSH and NSAH had very high accuracies (92% and 98%, respectively), whereas ISSH and FASH had poor assignment accuracies (43% and 47%, respectively) (Figure 7). The majority of the ISSH and FASH individuals that were wrongly assigned were from FASH and ISSH, respectively. This indicated that the FASH and ISSH stocks are highly similar and not enough discriminatory power existed to distinguish them. Another assignment test with FASH and

ISSH merged to a single putative population in the reference was conducted. The accuracy improved to 89% overall; the accuracy for the merged FASH_ISSH putative population was 88%, whereas the accuracy for NSAH and NSSH decreased slightly to 89% and 91%, respectively (Figure 8). These results partly confirmed the Monto-Carlo cross-validation assignment results using the baseline samples only; that is, they confirmed that it is possible to assign NSSH and NSAH individuals to the correct putative population with high accuracy using the selected SNPs. However, they also showed that it is not possible to consistently assign FASH and ISSH individuals correctly using these SNPs, yet it is possible to assign them to a single merged putative population (FASH_ISSH) with high accuracy.

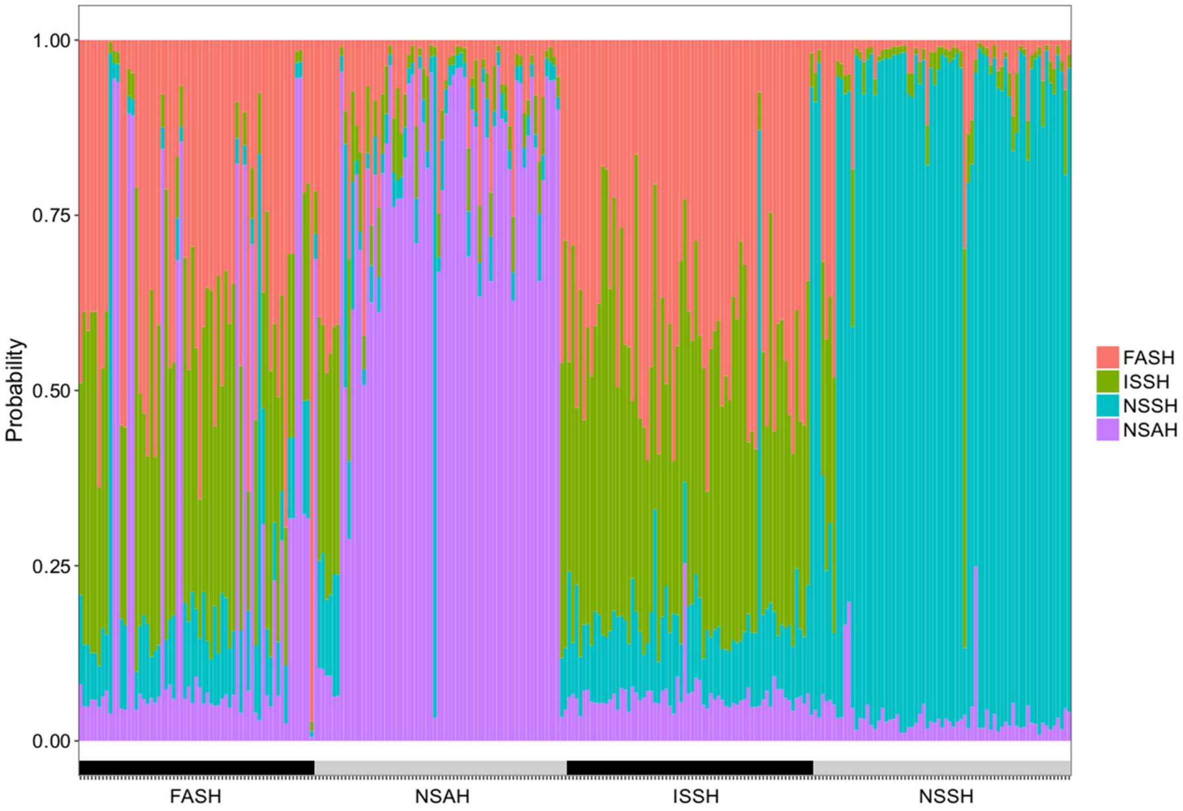


Figure 7. Assignment of the genotyped Atlantic herring individuals using assignPOP and the sequenced individuals as a baseline. The acronyms at the X-axis show which stock the individuals are from and the colours indicate which stock they were assigned to. NSSH = Norwegian spring-spawning herring, NSAH = North Sea autumn-spawning herring, FASH = Faroese autumn spawning herring, and ISSH = Icelandic summer-spawning herring.

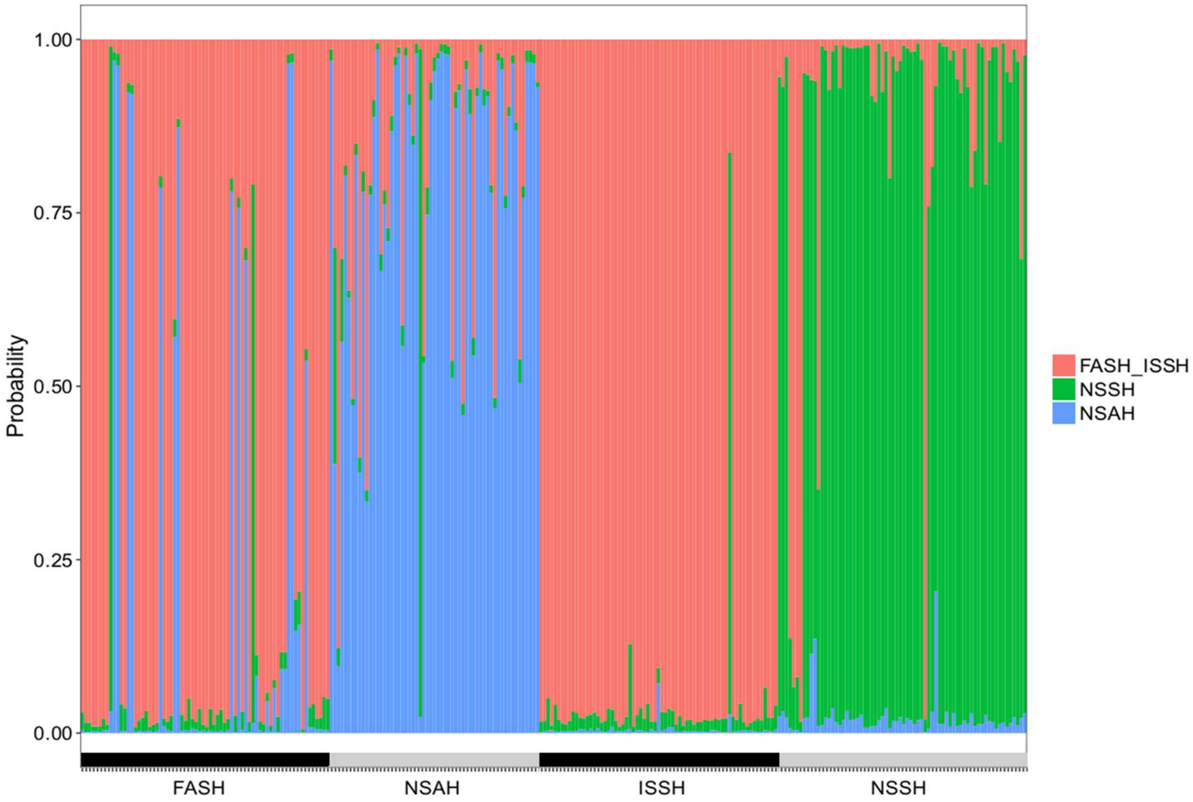


Figure 8. Assignment of the genotyped Atlantic herring individuals using assignPOP and the sequenced individuals as a baseline with FASH and ISSH as one putative population. The acronyms at the X-axis show which stock the individuals are from and the colours indicate which stock they were assigned to. NSSH = Norwegian spring-spawning herring, NSAH = North Sea autumn-spawning herring, and FASH_ISSH = Faroese autumn spawning herring and Icelandic summer-spawning herring merged.

To further investigate the ISSH and FASH results, a Monto-Carlo cross-validation assignment test was performed on the test samples. Based on this analysis, the assignment accuracy for FASH was also poor (21%), whereas it was relatively high for ISSH (86%) (Table 5). Some of the population structure analyses indicated that ISSH and FASH are one panmictic population. If this was the case, one would expect the assignment of FASH and ISSH individuals to be assigned to FASH 50% of the time and to ISSH 50% of the time, just by chance.

Table 5. Mean assignment ± standard deviation of Atlantic herring across 180 tests from the Monte-Carlo cross-validation. Based on 377 genotyped SNPs. NSSH = Norwegian spring-spawning herring, NSAH = North Sea autumn-spawning herring, FASH = Faroese autumn spawning herring, and ISSH = Icelandic summer-spawning herring.

	FASH	ISSH	NSAH	NSSH
FASH	0.21 ± 0.12	0.58 ± 0.14	0.18 ± 0.08	0.03 ± 0.04
ISSH	0.12 ± 0.13	0.86 ± 0.14	0.00 ± 0.00	0.02 ± 0.03
NSAH	0.04 ± 0.06	0.00 ± 0.00	0.94 ± 0.07	0.02 ± 0.03
NSSH	0.01 ± 0.03	0.07 ± 0.07	0.00 ± 0.01	0.92 ± 0.07

3.4.5. Discussion

Our results showed that all four stocks were significantly distinct from each other (Table 3). In the STRUCTURE analysis, K = 3 had the highest likelihood; however, when the population structure was investigated using the Evanno method, the most likely number of K was 2. An investigation of these two clusters indicated substructure in one of them. Once the five suspected migrant herring were removed from the analysis, further substructure was detected. These results indicated hierarchical clustering.

When individual assignment was performed using the baseline samples, it was possible to assign the individuals to the correct putative population with an accuracy of approximately 90%. When assigning the test samples, the samples from the NSSH and NSAH stocks were assigned with high accuracies (92% and 98%, respectively) but the samples from the FASH and ISSH stocks were less successfully assigned (43% and 47%, respectively). However, combining these two stocks into a putative population improved the assignment and raised the overall assignment accuracy to 89%.

Is the FASH stock a true population?

The results regarding the FASH sample are inconsistent. On the one hand, it was shown to be significantly different from the other stocks (Table 3), and the F_{ST} between FASH and ISSH was 0.1352 in the baseline samples (0.0146 using the test samples). These results, together with the assignment of the baseline samples (89% accuracy for FASH), indicated that four distinct populations exist. However, on the other hand, the STRUCTURE and Evanno results showed that the most likely number of K was 3 (K = 2 with substructure in one cluster). NSSH formed one cluster, NSAH a second cluster, and ISSH and FASH formed a third cluster. The assignment accuracy of the test samples was in agreement with these results. Only 47% of the test FASH individuals were correctly assigned using the sequenced individuals as baseline and 21% using Monte-Carlo cross-validation. Furthermore, the NGSadmix results indicated that FASH and ISSH formed a single cluster. Nonetheless, when the five FASH individuals believed to be migrant NSAH were removed from the STRUCTURE analyses, substructure was detected in the ISSH–FASH cluster. These five migrant individuals clustered with NSAH in every analysis and in all the assignment experiments. They were caught in a Faroese fjord, and from their otoliths it was evident that they were autumn-spawners; therefore, they were suspected to be from the FASH stock. Furthermore, NSAH samples have previously been caught in Faroese waters [42, 43]. Thus, we believed enough evidence existed to exclude these five herring from the analyses.

FASH and ISSH evidently showed high levels of similarities. These two stocks are geographically close, and their spawning times could also overlap. ISSH are summer-spawners (July) while FASH are early autumn-spawners (August–September). These are good conditions for gene flow between the two stocks. It is likely that the two stocks represent only one biological population.

Moreover, it is possible that they are two distinct populations that separated too recently for enough differences to have evolved for this study to detect them. In addition to time since separation, the effective population size, rate of gene flow, and hybrid fitness affect how fast two populations diverge. Herring has a large effective population size, and because of the close geographical habitats, high gene flow is expected between these two stocks. Thus, more generations since separation and a larger sample size are required for a detectable difference to have evolved. Overall, we could not confidently conclude whether the FASH and ISSH samples represent one or two populations.

Weaknesses of the study

The sequenced individuals were sequenced at low coverage (average 3.75x), which makes the genotypes inferred from these data more uncertain. For example, sequencing errors could be incorporated as a heterozygous genotype. More worryingly, true heterozygous genotypes could be called as homozygous because by chance only one allele was sequenced. Theoretically, these uncertainties should be distributed equally throughout the whole data set, but the coverage for the different stocks was not equal (Table 1), making the called genotypes less uncertain for the stocks with lower coverage. This could be why a difference existed in the assignment results using data from sequenced and genotyped individuals. It could also be because of differences between the two sample sets. Kukekova *et al.* [44] used low coverage sequencing for a similar study with foxes with good results, but foxes might show stronger genetic differentiation because of factors such as different effective population size and geographical migration patterns to those of herring.

The DNA quality of the genotyped test samples was not high enough for LGC Genomics. Several of our samples are from fisheries catches and had been frozen and thawed twice before the DNA was extracted. This type of sample is, however, exactly what we aim to be able to assign in the future, and therefore we conducted the analyses on these samples despite their low DNA quality.

Another weakness of this study was the samples. Firstly, the sampling size was quite small, especially for the baseline samples. Population studies with other species (*e.g.,* mackerel [45]) have used small samples sizes, but most herring population studies use larger sample sizes [4, 46, 47]. Herring have a large effective population size, making larger samples sizes necessary to detect population structure. The small sample size could be a reason for the poor individual assignment in this study. In addition to the sample size, using nonspawning individuals is not ideal because the assignment to specific stocks is more uncertain. A future study using larger sample sizes and only spawning individuals as a baseline is required to confirm the present study's results, as well as to determine whether FASH is a true population.

Uses and implications in industry and monitoring

The correct assignment of individual fish to a population would be of great use in maintaining sustainable herring fisheries. The SNPs identified in this study could be highly useful for monitoring the herring fisheries' catches and obtaining a realistic picture of the degree of population mixing of these fisheries. Not being able to distinguish between FASH and ISSH is a problem, but the FASH fishery is very small and insignificant. The SNPs from this study could be used to distinguish between populations in large-scale herring fisheries of the other three stocks. Bekkevold *et al.* [22] developed a panel with 156 SNPs (not overlapping with the SNPs in this study) for individual assignment to a geographical region (North Atlantic, North

Sea, and the British Isles, “Transition” [North Sea–Baltic transition area] and Baltic Sea), with an assignment accuracy of 92% with their restricted assignment. Nevertheless, 26% of their samples were unassigned. Our study involved fewer populations, but with our SNP panel we could assign individuals to one of the three stocks rather than a geographic region, with almost the same accuracy and a lower number of unassigned individuals. This makes our panel particularly useful for dealing with mixed herring catches. Combining both these panels would probably result in a powerful tool that could be used to investigate mixed herring catches in most of the North Atlantic Ocean.

When a stock is small and cryptic, such as FASH, being cautious is of great importance to prevent over-exploitation. Therefore, establishing for certain whether FASH is a true population would be desirable, as would being able to distinguish this population from the others.

Furthermore, the SNPs identified in this study could be useful for monitoring the herring populations of the North Atlantic. When conducting the annual herring surveys, using this knowledge together with existing nongenetic methods to assign individuals to a population would improve the accuracy of these data as well as their results and analysis.

3.4.6. References

1. www.hagstovan.fo. 2017. http://statbank.hagstova.fo/pxweb/fo/H2/H2__VV__VV01/fv_heild.px/table/tableViewLayout1/?rxid=fb9148fa-6b94-45c4-92fe-2094d92fe1ed.

2. Whitehead PJ. FAO species catalogue, Vol. 7. Clupeoid fishes of the world. An annotated and illustrated catalogue of the herrings, sardines, pilchards, sprats, anchovies and wolf herrings. Part 1-Chirocentridae, Clupeidae and Pristigasteridae. FAO Fish Synop. 1985;125:303.

3. Hay D, Toresen R, Stephenson R, Thompson M, Claytor R, Funk F, *et al*. Taking stock: an inventory and review of world herring stocks in 2000. In: Funk F, Blackburn J, Hay D, Paul AJ, Stephenson R, Toresen R, *et al*., editors. Herring: Expectations for a new millennium. University of Alaska Sea Grant, Fairbanks; 2001. p. 381-454.

4. Jørgensen HB, Hansen MM, Bekkevold D, Ruzzante DE and Loeschcke V. Marine landscapes and population genetic structure of herring (*Clupea harengus* L.) in the Baltic Sea. Molecular Ecology. 2005;14 10:3219-34.

5. Ruzzante DE, Mariani S, Bekkevold D, André C, Mosegaard H, Clausen LA, *et al*. Biocomplexity in a highly migratory pelagic marine fish, Atlantic herring. Proceedings of the Royal Society B: Biological Sciences. 2006;273 1593:1459-64.

6. Johannessen A, Nøttestad L, Fernö A, Langård L and Skaret G. Two components of Northeast Atlantic herring within the same school during spawning: support for the existence of a metapopulation? ICES Journal of Marine Science. 2009;66 8:1740-8.

7. Husebø Å, Slotte A, Clausen L and Mosegaard H. Mixing of populations or year class twinning in Norwegian spring spawning herring? Marine and Freshwater Research. 2005;56 5:763-72.

8. Jakobsson J, Vilhjálmsson H and Schopka SA. On the biology of the Icelandic herring stocks. Hafrannsóknastofnunin; 1969.

9. Tåning ÅV. Fiskeri- og Havundersøgelser ved Færøerne. Skrifter fra Komm f Danm Fiskeri- og Havundersøgelser. 1943;12:92-4.

10. Joensen J and Taning AV. Marine and freshwater fishes. Vald. Pedersens Bogtrykkeri; 1970.

11. Pampoulie C, Slotte A, Óskarsson GJ, Helyar SJ, Jónsson Á, Ólafsdóttir G, *et al*. Stock structure of Atlantic herring *Clupea harengus* in the Norwegian Sea and adjacent waters. Marine Ecology Progress Series. 2015;522:219-30. doi:10.3354/meps11114.

12. Smith P, Francis R and McVeagh M. Loss of genetic diversity due to fishing pressure. Fisheries Research. 1991;10 3-4:309-16.

13. Libungan LA, Slotte A, Husebø Å, Godiksen JA and Pálsson S. Latitudinal gradient in otolith shape among local populations of Atlantic herring (*Clupea harengus* L.) in Norway. PLoS One. 2015;10 6:e0130847.

14. Hulme T. The use of vertebral counts to discriminate between North Sea herring stocks. ICES Journal of Marine Science. 1995;52 5:775-9.

15. McPherson AA, O'Reilly PT and Taggart CT. Genetic differentiation, temporal stability, and the absence of isolation by distance among Atlantic herring populations. Transactions of the American Fisheries Society. 2004;133 2:434-46.

16. Lamichhaney S, Barrio AM, Rafati N, Sundström G, Rubin C-J, Gilbert ER, *et al*. Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. Proceedings of the National Academy of Sciences. 2012;109 47:19345-50.

17. Limborg MT, Helyar SJ, De Bruyn M, Taylor MI, Nielsen EE, Ogden R, *et al*. Environmental selection on transcriptome ‐ derived SNPs in a high gene flow marine fish, the Atlantic herring (*Clupea harengus*). Molecular Ecology. 2012;21 15:3686-703.

18. Teacher AG, André C, Jonsson PR and Merilä J. Oceanographic connectivity and environmental correlates of genetic structuring in Atlantic herring in the Baltic Sea. Evolutionary Applications. 2013;6 3:549-67.

19. Corander J, Majander KK, Cheng L and Merilä J. High degree of cryptic population differentiation in the Baltic Sea herring *Clupea harengus*. Molecular Ecology. 2013;22 11:2931-40.

20. Bekkevold D, Gross R, Arula T, Helyar SJ and Ojaveer H. Outlier loci detect intraspecific biodiversity amongst spring and autumn spawning herring across local scales. PLoS One. 2016;11 4:e0148499.

21. Skırnisdóttir S, Ólafsdóttir G, Helyar S, Pampoulie C and Óskarsson GJ. A Nordic network for the stock identification and increased value of Northeast Atlantic herring (HerMix). Matıs ohf, Reykjavık, Iceland. 2012.

22. Bekkevold D, Helyar SJ, Limborg MT, Nielsen EE, Hemmer-Hansen J, Clausen LA, *et al*. Gene-associated markers can assign origin in a weakly structured fish, Atlantic herring. ICES Journal of Marine Science. 2015;72 6:1790-801.

23. Shaw P, Turan C, Wright JM, O'connell M and Carvalho G. Microsatellite DNA analysis of population structure in Atlantic herring (*Clupea harengus*), with direct comparison to allozyme and mtDNA RFLP analyses. Heredity. 1999;83 4:490.

24. Postuma K. The nucleus of the herring otolith as a racial character. ICES Journal of Marine Science. 1974;35 2:121-9.

25. Bolger AM, Lohse M and Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30 15:2114-20. doi:10.1093/bioinformatics/btu170.

26. Chen S, Huang T, Zhou Y, Han Y, Xu M and Gu J. AfterQC: automatic filtering, trimming, error removing and quality control for fastq data. BMC Bioinformatics. 2017;18 3:80.

27. Andrews S. FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc, 2010.

28. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:13033997*. 2013.

29. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al*. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25 16:2078-9.

30. Garrison E and Marth G. Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:12073907. 2012.

31. Weir BS and Cockerham CC. Estimating F‐statistics for the analysis of population structure. Evolution. 1984;38 6:1358-70.

32. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, *et al*. The variant call format and VCFtools. Bioinformatics. 2011;27 15:2156-8.

33. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics. 2007;81 3:559-75.

34. Falush D, Stephens M and Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 2003;164 4:1567-87.

35. Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA and Mayrose I. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. Molecular Ecology Resources. 2015;15 5:1179-91.

36. Evanno G, Regnaut S and Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology. 2005;14 8:2611-20.

37. Rousset F. Genepop’007: a complete re‐implementation of the Genepop software for Windows and Linux. Molecular Ecology Resources. 2008;8 1:103-6.

38. Skotte L, Korneliussen TS and Albrechtsen A. Estimating individual admixture proportions from next generation sequencing data. Genetics. 2013;195 3:693-702.

39. Korneliussen TS, Albrechtsen A and Nielsen R. ANGSD: analysis of next generation sequencing data. BMC Bioinformatics. 2014;15 1:356.

40. Meisner J and Albrechtsen A. Inferring population structure and admixture proportions in low-depth NGS data. Genetics. 2018;210 2:719-31.

41. Chen KY, Marschall EA, Sovic MG, Fries AC, Gibbs HL and Ludsin SA. AssignPOP: An R package for population assignment using genetic, non‐genetic, or integrated data in a machine‐learning framework. Methods in Ecology and Evolution. 2018;9 2:439-46.

42. Jacobsen JA. Autumn spawning herring around Faroes during summer 1991. ICES, CM. 1991.

43. Jacobsen JA. A survey of herring south of the Faroes in June 1990. ICES, Doc CM. 1990.

44. Kukekova AV, Johnson JL, Xiang X, Feng S, Liu S, Rando HM, *et al*. Red fox genome assembly identifies genomic regions associated with tame and aggressive behaviours. Nature Ecology & Evolution. 2018;2 9:1479.

45. Rodríguez‐Ezpeleta N, Bradbury IR, Mendibil I, Álvarez P, Cotano U and Irigoien X. Population structure of Atlantic mackerel inferred from RAD‐seq‐derived SNP markers:

effects of sequence clustering parameters and hierarchical SNP selection. Molecular Ecology Resources. 2016;16 4:991-1001.

46. Bonanomi S, Pellissier L, Therkildsen NO, Hedeholm RB, Retzel A, Meldrup D, *et al*. Archived DNA reveals fisheries and climate induced collapse of a major fishery. Scientific Reports. 2015;5:15395.

47. Bekkevold D, Clausen LA, Mariani S, André C, Christensen TB and Mosegaard H. Divergent origins of sympatric herring population components determined using genetic mixture analysis. Marine Ecology Progress Series. 2007;337:187-96.

3.4.7. Author contributions

SíK: Conducted the laboratory work and the analysis and interpretation, as well as wrote the manuscript. SOM: Contributed to the design of the study, writing of the manuscript, and supervised the laboratory work and analysis and interpretation of the data. EíH and JAJ: Contributed to the acquisition and interpretation of the data. TDA: Helped with the design of the study, interpretation of the data, and writing of the manuscript. HG: Helped with the interpretation of the data and writing of the manuscript. PF: Contributed to the design of the study and analysis and interpretation of the data. HAD: Designed the study, acquired funding, and supervised the laboratory work and analysis and interpretation of data. All authors contributed to revising the manuscript and approved the final version.

4. Discussion

Several reasons for assembling a genome exist. Some are basic reasons such as discovering the genome configuration (*i.e.,* the gene content, heterogeneity, repeat content, and other genetic variations). This can provide indications about the evolution of the species and genomes in general. More practical reasons also exist for assembling a genome. The most common is the ability to use the genome assembly to answer biological questions about the species. Here, we wanted to assemble the Atlantic herring genome, as well as answer questions about herring sex determination and population structure in the Northeast Atlantic.

4.1. Herring genome assembly

In this study, a *de novo* herring genome was assembled (A1), then scaffolded with long and linked reads (A2) and finally combined with the previously published draft assembly, resulting in an improved assembly (A3; Manuscript 1). The assembly generated from data from this study alone (A2) was highly similar to the published draft assembly [122]. Thus, we reproduced the herring assembly using different individuals, data types, and assembly software. Thus, our work is in essence a validation of the herring assembly. Reproducibility has become an important topic in science in recent years [124]. Being able to reproduce results is a vital part of the scientific process but has to some degree been underappreciated. Korhonen *et al.* [125] produced a common workflow language (CWL)-based software pipeline for *de novo* genome assembly and suggested publishing assembly pipelines in this manner to achieve the repeatability and reproducibility of assembly results. This would also be an effective method for scientists who are new to the field to learn how to generate assemblies.

In addition, the assemblies from this study and the published draft assembly were compared (Manuscript 1). The comparison was based on fragmentation, correctness, and completeness, and A3 was shown to be the assembly with the best overall quality. As discussed in Manuscript 1, comparing assemblies is a complex task. Determining which assembly has the best quality is not always straightforward, because assemblies can have different strengths and weaknesses. Luckily, user-friendly software packages are available for comparing genome assemblies, such as QUAST (used in Manuscript 1). However, manual comparisons can also be useful, as shown by our manual connexin analysis (Manuscripts 1 and 2).

The process of assembling an organism's genome has evolved together with the development of sequencing technology. The first genome assemblies were generated by cloning fragments of the genome and sequencing them using Sanger sequencing [126]. This was labour-intensive, expensive, and time-consuming, but researchers generally knew which chromosomal segment the various sequencing reads came from, and the sequences were put together in a systematic manner. When NGS sequencing was introduced, the whole genome could be sequenced faster and at lower cost. However, the sequencing reads were short and numerous with no indication of which chromosomal segment they were from. Thus, billions of sequencing reads had to be compared with each other, and handling the large dataset together with resolving repeats in the genome were the new challenges [127, 128]. The relatively recent introduction of TGS has made hybrid (short and long reads) assembly approaches desirable and has resulted in several high-quality assemblies [129-131]. Newer technology has even made it possible to produce high-quality diploid (where both parental alleles from a diploid organism are assembled separately) assemblies using linked reads [132, 133]. This development of different sequencing data types means that several different methods exist of sequencing a genome with the intent to assemble it. The type of sequencing data, as well as their quantity and quality, will result in slightly different assemblies for the same species and individual. This was shown in the present study with the same species but different individuals (Manuscript 1), as well as by Warren *et al.* (and the references therein) using the same individual to produce new versions of the chicken genome assembly [134]. The evolution of assembly software or assemblers has also resulted in several different types of assembly software that use different assembly methods (see subsection 1.6.1). Therefore, even with the same data, using different types of assembly software can result in different outcomes [109-111].

4.1.1. Differences between assembly software

For our *de novo* herring assembly (A1) we used the AllPaths-LG assembler, which is a de Bruijn graph assembler [101]. We had one paired-end and two mate-pair sequencing data sets. Martinez Barrio *et al.* [122] used the SOAPdenovo assembler and libraries of eight different insert sizes for their herring assembly, ranging from 170 bp to 20 kb. The SOAPdenovo assembler is also a de Bruijn graph assembler designed for large genomes [107]. The two assemblies are for the same species but different individuals and are generated from similar data (paired-end and mate-pair) and different assemblers based on the same principal (de Bruijn graphs). When the assemblies were compared using FRCs, it was obvious that they had different strengths and weaknesses (Manuscript 1).

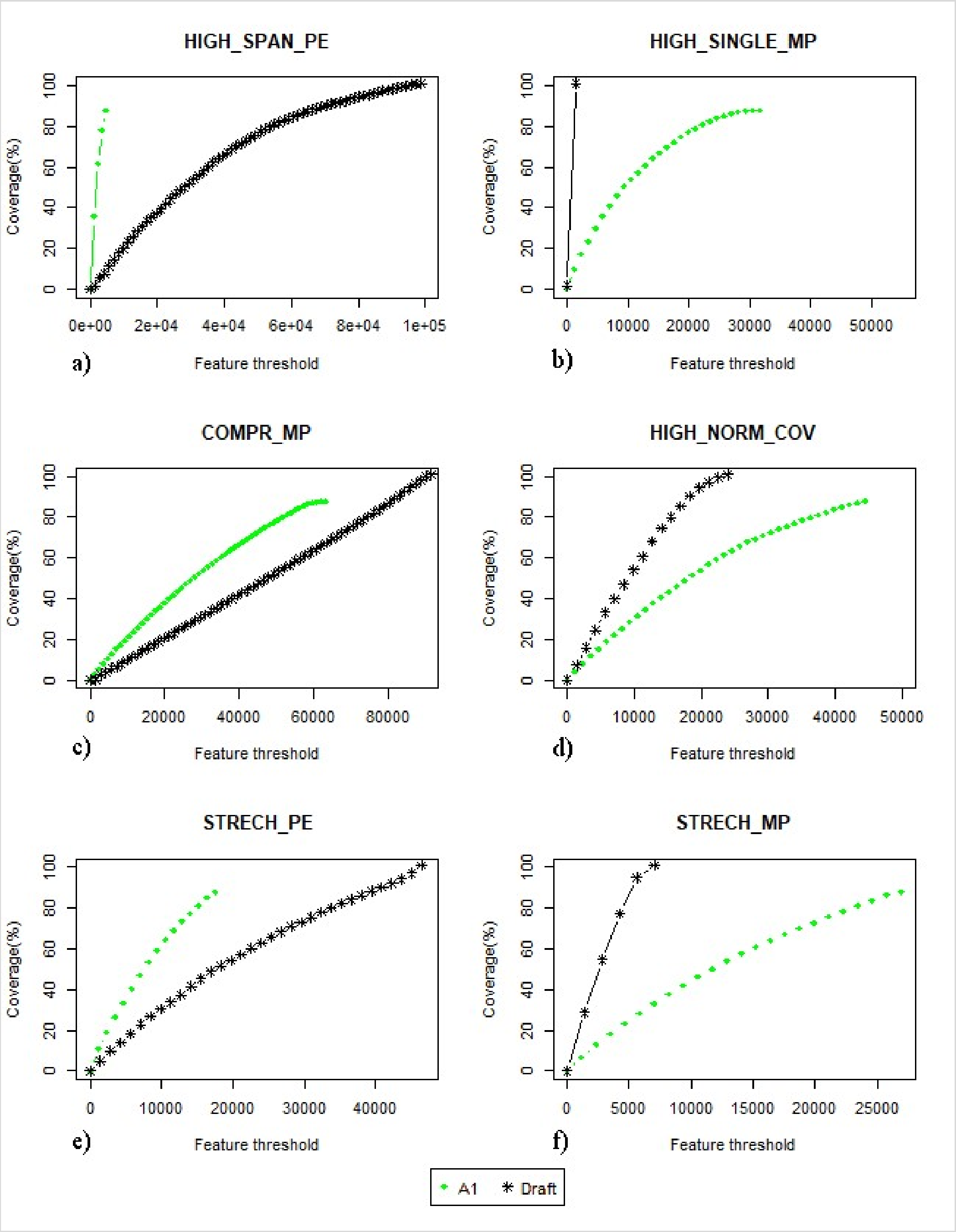


Figure 4.1. FRCs for selected feature types where a difference existed between the A1 and draft assemblies. The FRCs are the same as the FRCs in the supplementary materials for Manuscript 1, but with the other assemblies removed for clarity.

In Figure 4.1, FRCs for a few selected feature types from the Supplementary Information in Manuscript 1 are shown in modified form. Evidently, A1 was superior regarding HIGH_SPAN_PE, COMPR_MP, and STRECH_PE features (Figure 4.1a, c, and e), whereas the draft assembly was superior in terms of HIGH_SINGLE_MP, HIGH_NORM_COV, and STRECH_MP features (Figure 4.1b, d, and f).

HIGH_NORM_COV features describe areas with higher than expected coverage, computed using only reads where both reads in a pair are aligned; HIGH_SINGLE_MP features describe areas where a high number of single reads from mate-pairs align; and HIGH_SPAN_PE features describe areas where a high number of paired-end reads align on different scaffolds [135]. COMP and STRECH features are areas with low and high CE-statistics, respectively. Low CE-statistics indicate compressed sequences, whereas high CE-statistics indicate stretched sequences [135]. The MP or PE states if mate-pair (MP) or paired-end (PE) data were used to calculate the CE-statistics. The A1 assembly had steeper FRCs and lower maximum feature thresholds for both COMPR_MP and STRECH_PE compared with the draft assembly. This indicated that AllPaths-LG was better at resolving repeats than SOAPdenovo. By contrast, the FRCs for the STRECH_MP feature indicated that SOAPdenovo was superior. This contradiction between mate-pair and paired-end data was also true for the majority of feature types (Suppl. File 1, Manuscript 1). This could be because of the inexact insert sizes of our mate-pair libraries (see Methods in Manuscript 1). The libraries were believed to be 4 kb and 7.5 kb when produced in the laboratory, but when investigated bioinformatically (aligned to the A1 assembly), it looked as though they were both roughly 2 kb. For the FRC analysis, 2 kb was used as the mean insert size (max insert size was set to 5 kb). Errors in the calculated insert size could influence the detected numbers of the features. It would have been preferable to produce new mate-pair data with intended insert sizes. Because of the degraded DNA from the sequenced individual, mate-pair libraries with larger insert sizes were not possible. Therefore, another individual would have been required for these libraries, further complicating the assembly.

In addition to the AllPaths-LG assembler, we tried the MaSuRCA assembler (Manuscript 1). MaSuRCA is a hybrid assembler for both short and long reads [136]. The MaSuRCA assembly was generated using the same paired-end, mate-pair, and long-read data as A2. For A2, we also used the linked reads; therefore, to ensure a fair comparison, we will examine the results for the assembly that did not include the linked reads (A1.5). Table 4.1 presents summary statistics for the two assemblies. The MaSuRCA assembly had fewer contigs but a shorter total length and more than seven times as many scaffolds as A1.5. The scaffold N50 was also much shorter than for A1.5. These results show how different assemblers obtain different assemblies even when using the same dataset, in line with previous indications [109-111]. However, having more long reads would probably have given a better assembly using the MaSuRCA assembler.

Table 4.1. Summary statistics for the A1.5 and MaSuRCA assemblies. A1.5 was A1 scaffolded with long reads.

Assembly	No. of Contigs	No. of Scaffolds	Scaffold N50 (kb)	Total length (Mb)
A1.5	117.857	10,354	262	729
MaSuRCA	91.352	74,436	28	588

4.1.2. What is the best assembly approach?

When it comes to genomes, in addition to all the different types of data and assembly software, there are numerous interspecies and intraspecies variations. There are small genomes, such as the 160 kb genome of the *Carsonella ruddii* proteobacteria [137], and there are large genomes, for example, the marbled lung fish (*Protopterus aethiopicus*), which has an estimated genome size of 130 Gigabases (Gb) [138]. Moreover, organisms can have different ploidy (for example, mammals and most animals are diploids), but plants show a variety of ploidy, such as the commercial strawberry (*Fragaria × ananassa*), which has an octoploid genome [139], and bread wheat (*Triticum aestivum*), which has a hexaploid genome [140]. Genomes also have a variety of heterogeneity and repeat content. All these different types of genomes have different bioinformatical requirements for obtaining the optimal assembly. Furthermore, intraspecies variations complicate the assembly process, such as SNPs, microsatellites, copy number variations, insertions, deletions, and inversions.

Using a single individual for a genome assembly is preferable because of the added complication of individual variations. When two assemblies of the same species but different individuals are compared, the assembly differences cannot necessarily be assigned to individual variations or assembly errors (without further laboratory work being required). Therefore, it is also preferable to use the same individual when new sequences are produced to improve an assembly. However, this is not always possible; the sample could be too small or too degraded for repeated experiments or alternative sequencing approaches.

Our final assembly (A3) was produced from four different individuals, because the initial sample had not been stored optimally and the DNA was too degraded to obtain long reads. Therefore, the A3 assembly approaches an average herring genome rather than the precise genome of a single individual fish. This could be interpreted as a weakness if the individual genome was of interest. However, having an average genome is also desirable. The current human reference genome assembly (GRCh38.p13) is a composite genome. Approximately 93% is derived from the sequence of 11 genomic clone libraries (generally considered a proxy for an individual's genome), whereas the remaining 7% represents sequences from more than 50 libraries [141]. In Denmark they have generated a regional reference human genome from 150 individuals, to improve local studies and clinical uses such as precision medicine [142]. There are of course accompanying databases with known differences between individuals. Similar databases with known differences in herring and other species would be desirable. They could, for example, be incorporated into genome browsers. Generating these databases would require much time and resources and might not be considered practical for non-model organisms. Nevertheless, generating a genome assembly for every species was not considered practical only a few years ago but is virtually a reality today. Perhaps in a few years genome assembly, annotation, and accompanying variation databases will be feasible to generate for all species of interest, model or non-model.

So, what is the optimal way to perform a genome assembly? The consensus seems to be that it is to combine long and short read data. Nevertheless, when it comes to the assembly software, trial and error seems to be the method for finding the best assembler. Should there be a standardised way to generate data for a genome assembly? A gold standard would certainly make the process easier, but it would not guarantee good assemblies. For some model organisms, high-quality reference assemblies are available. When new individuals from these species are sequenced, the reference assembly can be used in the assembly process of the new individual (reference-based assembly). In the same manner, a reference assembly from a closely related species can be used for a reference-based assembly. Thus, the availability of reference assemblies also affects the decision on the assembly method. Producing assemblies for the same genome using different methods and ending up with similar results provides more authority to the assembly. Having a gold standard assembly process would increase the repeatability of an assembly but using different methods would probably increase the biological validity if they showed similar results. Furthermore, sequence areas that show obvious differences between assemblies could be subject to closer scrutiny (see the discussion below).

4.2. Annotation

Once a satisfying genome assembly has been produced, annotation of the genome can begin. Gene annotation can be performed experimentally using, for example, cDNA or mRNA sequences from the species and mapping them back to the genome to find where the gene is positioned on the genome [143]. Furthermore, genes can be predicted using algorithms that find sequences on the genome where genes are most likely located [144]. Variations in the signatures of genes make the prediction complex, and it is difficult to make algorithms that cover every genetic possibility. We used BUSCO as an algorithmic approach to investigate genome completion and gene fragmentation, while we investigated connexin genes with a manual approach. Additionally, we conducted a comparative study of the annotated connexin genes in nine teleost species, covering the range of divergence times (Manuscript 2). The results showed that the annotation of connexin genes was not optimal. Some nonconnexin genes were wrongly predicted as being connexins, and some connexins were not predicted at all. Errors existed regarding the start and end position of genes and introns. Moreover, nomenclature was not consistent between the species and did not follow the rules from the nomenclature committees. Inconsistencies in the naming of genes cause problems when comparisons between species are made, because even though the genes have the same name, they might not be orthologues. Comparisons between species should be improved in annotation software to ensure that orthologues have consistent names. We did not investigate any other gene families; however, if the accuracy of gene annotation for other gene families has the same level of incorrectness as the annotation of the connexin genes, then there will be hundreds or even thousands of erroneous gene predictions.

As this study came to an end, a high-quality chromosome level assembly (CLA) of the herring genome was made public (GCA_900700415.1) [145]. This assembly had very few gaps and covered most of the herring genome, and thus had a high level of completeness. However, the picture was somewhat different when looking at the genes in this assembly. The BUSCO analysis in Manuscript 1 showed that there were fewer complete BUSCOs in this assembly compared with A3 and the draft assembly (4036 compared to 4258 and 4348, respectively). Moreover, two connexin genes (*gja9like*-XM_012824682 and *gjb1like*-XM_012819602) that were present in A3 and the draft assembly were missing in the CLA (Manuscript 2). This strongly suggested that A3 and the draft assembly are more complete than the CLA in terms of gene content, despite being more fragmented.

Because the CLA was made available so close to the submission of Manuscript 1, very few comparisons between CLA and the other assemblies were included in Manuscript 1. Therefore, some additional analyses are included here. To search for potential explanations for the missing connexin genes in this high-quality assembly, we aligned the A3 assembly and CLA. Overall, the two assemblies were not that different. There were some missing sequences in A3, a few inversions, and rearrangements (Figure 4.2). We then examined the location of the missing connexin genes more closely. *Gja9like*-XM_012824682 was located on scaffold116:904092-902563 in A3, which aligns to chromosome 11 (LR535867.1) in the CLA. Figure 4.3 presents a dotplot with an alignment between scaffold166 from A3 and the relevant part of chromosome 11 from CLA. They align for most of the scaffold, but a small part does not align very well (*i.e.,* there are some missing sequences and an inversion). The missing *gja9like*-XM_012824682 should have been at this location. This indicates an assembly error in the CLA.

The same was done for *gjb1like*-XM_012819602, which was located on scaffold160:71403-72125 in A3. This scaffold aligns to chromosome 8 (LR535864.1) in the CLA, and the alignment can be seen in Figure 4.4. Here it is also evident that the alignment is poor at the site of *gjb1like*-XM_012819602. Some of the sequence is missing and the remaining sequence is inverted.

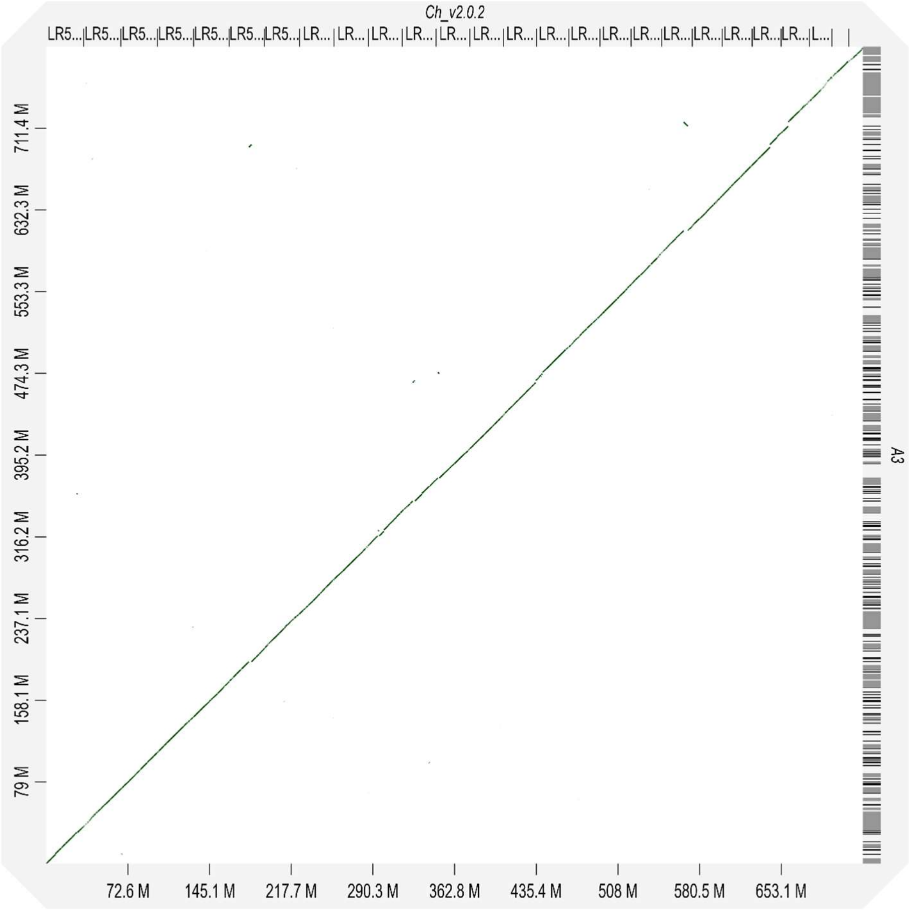


Figure 4.2. Whole genome alignment of A3 and the chromosome level assembly. Alignment and dotplot were generated using D-Genies [146].

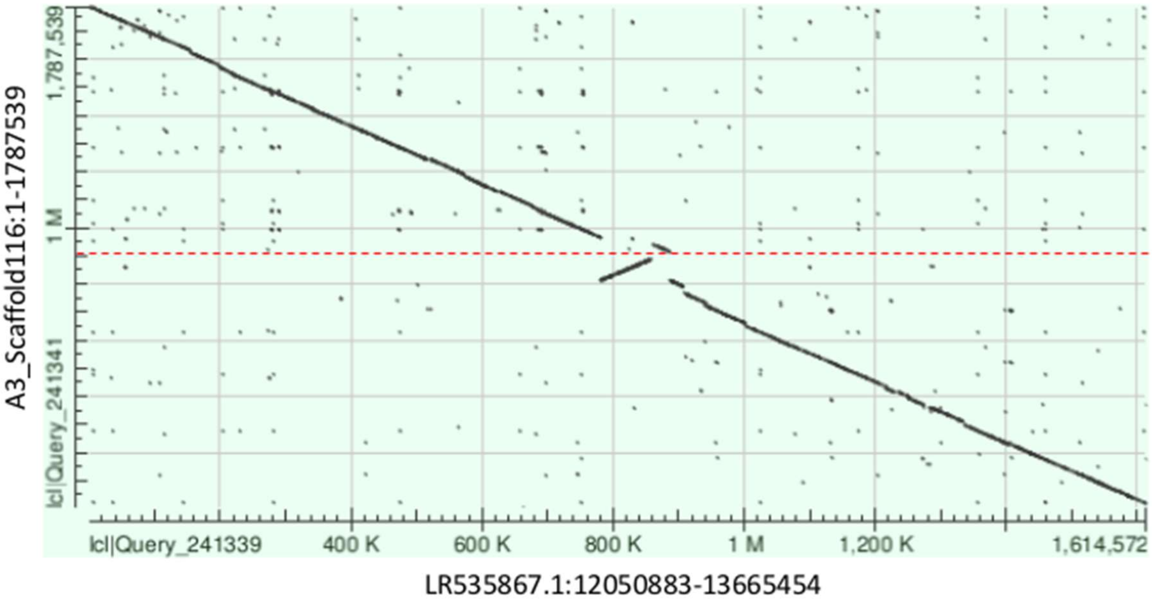


Figure 4.3. Alignments of scaffold166 (A3) and chromosome 11 (LR535867.1) (CLA). The red dotted line indicates the position of *gja9like*-XM_012824682 on A3. Alignment and dotplot were generated using the NCBI online blast tool (available at https://blast.ncbi.nlm.nih.gov/).

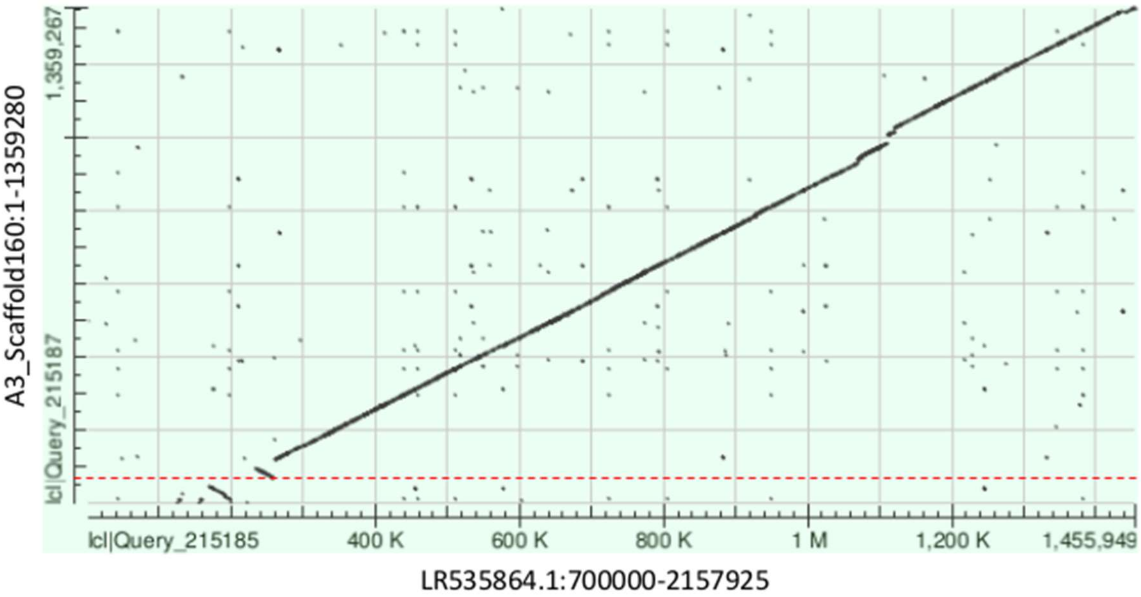


Figure 4.4. Alignment of scaffold160 (A3) and chromosome 8 (LR535864.1) (CLA). The red dotted line indicates the position of *gjb1like*-XM_012819602 on A3. Alignment and dotplot were generated using the NCBI online blast tool (available at https://blast.ncbi.nlm.nih.gov/).

Again, this indicated assembly errors in the CLA. In principle, these differences could be caused by errors in our assembly, but we find it unlikely that these genes are mistakenly present in two independent assemblies (A2 and the draft assembly); furthermore, the presence of these genes was expected as evidenced from other teleosts (Manuscript 2). Thus, these results strongly indicated that even though the CLA is of high quality, assembly errors are still present. Moreover, similar missing connexins caused by misassemblies were indicated in the new chromosome-level cod assembly (GCF_902167405.1) in Manuscript 2. These errors showed that even high-quality assemblies have issues. It is indeed difficult to produce the perfect assembly.

4.3. Answering biological questions

With an assembled and annotated genome, it is possible to investigate interesting biological questions. For example, genetic variation between individuals or species, evolution, or the genetic background for specific biological traits. We identified SNPs by sequencing individual herring and used them to investigate the sex determination in herring as well the population structure in the sequenced samples.

4.3.1. SNP calling

In both this study (Manuscripts 3 and 4) and that of Martinez Barrio *et al.* [122], SNPs were called based on sequencing data from individual fish. We sequenced 103 fish from four stocks from the Northeast Atlantic. Martinez Barrio *et al.* sequenced samples from 20 populations, originating from the Baltic Sea, Skagerrak, Kattegat, North Sea, Atlantic Ocean, and Pacific Ocean. Thus, their populations are from a more diverse range of marine environment, in relation to factors such as salinity and temperature. Approximately 10% of their samples were from the Northeast Atlantic. The SNPs called in this study and by Martinez Barrio *et al.* were investigated to determine whether the same SNPs were called by both studies (Table 4.2). Our study identified more SNPs per individual than Martinez Barrio *et al.* Because we sequenced each individual at a higher coverage; therefore, we most likely called more of the rare SNPs. Only approximately 7% of the SNPs were called by both studies, for both individual and pooled data. When data are pooled at the population/stock level, the coverage for each 'sample' increases and rare individual variations are not called as SNPs. Therefore, fewer SNPs were called when using pooled data.

Table 4.2. Comparison of SNPs called in this study and in that of Martinez Barrio *et al.*

SNPs called from	This study	Martinez Barrio *et al.*	Both studies
Individual data	13.455.776	14.485.088	2.052.190
Pooled data	5.698.051	7.639.919	960.778

4.3.2. Sex determination

We used the herring genome assembly to identify variations between individual herring, and more specifically SNPs (Manuscripts 3 and 4). We performed a GWAS and found SNPs significantly associated with sex in herring (Manuscript 3). The significant SNPs were clustered on six regions (sex regions; SRs) on the herring genome. Females had homozygous genotypes and males had heterozygous genotypes at these SNPs, indicating that herring have a male heterogametic sex determination system. As far as we know, this is the first time a sex determination system for Atlantic herring has been investigated. This novel finding adds to the knowledge of the evolution of sex determination systems in teleosts and fits well with the findings of Pennell *et al*. on the evolution of the various sex determinations systems [147]. The low sequencing coverage did add some uncertainties to the genotypes of the SNPs, as was discussed in Manuscript 3.

Unfortunately, no gene on the different SRs could be singled out as a master sex regulation (MSR) gene. Some genes could potentially play a role in sex determination or development, but closer investigations are required. Sequencing the identified SRs of both sexes could potentially reveal some structural differences between the sexes. Expression analyses of the genes on the SRs could answer whether any of these 20 genes are involved in sex determination. Moreover, it is possible that sex determination is controlled by noncoding genes or genes that have not yet been predicted.

In Manuscript 3, we hypothesised that SR1 and SR4 were on the same chromosome, but the assembly was too fragmented to show this. Once the CLA was made available, the SRs were aligned to this assembly to investigate this matter. Surprisingly, SR1 and SR4 were not on the same chromosome in the CLA. This could suggest that sex determination in herring is more complex than we anticipated. All identified SRs in herring seemed to be segregating together because we found no individual with one heterozygous SR and one homozygous SR. Individual SNPs on the SR could diverge from this, most likely because of errors caused by the low sequencing coverage. Chromosomes segregate independently, and thus by chance individuals must exist with a mixture of heterozygous and homozygous SRs. However, we found none among our 103 individuals. One possible explanation could be that these individuals are not viable. However, that would mean that 50% of the fertilised roe would not be viable, which is biologically improbable. Other explanations could be that these regions were wrongly placed on different chromosomes in the CLA, or that one or more of the SRs were wrongly associated in our GWAS. Further investigations are required to answer this question.

4.3.3. Population structure

In addition to using the herring genome and identified SNPs to investigate herring sex determination, we used them to unravel the population structure of herring in the Northeast Atlantic (Manuscript 4). We investigated if our four stocks represented four genetically distinct populations and found significant differences between all pairwise comparisons of the stocks. In contrast to these findings, cluster analyses indicated only three subpopulations: NSSH, NSAH, and a third with both the FASH and ISSH stocks. Removing suspected migrant herring from the FASH sample resulted in substructure being detected in the analysis of FASH and ISSH using the Evanno method [148]. Despite not being able to distinguish between the FASH and ISSH stocks in this study, evidence indicated that these two stocks do not form an entirely panmictic population. In addition to the aforementioned Evanno results, the fact that ISSH and FASH were significantly different and the F_{ST} between them is 0.135 (Manuscript 4, Table 3) support this. Wright [149] indicated that genetic differentiation as small as 0.05 is not negligible. In the PCA, FASH and ISSH overlapped, but a clear gradient existed with ISSH on one side and FASH on the other (Manuscript 4, Suppl. Figure S9). Furthermore, the mean assignment of FASH and ISSH test individuals across 180 tests from the Monte-Carlo cross-validation was not what would be expected if these individuals were from one panmictic population; that is, roughly a 50:50 assignment to both stocks (Manuscript 4, Table 5). As discussed in Manuscript 4, these two stocks might not have evolved enough differentiation for us to detect because of short time since divergence, high gene flow between them, and a large effective population size. Nonetheless, further studies with larger samples sizes and spawning individuals from these two stocks would be desirable.

In addition, to account for the uncertainty caused by the low sequencing coverage, the population structure was investigated using called genotypes (STRUCTURE) and genotype likelihoods (NGSadmix). Notably, these two analyses provided different results when asked to sort individuals into two clusters (*i.e.,* K = 2) (Manuscript 4, Figures 4 and 5). Using called genotypes, NSSH formed one cluster and the other three stocks formed a second cluster. Using genotype likelihoods, NSAH formed one cluster and the other three stocks formed a second. STRUCTURE is not able to handle a large number of markers; therefore, for this analysis, the SNPs were chosen beforehand based on pairwise F_{ST} between the stocks (see Manuscript 4, Methods section). We ended up with 154 SNPs that had the largest discriminatory power between the stocks, whereas in the NGSadmix analysis, 4.7 million SNPs were used. In the PCA plot (Manuscript 4, Figure 6), the NSAH population could be seen to be genetically diverse. A reason for the difference in K = 2 between the STRUCTURE and NGSadmix analyses could be that the 154 SNPs used in the STRUCTURE analysis were not enough for capturing the genetic diversity within the NSAH.

Lastly, we used population-specific SNPs (panel) to assign individuals to the putative populations (Manuscript 4). Initial results with the reference samples showed good discriminatory power (90% correctly assigned). When we validated the results using new samples and the genotyping method, we were not able to discriminate between the FASH and ISSH stocks. Nevertheless, when these two stocks were combined, the assignment accuracy was 89%. This means that the SNP panel is a potential tool for use in international fisheries, because FASH is a small stock only fished in the Faroe Islands. Our results indicated that NSSH, NSAH, and ISSH are distinct genetic populations, which is in agreement with relevant studies [62, 68, 73].

Limitations of the population study included low sequencing coverage of individuals, a small sample size, and nonideal sampling times of the reference populations. All the samples used represented the reality of fisheries samples, which is what the panel was aimed at. However, this type of samples is not ideal to use for the reference populations.

4.3.3.1 SNP panel as a stock management tool

More work is still required on the panel, especially to validate it and, if possible, reduce the number of SNPs included. We were able to genotype 500 SNPs for 240 individuals with a price of 200 DKK per individual at LGC Genomics. This is the set-up price, which includes the design of primers, which should only be necessary the first time, making subsequent genotyping experiments cheaper. Furthermore, the price per individual reduces as the number of individuals increases. However, if for example a sample of 10 fish from every haul during the fishing season was to be investigated, it would still add up to a large sum of money. Whether the fishing industry would want to invest that much money is questionable, unless it was required by law. If the number of SNPs could be reduced, a routine set up for sampling (to ensure higher-quality DNA), and agreements made with LGC Genomics about the quantity of samples, then the price could be reduced further. However, the turnaround time for sending samples to LGC Genomics, genotyping, and analysing the results is approximately 3 weeks if everything goes according to plan. By that time, the fish will have most likely reached the consumer or even been consumed, and results might not be very useful. Thus, in addition to reducing costs, reducing the analysis time is required to make the panel practical for routine use in the industry.

However, the SNP panel would be a useful tool in stock monitoring. For example, it could be used in the annual herring surveys together with current methods to distinguish between stocks. The results would strengthen estimations of the mixing of stocks as well as validate the panel. Furthermore, the panel could be used by the authorities to investigate mixed fisheries in herring landing. Martinsohn *et al*. [150] showed that the cost of genetic analysis in fisheries and aquaculture forensics was far less than the economic gain (*i.e.,* the fines given), and they encourage routinely use of such genetics techniques in fisheries and aquaculture monitoring. Other interesting uses for the panel would be to test finished herring products (traceability) or historical samples, to further elucidate the herring stock collapse of the 1960s.

5. Conclusions and future perspective

A *de novo* herring genome assembly was produced that could validate the already published herring assembly. To produce a herring genome assembly of even higher quality than the ones available now, we would have to produce more sequencing data. High-quality long reads such as PacBio reads would most likely be the optimal approach, because this type of data is most helpful for resolving repeat regions. For long-read sequencing, the input DNA must be of high quality. In the present study, the DNA quality of the herring samples limited the amount of long and linked reads produced. Thus, another individual would be required. An average herring genome could also be produced by merging all the available assemblies.

Annotation of the connexin gene family in teleost fish was shown to include many errors, which makes comparisons of connexin genes between species problematic. Improving the annotation for the connexin gene family would be desirable, as would reannotations of available genomes to correct the errors in public databases. Similar manual analysis of other gene families would be interesting to determine whether the state of the annotation of other gene families is the same as for the connexin gene family.

In this study, a male heterogametic sex determination system was suggested for herring for the first time. No gene or molecular mechanism could be identified, but several SRs being found indicated that the process might be complex and possibly polygenic. Further studies are required to confirm these SRs. A possibility would be to genotype the SNPs found to be significantly associated with sex in new individuals. Because the two SRs with the highest association with sex were on different chromosomes in the CLA, a new GWAS using SNPs called using the CLA would be appropriate. This could even be done using the individual sequencing data from our study.

In addition, three of the investigated stocks (NSSH, NSAH, and ISSH) were shown to be genetically distinct from each other. Some of the analyses suggested that the fourth population (FASH) is a part of the same population as ISSH, despite them being significantly different (Manuscript 4, Table 3). There were indications that FASH and ISSH are not totally panmictic and perhaps FASH is a subpopulation of the ISSH population. Spawning individuals from ISSH and FASH stocks should be sampled and sequenced at higher coverage to determine with more confidence if these two stocks form one or two genetical populations. The panel developed in this study to assign individual herring to a stock worked well with the three populations NSSH, NSAH, and ISSH. This panel could be highly useful in assigning fisheries samples to stocks, but it should be validated first. An effective method would be to use this panel in the annual herring surveys. Samples collected for these surveys are assigned using current methods and comparing these methods with our panel would give an indication of how effective the panel is.

Overall, the present study has produced more knowledge about herring genetics and evolution. The developed SNP panel could also be useful for keeping herring fisheries sustainable.

References

1. Hagstovan, www.hagstovan.fo. 2017. http://statbank.hagstova.fo/pxweb/fo/H2/H2__VV__VV01/fv_heild.px/table/tableViewLayout1/?rxid=fb9148fa-6b94-45c4-92fe-2094d92fe1ed.

2. California Academy of Sciences, San Francisco, CA, USA, http://researcharchive.calacademy.org/research/ichthyology/catalog/fishcatmain.asp. 2018.

3. Whitehead PJ. FAO species catalogue, Vol. 7. Clupeoid fishes of the world. An annotated and illustrated catalogue of the herrings, sardines, pilchards, sprats, anchovies and wolf herrings. Part 1-Chirocentridae, Clupeidae and Pristigasteridae. FAO Fish Synop. 1985;125:303.

4. Jay E, Bambara R, Padmanabhan R and Wu R. DNA sequence analysis: a general, simple and rapid method for sequencing large oligodeoxyribonucleotide fragments by mapping. Nucleic Acids Research. 1974;1 3:331-54.

5. Gervais H-P. Les poissons: synonymie, description, mœurs, frai, pêche, iconographie des espèces composant plus particulièrement la faune française.: J. Rothschild; 1876.

6. Holst JC, Røttingen I and Melle W. The herring. The Norwegian Sea Ecosystem. 2004:203-26.

7. Blaxter J and Holliday F. The behaviour and physiology of herring and other clupeids. Advances in Marine Biology. Elsevier; 1963. p. 261-394.

8. McQuinn IH. Metapopulations and the Atlantic herring. Reviews in Fish Biology and Fisheries. 1997;7 3:297-329.

9. Reid RN, Cargnelli LM, Griesbach SJ, Packer DB, Johnson DL, Zetlin CA, *et al*. Atlantic herring, *Clupea harengus,* life history and habitat characteristics. NOAA Technical Memorandum NMFS-NE-126: 1-48. 1999:1-48.

10. Geffen AJ. Advances in herring biology: from simple to complex, coping with plasticity and adaptability. ICES Journal of Marine Science. 2009;66 8:1688-95.

11. Bigelow HB and Schroeder WC. Fishes of the Gulf of Maine. US Government Printing Office Washington, DC; 1953.

12. Johansen ACJ. On the Large Spring-spawning Sea Herring -*Clupea Harengus*, L.- in the North-West European Waters.: København; 1919.

13. Jakobsson J. The biological position of the ‘Faeroese Bank’ herring within the Atlanto-Scandian herring stocks. ICES CM. 1970.

14. Dragesund O, Johannessen A and Ulltang Ø. Variation in migration and abundance of Norwegian spring spawning herring (*Clupea harengus* L.). Sarsia North Atlantic Marine Science. 1997;82 2:97-105.

15. Toresen R and Østvedt OJ. Variation in abundance of Norwegian spring-spawning herring *(Clupea harengus*, Clupeidae) throughout the 20th century and the influence of climatic fluctuations. Fish and Fisheries. 2000;1 3:231-56.

16. Husebø Å, Slotte A, Clausen L and Mosegaard H. Mixing of populations or year class twinning in Norwegian spring spawning herring? Marine and Freshwater Research. 2005;56 5:763-72.

17. Jørstad K, Dahle G and Paulsen O. Genetic comparison between Pacific herring (*Clupea pallasi*) and a Norwegian fjord stock of Atlantic herring (*Clupea harengus*). Canadian Journal of Fisheries and Aquatic Sciences. 1994;51 S1:233-9.

18. Jakobsson J, Vilhjálmsson H and Schopka SA. On the biology of the Icelandic herring stocks. Hafrannsóknastofnunin; 1969.

19. Joensen J and Taning AV. Marine and freshwater fishes. Vald. Pedersens Bogtrykkeri; 1970.

20. Jacobsen JA. A survey of herring south of the Faroes in June 1990. ICES, Doc CM. 1990.

21. Jacobsen JA. Autumn spawning herring around Faroes during summer 1991. ICES, CM. 1991.

22. Jørgensen HB, Hansen MM, Bekkevold D, Ruzzante DE and Loeschcke V. Marine landscapes and population genetic structure of herring (*Clupea harengus* L.) in the Baltic Sea. Molecular Ecology. 2005;14 10:3219-34.

23. Ruzzante DE, Mariani S, Bekkevold D, André C, Mosegaard H, Clausen LA, *et al*. Biocomplexity in a highly migratory pelagic marine fish, Atlantic herring. Proceedings of the Royal Society B: Biological Sciences. 2006;273 1593:1459-64.

24. Pampoulie C, Slotte A, Óskarsson GJ, Helyar SJ, Jónsson Á, Ólafsdóttir G, *et al*. Stock structure of Atlantic herring *Clupea harengus* in the Norwegian Sea and adjacent waters. Marine Ecology Progress Series. 2015;522:219-30. doi:10.3354/meps11114.

25. Hempel G. Die Temperaturabhängigkeit der Myomerenzahl beim Hering (*Clupea harengus* L.). Springer; 1953.

26. Hulme T. The use of vertebral counts to discriminate between North Sea herring stocks. ICES Journal of Marine Science. 1995;52 5:775-9.

27. Johannessen A, Nøttestad L, Fernö A, Langård L and Skaret G. Two components of Northeast Atlantic herring within the same school during spawning: support for the existence of a metapopulation? ICES Journal of Marine Science. 2009;66 8:1740-8.

28. Campana SE and Neilson JD. Microstructure of fish otoliths. Canadian Journal of Fisheries and Aquatic Sciences. 1985;42 5:1014-32.

29. Postuma K. The nucleus of the herring otolith as a racial character. ICES Journal of Marine Science. 1974;35 2:121-9.

30. Brophy D and Danilowicz BS. Tracing populations of Atlantic herring *(Clupea harengus* L.) in the Irish and Celtic Seas using otolith microstructure. ICES Journal of Marine Science. 2002;59 6:1305-13.

31. Geffen AJ, Nash RD and Dickey-Collas M. Characterization of herring populations west of the British Isles: an investigation of mixing based on otolith microchemistry. ICES Journal of Marine Science. 2011;68 7:1447-58.

32. Libungan LA, Slotte A, Husebø Å, Godiksen JA and Pálsson S. Latitudinal gradient in otolith shape among local populations of Atlantic herring (*Clupea harengus* L.) in Norway. PLOS ONE. 2015;10 6:e0130847.

33. Feet PØ, Ugland KI and Moksness E. Accuracy of age estimates in spring spawning herring (*Clupea harengus* L.) reared under different prey densities. Fisheries Research. 2002;56 1:59-67.

34. Cardinale M, Doering-Arjes P, Kastowsky M and Mosegaard H. Effects of sex, stock, and environment on the shape of known-age Atlantic cod (*Gadus morhua*) otoliths. Canadian Journal of Fisheries and Aquatic Sciences. 2004;61 2:158-67.

35. Fox CJ, Folkvord A and Geffen AJ. Otolith micro-increment formation in herring *Clupea harengus* larvae in relation to growth rate. Marine Ecology Progress Series. 2003;264:83-94.

36. Høie H, Folkvord A and Johannessen A. Maternal, paternal and temperature effects on otolith size of young herring (*Clupea harengus* L.) larvae. Journal of Experimental Marine Biology and Ecology. 1999;234 2:167-84.

37. Bowers A. Histological changes in the gonad associated with the reproductive cycle of the herring (*Clupea harengus* L.). Mar Res Depart Agric Fish Scotland. 1961;5:1-16.

38. Skırnisdóttir S, Ólafsdóttir G, Helyar S, Pampoulie C and Óskarsson GJ. A Nordic network for the stock identification and increased value of Northeast Atlantic herring (HerMix). Matıs ohf, Reykjavık, Iceland. 2012.

39. Mendel G. Versuche uber pflanzen-hybriden. Verhandlungen des naturforschenden Vereins in Brunn fur. 1866;4:3-47.

40. Olby RC. Origins of Mendelism. Univ of Chicago Pr; 1966.

41. Morgan TH. Sex limited inheritance in Drosophila. Science. 1910;32 812:120-2.

42. Creighton HB and McClintock B. A correlation of cytological and genetical crossing-over in Zea mays. Proceedings of the National Academy of Sciences of the United States of America. 1931;17 8:492.

43. Beadle GW and Tatum EL. Genetic control of biochemical reactions in Neurospora. Proceedings of the National Academy of Sciences of the United States of America. 1941;27 11:499.

44. Avery OT, MacLeod CM and McCarty M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. Journal of Experimental Medicine. 1944;79 2:137-58.

45. McClintock B. Chromosome organization and genic expression. In: *Cold Spring Harbor symposia on quantitative biology* 1951, pp.13-47. Cold Spring Harbor Laboratory Press.

46. Watson JD and Crick FH. Molecular structure of nucleic acids. Nature. 1953;171 4356:737-8.

47. Nirenberg MW and Matthaei JH. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. Proceedings of the National Academy of Sciences. 1961;47 10:1588-602.

48. Wu R. Nucleotide sequence analysis of DNA: I. Partial sequence of the cohesive ends of bacteriophage λ and 186 DNA. Journal of Molecular Biology. 1970;51 3:501-21.

49. Kleppe K, Ohtsuka E, Kleppe R, Molineux I and Khorana H. Studies on polynucleotides: XCVI. Repair replication of short synthetic DNA's as catalyzed by DNA polymerases. Journal of Molecular Biology. 1971;56 2:341-61.

50. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes JC, *et al*. Nucleotide sequence of bacteriophage φX174 DNA. Nature. 1977;265 5596:687.

51. Jeffreys AJ, Wilson V and Thein SL. Individual-specific 'fingerprints' of human DNA. Nature. 1985;316 6023:76.

52. Campbell KH, McWhir J, Ritchie WA and Wilmut I. Sheep cloned by nuclear transfer from a cultured cell line. Nature. 1996;380 6569:64.

53. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, *et al*. The sequence of the human genome. Science. 2001;291 5507:1304-51.

54. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001;409 6822:860-921. doi:10.1038/35057062.

55. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, *et al*. The complete genome of an individual by massively parallel DNA sequencing. Nature. 2008;452 7189:872.

56. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, *et al*. Real-time DNA sequencing from single polymerase molecules. Science. 2009;323 5910:133-8.

57. Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, Stiller M, *et al*. Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature. 2013;499 7456:74.

58. Jang S and Atkins MB. Which drug, and when, for patients with BRAF-mutant melanoma? The Lancet Oncology. 2013;14 2:e60-e9.

59. Cyranoski D and Ledford H. Genome-edited baby claim provokes international outcry. Nature. 2018;563 7733:607-8.

60. Richard G-F, Kerrest A and Dujon B. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol Mol Biol Rev. 2008;72 4:686-727.

61. O'connell M, Dillon M, Wright JM, Bentzen P, Merkouris S and Seeb J. Genetic structuring among Alaskan Pacific herring populations identified using microsatellite variation. Journal of Fish Biology. 1998;53 1:150-63.

62. André C, Larsson LC, Laikre L, Bekkevold D, Brigham J, Carvalho G, *et al*. Detecting population structure in a high gene-flow species, Atlantic herring (*Clupea harengus*): direct, simultaneous evaluation of neutral vs putatively selected loci. Heredity. 2011;106 2:270.

63. Strachan T and Read A. Human Molecular Genetics. 4th ed. New York, USA: Garland Science; 2011.

64. Bekkevold D, Helyar SJ, Limborg MT, Nielsen EE, Hemmer-Hansen J, Clausen LA, *et al*. Gene-associated markers can assign origin in a weakly structured fish, Atlantic herring. ICES Journal of Marine Science. 2015;72 6:1790-801.

65. Lamichhaney S, Fuentes-Pardo AP, Rafati N, Ryman N, McCracken GR, Bourne C, *et al*. Parallel adaptive evolution of geographically distant herring populations on both sides of the North Atlantic Ocean. Proceedings of the National Academy of Sciences. 2017;114 17:E3452-E61.

66. McPherson AA, O'Reilly PT and Taggart CT. Genetic differentiation, temporal stability, and the absence of isolation by distance among Atlantic herring populations. Transactions of the American Fisheries Society. 2004;133 2:434-46.

67. Lamichhaney S, Barrio AM, Rafati N, Sundström G, Rubin C-J, Gilbert ER, *et al*. Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. Proceedings of the National Academy of Sciences. 2012;109 47:19345-50.

68. Limborg MT, Helyar SJ, De Bruyn M, Taylor MI, Nielsen EE, Ogden R, *et al*. Environmental selection on transcriptome-derived SNPs in a high gene flow marine fish, the Atlantic herring (*Clupea harengus*). Molecular Ecology. 2012;21 15:3686-703.

69. Teacher AG, André C, Jonsson PR and Merilä J. Oceanographic connectivity and environmental correlates of genetic structuring in Atlantic herring in the Baltic Sea. Evolutionary Applications. 2013;6 3:549-67.

70. Corander J, Majander KK, Cheng L and Merilä J. High degree of cryptic population differentiation in the Baltic Sea herring *Clupea harengus*. Molecular Ecology. 2013;22 11:2931-40.

71. Bekkevold D, Gross R, Arula T, Helyar SJ and Ojaveer H. Outlier loci detect intraspecific biodiversity amongst spring and autumn spawning herring across local scales. PLOS ONE. 2016;11 4:e0148499.

72. Mariani S, Hutchinson WF, Hatfield EM, Ruzzante DE, Simmonds EJ, Dahlgren TG, *et al*. North Sea herring population structure revealed by microsatellite analysis. Marine Ecology Progress Series. 2005;303:245-57.

73. Shaw P, Turan C, Wright JM, O'connell M and Carvalho G. Microsatellite DNA analysis of population structure in Atlantic herring (*Clupea harengus*), with direct comparison to allozyme and mtDNA RFLP analyses. Heredity. 1999;83 4:490.

74. Lart W. *Fish Stock assessment models and ICES reference points*. 2015.

75. The Faroese Ministry of Fisheries: Loyvisregulering. https://www.fisk.fo/fo/arbeidsoki/fiskivinna/foroyska-fiskiveidiskipanin/loyvisregulering/. Accessed 21-05-2019.

76. Fiskiveiðieftirlitið: http://www.fve.fo. Accessed 21-05-2019.

77. Havstovan: www.hav.fo. Accessed 21-05-2019.

78. European Commisson: https://ec.europa.eu/commission/index_en. Accessed 01-06-2019.

79. ICES. *Herring (Clupea harengus) in subareas 1, 2, and 5, and in divisions 4.a and 14.a, Norwegian spring-spawning herring (the Northeast Atlantic and the Arctic Ocean)*. 2019.

80. ICES. *Mackerel (Scomber scombrus) in subareas 1–8 and 14, and in Division 9.a (the Northeast Atlantic and adjacent waters)*. 2019.

81. ICES. *Blue whiting (Micromesistius poutassou) in subareas 1–9, 12, and 14 (Northeast Atlantic and adjacent waters)*. 2019.

82. Lassen H and Medley P. Virtual population analysis: a practical manual for stock assessment. Food & Agriculture Org.; 2001.

83. Gulland J. Estimation of mortality rates. Annex to Arctic fisheries working group report. International Council for the Exploration of the Sea, CM. 1965.

84. Devlin RH and Nagahama Y. Sex determination and sex differentiation in fish: an overview of genetic, physiological, and environmental influences. Aquaculture. 2002;208 3-4:191-364.

85. Shen Z-G and Wang H-P. Molecular players involved in temperature-dependent sex determination and sex differentiation in Teleost fish. Genetics Selection Evolution. 2014;46 1:26.

86. Ashman T-L, Bachtrog D, Blackmon H, Goldberg EE, Hahn MW, Kirkpatrick M, *et al*. Tree of Sex: A database of sexual systems. Scientific Data. 2014;1:140015.

87. Bachtrog D, Mank JE, Peichel CL, Kirkpatrick M, Otto SP, Ashman T-L, *et al*. Sex determination: why so many ways of doing it? PLOS Biology. 2014;12 7:e1001899.

88. Bull JJ. Evolution of sex determining mechanisms. The Benjamin/Cummings Publishing Company, Inc.; 1983.

89. Brykov VA. Mechanisms of sex determination in fish: evolutionary and practical aspects. Russian Journal of Marine Biology. 2014;40 6:407-17.

90. Brum M. Multiple sex chromosomes in South Atlantic fish, *Brevoortia aurea*, Clupeidae. Brazilian Journal of Genetics. 1992;15 3:547-53.

91. Doucette Jr AJ and Fitzsimons JM. Karyology of elopiform and clupeiform fishes. Copeia. 1988:124-30.

92. Luscombe NM, Greenbaum D and Gerstein M. What is bioinformatics? A proposed definition and overview of the field. Methods of Information in Medicine. 2001;40 04:346-58.

93. Wetterstrand KA: DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) www.genome.gov/sequencingcostsdata (2018). Accessed 06-06-2019.

94. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010;467 7319:1061.

95. Caulfield M, Davies J, Dennys M, Elbahy L, Fowler T, Hill S, *et al*. The 100,000 Genomes Project Protocol v3. Figshare. 2017; doi:https://doi.org/10.6084/m9.figshare.4530893.v4.

96. Sun Y, Huang Y, Li X, Baldwin CC, Zhou Z, Yan Z, *et al*. Fish-T1K (Transcriptomes of 1,000 Fishes) Project: large-scale transcriptome data for fish evolution studies. GigaScience. 2016;5 1:18.

97. Gilbert JA, Jansson JK and Knight R. The Earth Microbiome project: successes and aspirations. BMC Biology. 2014;12 1:69.

98. FarGen: https://www.fargen.fo/. Accessed 21-05-2019.

99. Schatz MC, Witkowski J and McCombie WR. Current challenges in *de novo* plant genome sequencing and assembly. Genome Biology. 2012;13 4:243.

100. Simpson JT and Pop M. The theory and practice of genome sequence assembly. Annual Review of Genomics and Human Genetics. 2015;16:153-72.

101. Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, *et al*. ALLPATHS: *de novo* assembly of whole-genome shotgun microreads. Genome Research. 2008;18 5:810-20. doi:10.1101/gr.7337908.

102. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, *et al*. A whole-genome assembly of Drosophila. Science. 2000;287 5461:2196-204.

103. de la Bastide M and McCombie WR. Assembling genomic DNA sequences with PHRAP. Current Protocols in Bioinformatics. 2007;17 1:11.4. 1-.4. 5.

104. Treangen TJ and Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature Reviews Genetics. 2012;13 1:36.

105. Fox EJ, Reid-Bayliss KS, Emond MJ and Loeb LA. Accuracy of next generation sequencing platforms. Journal of Next Generation Sequencing & Applications. 2014;1.

106. Kelley DR, Schatz MC and Salzberg SL. Quake: Quality-aware detection and correction of sequencing errors. Genome Biology. 2010;11 11:R116.

107. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, *et al*. *De novo* assembly of human genomes with massively parallel short read sequencing. Genome Research. 2010;20 2:265-72.

108. Vezzi F, Narzisi G and Mishra B. Feature-by-Feature – Evaluating *de novo* sequence assembly. PLOS ONE. 2012;7 2:e31002. doi:10.1371/journal.pone.0031002.

109. Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, *et al*. GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Research. 2012;22 3:557-67. doi:10.1101/gr.131383.111.

110. Earl D, Bradnam K, John JS, Darling A, Lin D, Fass J, *et al*. Assemblathon 1: a competitive assessment of *de novo* short read assembly methods. Genome Research. 2011;21 12:2224-41. doi:10.1101/gr.126599.111.

111. Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, *et al*. Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. GigaScience. 2013;2 1:10. doi:10.1186/2047-217X-2-10.

112. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV and Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31 19:3210-2.

113. Mikheenko A, Prjibelski A, Saveliev V, Antipov D and Gurevich A. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics. 2018;34 13:i142-i50. doi:10.1093/bioinformatics/bty266.

114. Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M and Otto TD. REAPR: a universal tool for genome assembly evaluation. Genome Biology. 2013;14 5:R47.

115. Garrison E and Marth G. Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:12073907. 2012.

116. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 2010;20 9:1297-303.

117. Kukekova AV, Johnson JL, Xiang X, Feng S, Liu S, Rando HM, *et al*. Red fox genome assembly identifies genomic regions associated with tame and aggressive behaviours. Nature Ecology & Evolution. 2018;2 9:1479.

118. Ida H, Oka N and Hayashigaki K-i. Karyotypes and cellular DNA contents of three species of the subfamily Clupeinae. Japanese Journal of Ichthyology. 1991;38 3:289-94.

119. Hardie DC and Hebert PD. Genome-size evolution in fishes. Canadian Journal of Fisheries and Aquatic Sciences. 2004;61 9:1636-46.

120. Ohno S, Muramoto J, Klein J and Atkin N. Diploid-tetraploid relationship in clupeoid and salmonoid fish. Chromosomes Today. 1969;2:139-47.

121. Hinegardner R and Rosen DE. Cellular DNA content and the evolution of teleostean fishes. The American Naturalist. 1972;106 951:621-44.

122. Martinez Barrio A, Lamichhaney S, Fan G, Rafati N, Pettersson M, Zhang H, *et al*. The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. eLife. 2016;5:e.12081. doi:10.7554/eLife.12081.

123. Smith P, Francis R and McVeagh M. Loss of genetic diversity due to fishing pressure. Fisheries Research. 1991;10 3-4:309-16.

124. Baker M. 1,500 scientists lift the lid on reproducibility. Nature News. 2016;533 7604:452. doi:10.1038/533452a.

125. Korhonen PK, Hall RS, Young ND and Gasser RB. Common workflow language (CWL)-based software pipeline for *de novo* genome assembly from long- and short-read data. GigaScience. 2019;8 4 doi:10.1093/gigascience/giz014.

126. Consortium TCeS. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science. 1998:2012-8.

127. Eichler EE, Clark RA and She X. An assessment of the sequence gaps: unfinished business in a finished human genome. Nature Reviews Genetics. 2004;5 5:345.

128. Alkan C, Sajjadian S and Eichler EE. Limitations of next-generation genome sequence assembly. Nature Methods. 2011;8 1:61.

129. Ye C, Hill CM, Wu S, Ruan J and Ma Z. DBG2OLC: Efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. Scientific Reports. 2016;6:31900. doi:10.1038/srep31900.

130. Jansen HJ, Liem M, Jong-Raadsen SA, Dufour S, Weltzien F-A, Swinkels W, *et al*. Rapid *de novo* assembly of the European eel genome from nanopore sequencing reads. Scientific Reports. 2017;7 1:7213. doi:10.1038/s41598-017-07650-6.

131. Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, *et al*. High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. Nature Communications. 2018;9 1:541. doi:10.1038/s41467-018-03016-2.

132. Ozerov MY, Ahmad F, Gross R, Pukk L, Kahar S, Kisand V, *et al*. Highly Continuous Genome Assembly of Eurasian Perch (*Perca fluviatilis*) Using Linked-Read Sequencing. G3: Genes, Genomes, Genetics. 2018;8 12:3737-43.

133. Seo J-S, Rhie A, Kim J, Lee S, Sohn M-H, Kim C-U, *et al*. *De novo* assembly and phasing of a Korean human genome. Nature. 2016;538 7624:243.

134. Warren WC, Hillier LW, Tomlinson C, Minx P, Kremitzki M, Graves T, *et al*. A new chicken genome assembly provides insight into avian genome structure. G3: Genes, Genomes, Genetics. 2017;7 1:109-17.

135. Vezzi F, Narzisi G and Mishra B. Reevaluating assembly evaluations with feature response curves: GAGE and assemblathons. PLOS ONE. 2012;7 12:e52210. doi:10.1371/journal.pone.0052210.

136. Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL and Yorke JA. The MaSuRCA genome assembler. Bioinformatics. 2013;29 21:2669-77. doi:10.1093/bioinformatics/btt476.

137. Nakabachi A, Yamashita A, Toh H, Ishikawa H, Dunbar HE, Moran NA, *et al*. The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. Science. 2006;314 5797:267-.

138. Pedersen RA. DNA content, ribosomal gene multiplicity, and cell size in fish. Journal of Experimental Zoology. 1971;177 1:65-78.

139. Edger PP, Poorten TJ, VanBuren R, Hardigan MA, Colle M, McKain MR, *et al*. Origin and evolution of the octoploid strawberry genome. Nature Genetics. 2019;51 3:541.

140. Eckardt NA. Grass Genome Evolution. The Plant Cell. 2008;20 1:3-4. doi:10.1105/tpc.108.058586.

141. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, Kitts PA, *et al*. Evaluation of GRCh38 and *de novo* haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Research. 2017;27 5:849-64.

142. Maretty L, Jensen JM, Petersen B, Sibbesen JA, Liu S, Villesen P, *et al*. Sequencing and *de novo* assembly of 150 genomes from Denmark as a population reference. Nature. 2017;548 7665:87.

143. Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S, *et al*. The Ensembl gene annotation system. Database. 2016;2016.

144. Salzberg SL, Delcher AL, Kasif S and White O. Microbial gene identification using interpolated Markov models. Nucleic Acids Research. 1998;26 2:544-8. doi:10.1093/nar/26.2.544.

145. Pettersson ME, Rochus CM, Han F, Chen J, Hill J, Wallerman O, *et al*. A chromosome-level assembly of the Atlantic herring – detection of a supergene and other signals of selection. bioRxiv. 2019:668384. doi:10.1101/668384.

146. Cabanettes F and Klopp C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ. 2018;6:e4958. doi:10.7717/peerj.4958.

147. Pennell MW, Mank JE and Peichel CL. Transitions in sex determination and sex chromosomes across vertebrate species. Molecular Ecology. 2018;27:3950-63.

148. Evanno G, Regnaut S and Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology. 2005;14 8:2611-20.

149. Wright S. Evolution and the genetics of populations, volume 4: variability within and among natural populations. University of Chicago press; 1984.

150. Martinsohn JT, Raymond P, Knott T, Glover KA, Nielsen EE, Eriksen LB, *et al*. DNA-analysis to monitor fisheries and aquaculture: Too costly? Fish and Fisheries. 2019;20 2:391-401.

Appendix

Supplementary material for Manuscript 1.

Supplementary material

Title

Using long and linked reads to improve an Atlantic herring (*Clupea harengus*) genome assembly

Authors and institutional addresses

Sunnvør í Kongsstovu[1,2,4,*], Svein-Ole Mikalsen[2], Eydna í Homrum[3], Jan Arge Jacobsen[3], Paul Flicek[4], Hans Atli Dahl [1]

[1] Amplexa Genetics A/S, Hoyvíksvegur 51, FO-100 Tórshavn, Faroe Islands.

[2] University of the Faroe Islands, Dept. of Science and Technology, Vestara Bryggja 15, FO-100 Tórshavn, Faroe Islands.

[3] Faroe Marine Research Institute, Nóatún 1, FO-100 Tórshavn, Faroe Islands.

[4] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.

* Corresponding author email: skik@amplexa.com

Supplementary Table S1. The assembler, parameters and data used in the process of finding the optimal assembly (A1). The assembly that gave the best results is indicated with italics. PE indicates paired-end data and MP indicates mate-pair data.

Assembler	**Parameters**	**Data**
AllPaths-LG	Default + ploidy=2	160x PE and 37x MP
AllPaths-LG	Default + ploidy=2	120x PE[1] and 37x MP
AllPaths-LG	Default + haploidify=true + ploidy=2	90x PE[2] and 28x MP
AllPaths-LG	*Default + haploidify=true + ploidy=2*	*63x PE*[2] *and 37x MP*
AllPaths-LG	Default + ploidy=2	63x PE[2] and 37x MP
AllPaths-LG	Default + haploidify=true + ploidy=2	50x PE[3] and 37x MP
SGA	Default + k=41 + OL=75	160x PE and 37x MP
SGA	Default + k=31 + OL=65	160x PE and 37x MP
SGA	Default + k=31 + OL=75	160x PE and 37x MP
SGA	Default + k=31 + OL=85	160x PE and 37x MP
SGA	Default + k=51 + OL=65	160x PE and 37x MP
SGA	Default + k=51 + OL=75	160x PE and 37x MP
SGA	Default + k=51 + OL=85	160x PE and 37x MP
SGA	Default + k=61 + OL=75	160x PE and 37x MP
SGA	Default + k=61 + OL=85	160x PE and 37x MP
SGA	Default + k=71 + OL=30	160x PE and 37x MP
MaSuRCA	Default	160x PE, 37x MP and 2.4x MinION
Supernova	Default + bcfrac=0.5 + maxreads=300M + style=pseudohap	78.5x 10x Genomics
Supernova	Default + bcfrac=0.75 + maxreads=450M + style=pseudohap	78.5x 10x Genomics

[1] Only used the second run of the PE data.

[2] PE data selected on quality to only include 90x and 63x of the highest quality.

[3] PE data selected randomly.

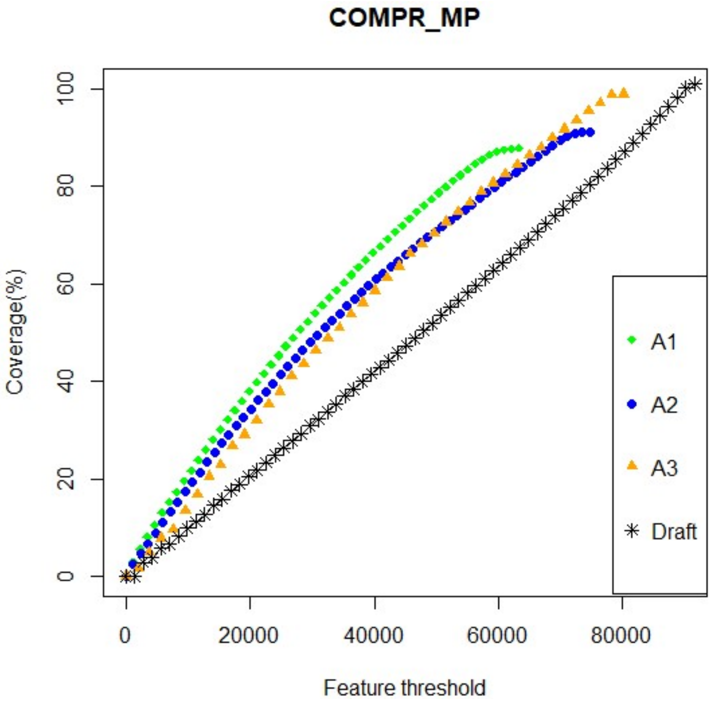
Supplementary Table S2. Herring connexin genes. The predicted genes from the published herring genome assembly are indicated with a Genbank accession number, with the duplicate accession numbers in the comment section.

#	**Abbreviated name**	**Accession number**	**Comments or Accession number of near duplicates (>98% id at nucleotide level)**
1	gja1-cx43	XM_012829211	
2	gja1like	XM_012836783	
3	gja3like	XM_012842347	
4	gja3like	XM_012840585	
5	gja3like	XM_012834366	
6	gja3like	XM_012819598	
7	gja6like	XM_012822071	
8	gja5like	XM_012816449	
9	gja5like	XM_012840593	
10	gja8	XM_012840595	
11	NP-gja8	XM_012816450	Named as histone acetyltransferase KAT6B-like
12	gja9like	XM_012824682	
13	gja9like	XM_012816385	
14	gja10-cx62	XM_012821374	
15	gja10like	XM_012836705	
16	cx32.7like	XM_012829360	
17	cx32.2like	XM_012829221	
18	cx32.2like	XM_012829260	
19	cx32.2like	XM_012828709	
20	gjb1like	XM_012819602	
21	gjb2like	XM_012834339	
22	gjb2like	XM_012842299	
23	gjb2like	XM_012820173	
24	gjb2like	XM_012840586	
25	gjb3like	XM_012822385	100% identical to XM_012822374 100% identical to XM_012822365
26	gjb3like	XM_012818491	100% identical to XM_012818489
27	gjb4like	XM_012822073	
28	gjb4like	XM_012826764	
29	gjb4like	XM_012822396	
30	gjb4like	XM_012818492	99.9% identical to XM_012818490
31	gjb7-cx25	XM_012823856	
32	gjc1-cx45	XM_012816830	
33	gjc1like	XM_012817598	
34	gjc1like	XM_012821065	
35	gjc1like	XM_012836489	
36	gjc2-cx47	XM_012827872	
37	gjd2-cx36	XM_012823340	
38	gjd2	XM_012819299	
39	gjd2like	XM_012828866	

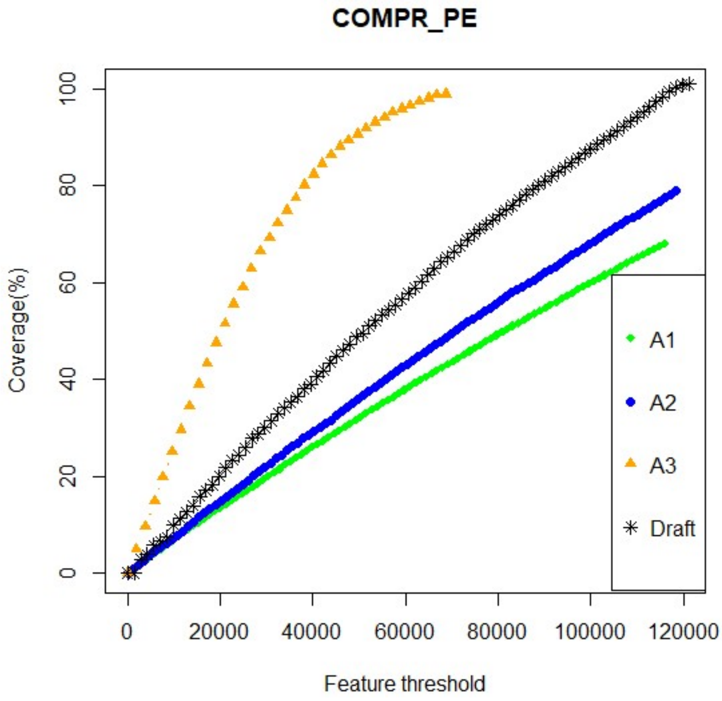
40	gjd2like	XM_012817227	
41	gjd2like	XM_012838313	
42	NP-cx39.2*		Identified by Blast using orthologs* from other teleosts
43	gjd3like	XM_012837668	98.4% identical toXM_012837669
44	gjd3like	XM_012837670	95.1% id to full length XM_012837668
45	gjd4-cx40.1	XM_012823059	
46	gje1like	XM_012822376	

* Sequences will be detailed elsewhere (Mikalsen SO, Tausen M, í Kongsstovu S, submitted).

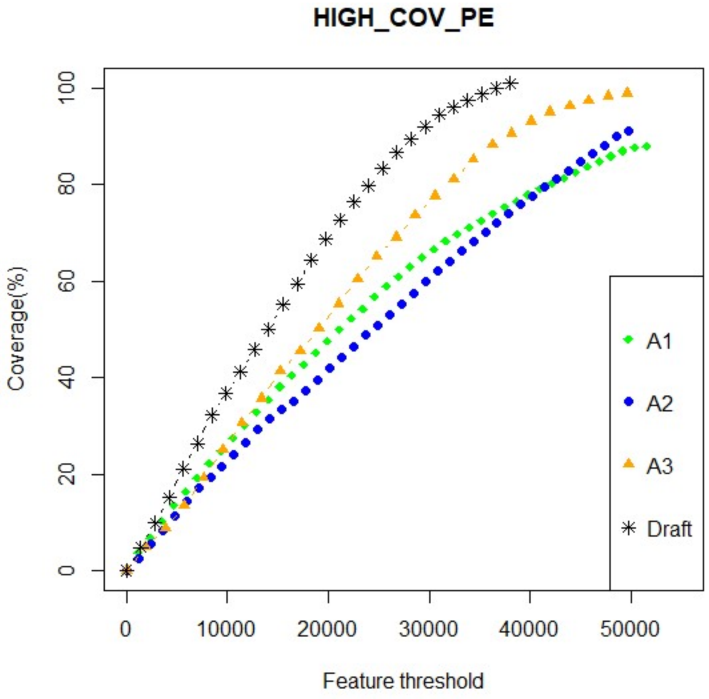
Supplementary Figures



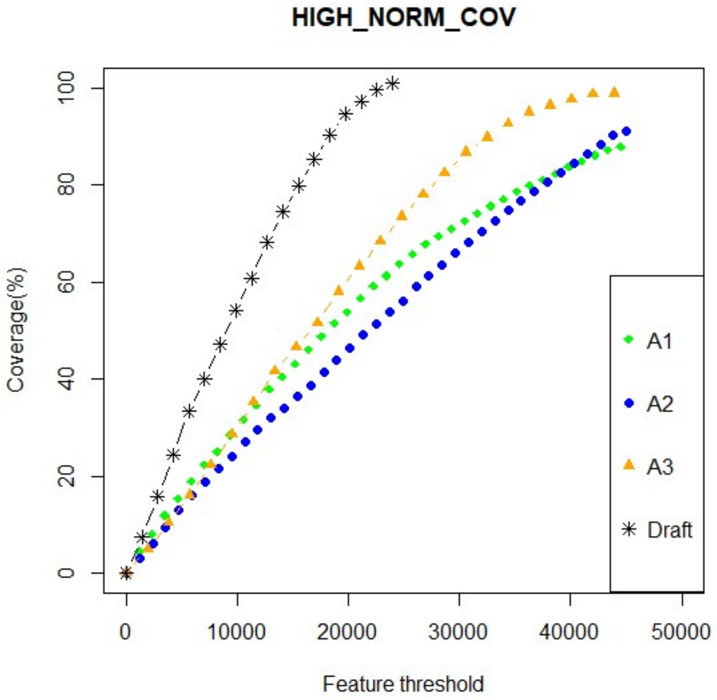
Supplementary Figure S1. Feature response curves (FRCs) for assemblies A1, A2, and A3 and the published draft herring assembly, showing COMPR_MP features. COMPR_MP features describe areas with low CE-statistics; that is, compressed sequences computed with mate-pair data[25]. The FRCs were generated using FRC$^{\text{bam}}$[25] and plotted in R v3.4.3[50].



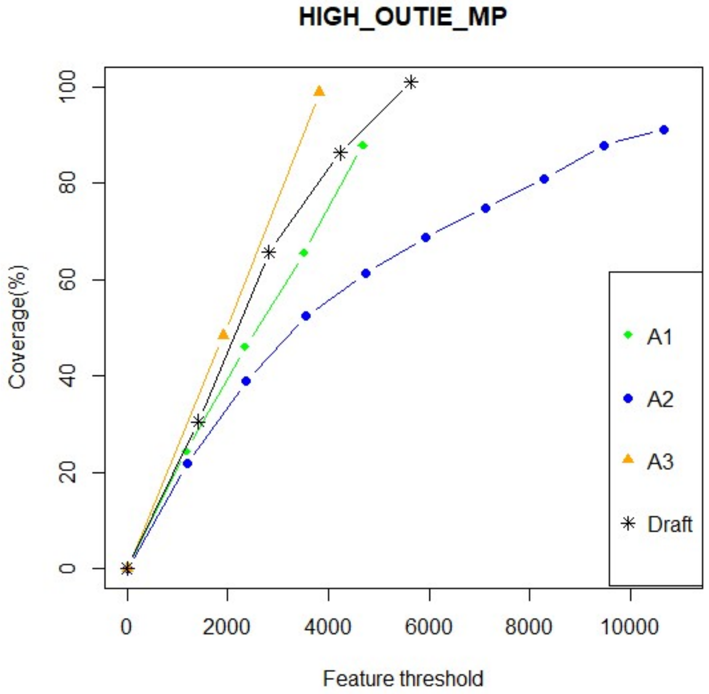
Supplementary Figure S2. Feature response curves (FRCs) for assemblies A1, A2, and A3 and the published draft assembly, showing COMPR_PE features. COMPR_PE features describe areas with low CE-statistics; that is, compressed sequences computed with paired-end data[25]. The FRCs were generated using FRC$^{\text{bam}}$[25] and plotted in R v3.4.3[50].



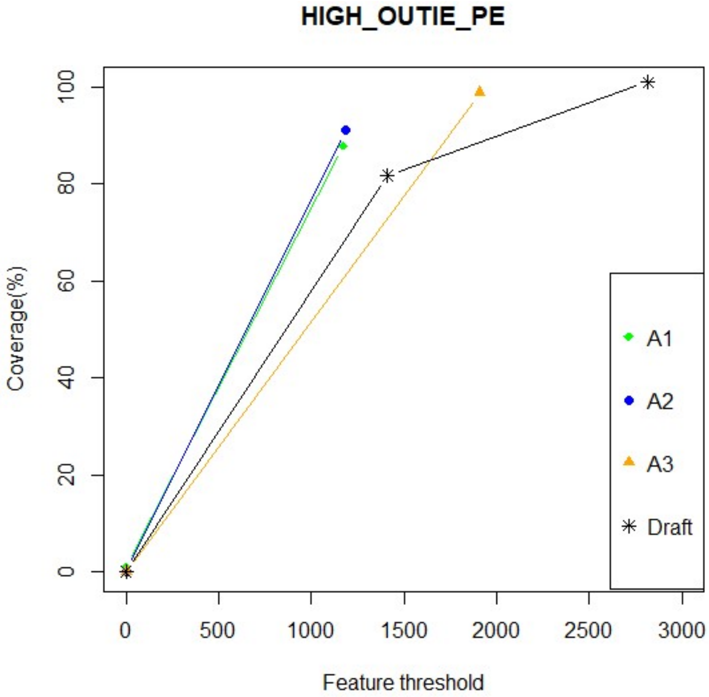
Supplementary Figure S3. Feature response curves (FRCs) for assemblies A1, A2, and A3 and the published draft assembly, showing HIGH_COV_PE features. HIGH_COV_PE features describe areas with high coverage, computed using all aligned reads[25]. The FRCs were generated using FRC[bam25] and plotted in R v3.4.3[50].



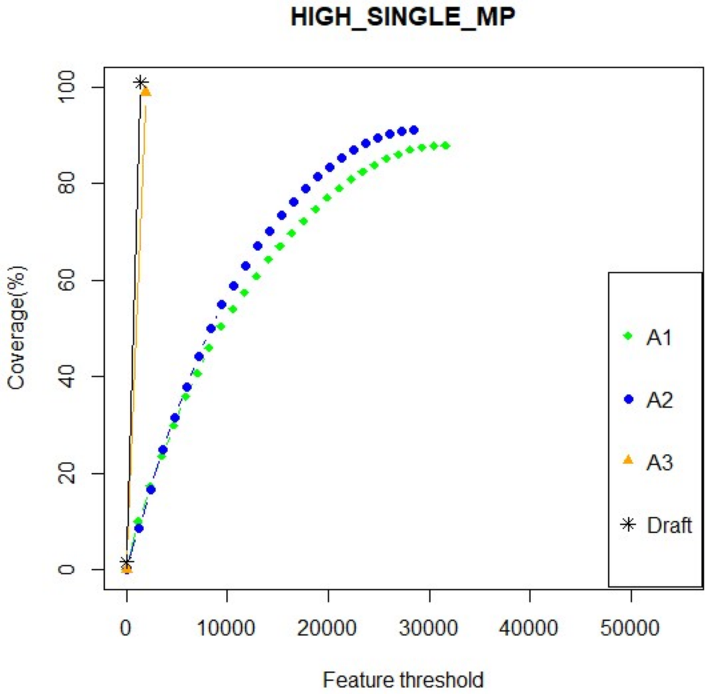
Supplementary Figure S4. Feature response curves (FRCs) for assemblies A1, A2, and A3 and the published draft assembly, showing HIGH_NORM_COV features. HIGH_NORM_COV features describe areas with high coverage, computed using only properly aligned read pairs[25]. The FRCs were generated using FRC[bam25] and plotted in R v3.4.3[50].



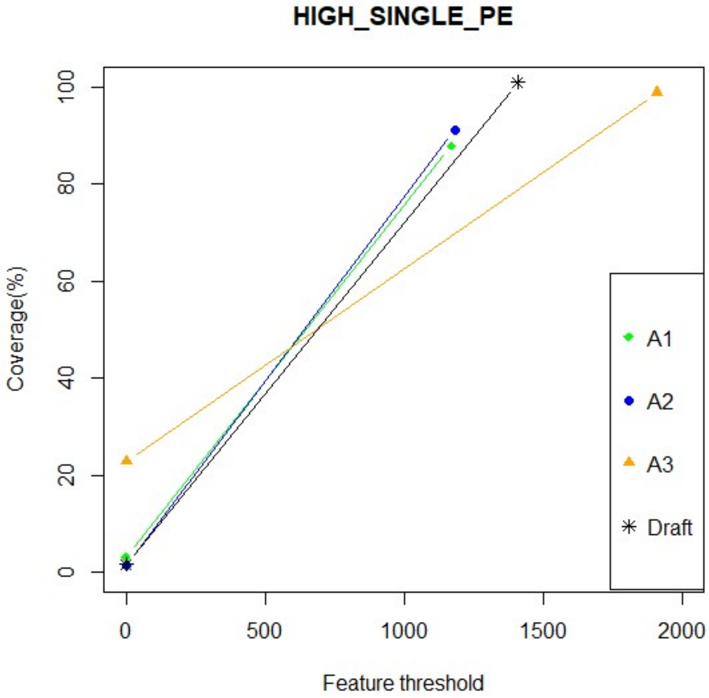
Supplementary Figure S5. Feature response curves (FRC) for assemblies A1, A2, and A3 and the published draft assembly, showing HIGH_OUTIE_MP features. HIGH_OUTIE_MP features describe areas with a high number of misoriented or overly distant mate-pair reads[25]. The FRCs were generated using FRC^{bam}[25] and plotted in R v3.4.3[50].



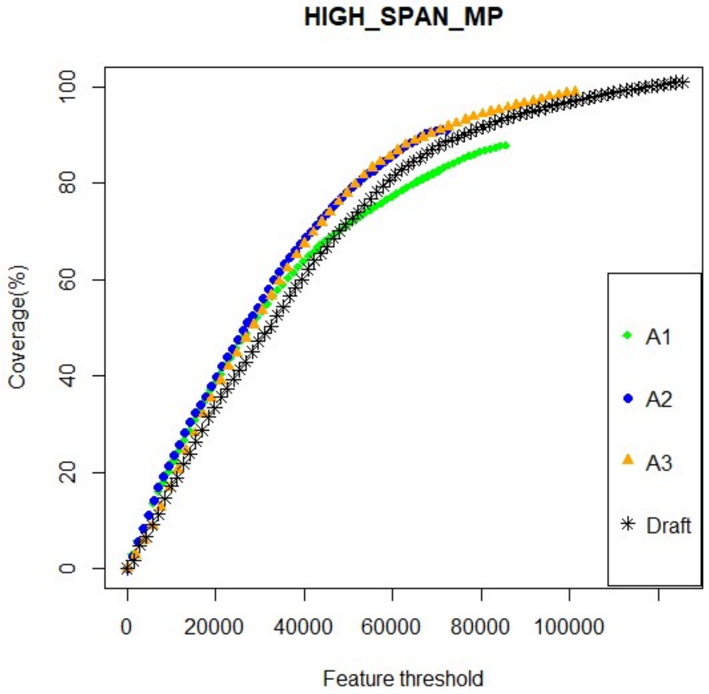
Supplementary Figure S6. Feature response curves (FRCs) for assemblies A1, A2, and A3 and the published draft assembly, showing HIGH_OUTIE_PE features. HIGH_OUTIE_PE features describe areas with a high number of misoriented or overly distant paired-end reads[25]. The FRCs were generated using FRC^{bam}[25] and plotted in R v3.4.3[50].



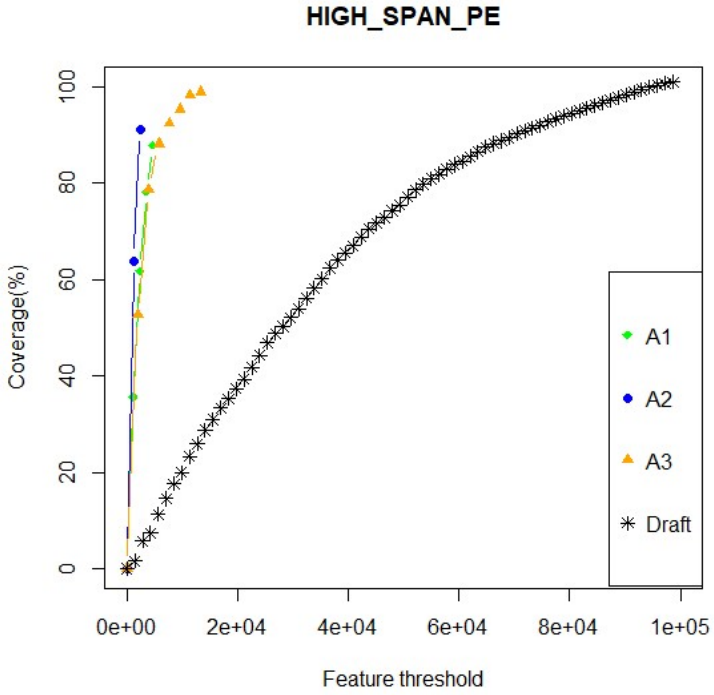
Supplementary Figure S7. Feature response curves (FRC) for assemblies A1, A2, and A3 and the published draft assembly, showing HIGH_SINGLE_MP features. HIGH_SINGLE_MP features describe areas with a high number of mate-pair reads with unmapped pairs[25]. The FRCs were generated using FRC^{bam25} and plotted in R v3.4.3[50].



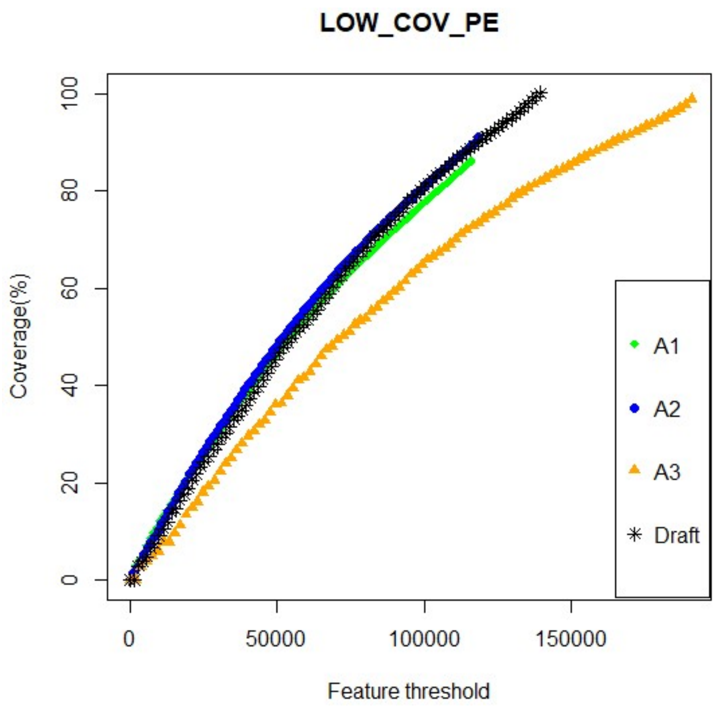
Supplementary Figure S8. Feature response curves (FRCs) for assemblies A1, A2, and A3 and the published draft assembly, showing HIGH_SINGLE_PE features. HIGH_SINGLE_PE features describe areas with a high number of paired-end reads with only one mapped read[25]. The FRCs were generated using FRC^{bam25} and plotted in R v3.4.3[50].



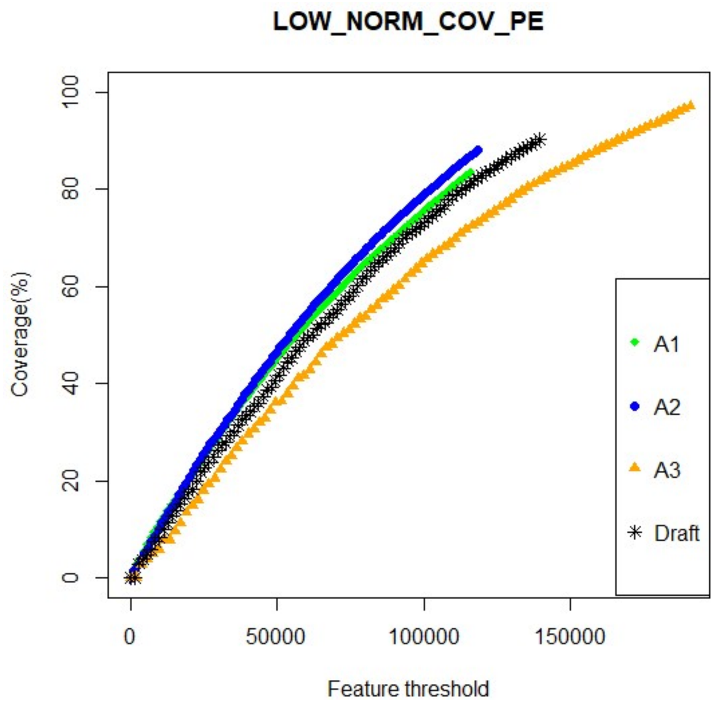
Supplementary Figure S9. Feature response curves (FRCs) for assemblies A1, A2, and A3 and the published draft assembly, showing HIGH_SPAN_MP features. HIGH_SPAN_MP features describe areas with a high number of mate-pairs mapping on different scaffolds[25]. The FRCs were generated using FRC^{bam25} and plotted in R v3.4.3[50].



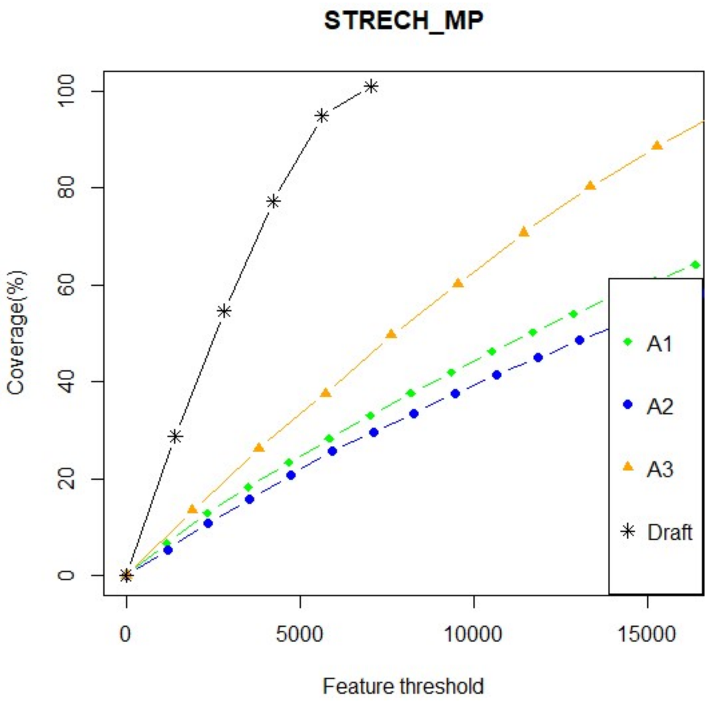
Supplementary Figure S10. Feature response curves (FRCs) for assemblies A1, A2, and A3 and the published draft assembly, showing HIGH_SPAN_PE features. HIGH_SPAN_PE features describe areas with a high number of paired-end reads mapping on different scaffolds[25]. The FRCs were generated using FRC^{bam25} and plotted in R v3.4.3[50].



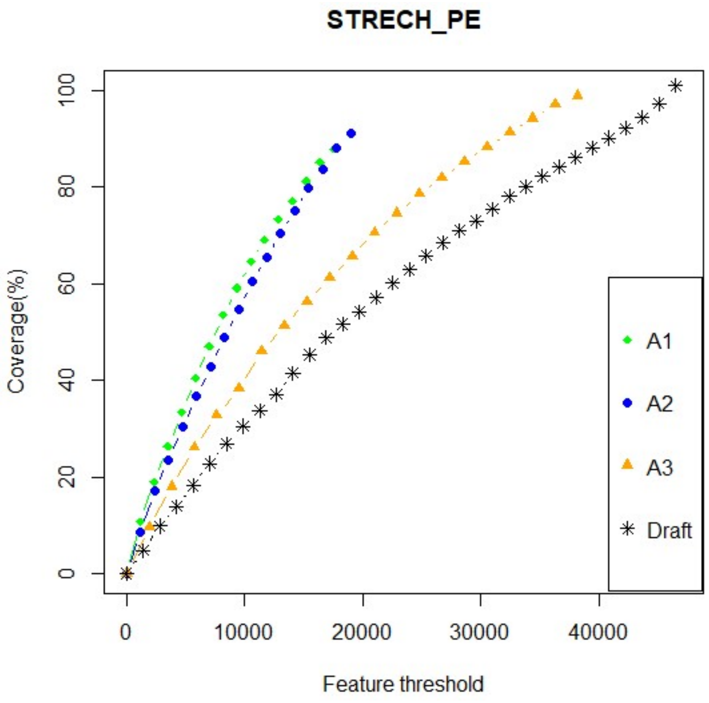
Supplementary Figure S11. Feature response curves (FRCs) for assemblies A1, A2, and A3 and the published draft assembly, showing LOW_COV_PE features. LOW_COV_PE features describe areas with low coverage, computed using all aligned reads[25]. The FRCs were generated using FRC^{bam25} and plotted in R v3.4.3[50].



Supplementary Figure S12. Feature response curves (FRCs) for assemblies A1, A2, and A3 and the published draft assembly, showing LOW_NORM_COV_PE features. LOW_NORM_COV_PE features describe areas with low coverage, computed using only properly aligned pairs[25]. The FRCs were generated using FRC^{bam25} and plotted in R v3.4.3[50].



Supplementary Figure S13. Feature response curves (FRC) for assemblies A1, A2, and A3 and the published draft assembly, showing STRECH_MP features. STRECH_MP features describe areas with high CE-statistics; that is, stretched sequences computed with mate-pair data[25]. The FRCs were generated using FRC^{bam25} and plotted in R v3.4.3[50].



Supplementary Figure S14. Feature response curves (FRCs) for assemblies A1, A2, and A3 and the published draft assembly, showing STRECH_PE features. STRECH_PE features describe areas with high CE-statistics; that is, stretched sequences computed with paired-end data[25]. The FRCs were generated using FRC^{bam25} and plotted in R v3.4.3[50].

Supplementary material for Manuscript 2.

The supplementary material for Manuscript 2 is very long (>150 pages), containing more than 350 connexin gene sequenced with descriptions. Therefore, for brevity, only some of the central parts are included here for illustrative purposes. The included figures and tables are listed as follows.

Parts of **Suppl. Figure 10.** Atlantic cod (*Gadus morhua*) connexins.

Suppl. Fig. 13. Comparisons of human "*GJA4P*" against *connexin39.2* and *GJA4*. **A.** Alignment of conserved domains in human "*GJA4P*" (NG_026166) against *connexin39.2* ("*gjd2like*") in various species at protein level. **B.** Alignment of conserved domains in human "GJA4P" (NG_026166) against GJA4 (connexin37) from human and eel at protein level.

Suppl. Figure 14. Expanded branches from the phylogenetic tree shown in Figure 1. For simplicity, in the title of the figures we often refer to both the mammalian and teleost sequences using the mammalian annotation. To save space, several orthologous groups are shown together in this abstracted version of the Supplement, and the mammalian groups have generally not been expanded, with the exception of Cx39.2 (Suppl. Figure 14G).

Suppl. Table 1. Statistical support for clade grouping.

Suppl. Table 2. Parameter overview for statistical analyses of phylogenetic trees.

Suppl. Table 9. Ohnology among teleost connexins.

>Gm-cx43-G20304 Our modification. Extended in 3’-direction. No reasonable stop codon in frame, but in other reading frames, there are translated sequences that become reasonable similar with other GJA1 orthologs. Hence, potential small intron or sequencing error towards 3’-end.

ATGGGTGACTGGAGTGCTCTGGGCCGCCTGCTGGACAAGGTCCAGGCCTACTCCACCGCTGGGGGAAGGTGTGGCTCTCCGT
CCTCTTCATCTTCAGGATCCTGGTCCTTGGGACGGCCGTGGAGTCCGCCTGGGGCGACGAGCAGTCGGCCTTCAACTGCAACA
CTCAGCAGCCCGGCTGCGAGAACGTATGCTATGACAAATCCTTCCCCATCTCCCATGTGCGCTTCTGGGTGCTGCAGATCATC
TTCGTGTCCACGCCCACGCTGCTGTACCTGGCCCACGTCTTCTACCTGATGAGGAAGGAGCAGAAGCTGAACAGGAAGGAGGA
AATGCTGAAGGCCGTGCAGAACGATGGCGGCGACGTTGACATCCCGCTGAGGAAGATCGAGATGAAGAAGCTGAAGCACGGCC
TGGAGGAGCACGGCAAGGTGAAGATGAAGGGCGCCCTGCTGAGAACCTACATCGTCAGCATCTTCTTCAAGTCCATGTTCGAG
GTGGGCTTCCTGGTCATCCAGTGGTACATATACGGCTTCAGTCTGGCAGCGGTGTACACCTGCGAGAGAGAACCCTGTCCCCA
CAGGGTGGACTGTTTCCTGTCTCGGCCCACAGAGAAGACGGTGTTCATCATCTTCATGCTGGTGGTGTCGCTGGTGTCCCTGC
TGCTCAACGTCATCGAGCTCTTCTACGTGTTCTTCAAGAGGATCAAGGACCGTGTGAAGGGCCGCCAGCCGCCCACCCTCTAC
CCCAGCGCTGGCACCCTGAGCCATACCCCCAAAGATCTTTCCACAGCCAAGTACGCCTACTACAATGGCTGCTCCTCCCCCAC
CGCCCCGCTCTCGCCCATGTCCCCGCCGGGCTACAAGCTGGCCACGGGCGAGCGCGGTACCGGCTCATGTCGCAACTACAACA
AGCAAGCCACCGAGCAGAACTGGACCAACTATTCCACGGAGCAGAACAGCTGGGCCAGCACGGCGCGGGCAGCACTATCTCAA
ACTCCCACGCGCAGGCTTTTGATTCCCCGACGATACGCACGAGCATAAGAAACTGACGTCATCCGCAGCTGCACACGAGATG

>Gm-NN-gja3-G09100-2 Our modification. Splice sites. This Ensembl prediction contains two separate and unique connexins sequences, the present and a cx30.3 sequence.

atgggtgactggagctttctgggacgccttctggagaatgctcaggaacactcaactgtgatcggcaaggtgtggctgaccgt
cctcttcatcttccgcattctggtgctgggcgcggccgcagaggaggtgtggggagacgagcagtcggacttcacctgcaaca
cgcagcagcccggttgcgagaacgtctgctacgaccaggccttccccatctcccacgtgcgcttctgggtgctgcagatcatc
ttcgtgtccacgcccacgctcatctacctgggccacgtgctgcacatcgtgcgcatggaggagaagcggcgtgagaaggagga
ggagctgcggaaggcgggctggcgcagcgaggagctcctcgggcaNNNNGGAGGCGGGAAGAAGGAGAGGCCGCCGATCCGCG
ACGAGCACGGGAAGATCCGCATCCGCGGGGCGCTGCTCCGGACCTACGTCTTCAACATCATCTTCAAGACCCTTCTGGAGGTG
GGCTTCATCCTGGGCCAGTACTCCCTCTACGGCTTCCGCCTCAAGCCGCTGTACAAGTGCGGCCGCTGGCCTTGCCCCAACAC
GGTGGACTGCTTCATCTCCAGGCCCACTGAGAAAACCATCTTCATCATCTTCATGCTGGTGGTGGCCTGCATCTCCCTGCTGC
TCAACCTGCTAGAGATGTACCACCTGGGCTGGAAGAAGGTCAAACACAGCGTCACCCACAAGTTCGCGGCTGACTGCGGGTCC
CTGCGGCTGGGCCCCGGCGACGACGCCGGCGACCCCCGGGCGGTCCCCGAGTGCGCCACCCTGGTTTCGGACCACTGCCTGCA
AGGCTACACCGGCAGGAGCACCATGGAGCGGGTCCGCTACCTGCCCGTCCAGAACTCCTC

>Gm-gja3-G04087 Our modification. Ensembl-predicted introns are included (underlined). There is probably an intron or something wrong in the 3’-end (after the conserved domain), but we have not tried to solve the problem here. In the first conserved domain at the position indicated by lower case “ga”, the Ensembl sequence indicates a row of approx. 100 Ns. “ga” has been found by Blast against GenBank cod wgs.

ATGGGCGACTGGAGCTTTCTGGGCCGGCTTCTTGAGAACGCGCAGGAGCACTCGACGGTGATCGGCAAGGTCTGGCTCACCGT
CCTCTTCATCTTCCGCATCCTAGTGCTGGGTGCCGCAGCAGAGGAGGTGTGGGGCgaCGAGCAGTCGGACTTCACCTGCAACA
CGCAGCAGCCCGGTTGCGAGAACGTCTGCTATGACCAGGCCTTCCCCATCTCCCACATCCGCTTCTGGGTGCTGCAGATCATC
TTTGTGTCCACTCCCACGCTCATCTACCTGGGCCACGTGCTGCACATCGTGCGCATGGAGGAGAAGCGCAAGGAGAAGGAGGA
GGAGCACCGCAAGGTCAGCGGGTTCCCCGATGACAAGGAGCTGCCGTACCGGAACGGGGGCGGCGGTAAAAAGGTGAAGCCGC
CGATCAGAGACGAGCACGGCAAAATCCGCATCCGCGGGGCCTTGCTGCGTACCTACGTGTTCAACATCATCTTCAAGACTCTG
TTTGAGGTGGGCTTCATCCTGGGCCAGTACTTCCTGTACGGCTTCTCGCTGCGGCCGCTCTACAAGTGCTCCCGTTGGCCGTG
CCCCAACACGGTGGACTGCTTTATCTCCAGGCCCACGGAGAAGACTATCTTCATCATATTCATGCTTGTTGTGGCTTGTGTGT
CGCTTTTACTCAACCTGCTGGAGATCTACCACCTGGGCTGGAAGAAGCTGAAGCAGGGCGTGTACCACCCCGACCACCTGCTG
CGGGCCGCCGGCCAGCTGGCCACGCCGGAGGGCGTGGCCTCGCTAGGGGCCCCGGCTCTCCTCAACTACCCCCCCACCTACAG
CCACATAGCGGCCGGCATGGGGTCCCCCACCGACGCCGA**G**TTCAAGATGGAGGAGCTCCAGCGGGAGGAG**GGGGCGCGGA**CGC
CTCCCCCGACTCCCCCGGCCGCCCACTACTACATCAGCAGCAACAACAACCACCGTCTGGCCGCAGAGCAGAACTGGGCCAAC
CTGGCCACCGAGCAGCACACCCGCCAGATGAAGGCCACCTCCCCCACCCCCACGTCCTTCTCCTCCTCAAGCAGTGAAGCGGC
CCCGCCCTGCTCAACTAGCCCCACCCCCTTAATGGCAACCCCGGGCAACGCTGCAGCCCCCGGTGATGTGGCGACCAGCGGCG
ACGGAGCCGGCCTGACCCCCGAGCCGGGCCAGCGGGAGGAAGAGGATGTCACCATGGCGACGGTGGAGATGCACCTGGAGGGG
GTGTTCCCGGACCCCCGGCGTCTTAGCAGAGCCAGTAGAAGCAGCATCCGCGCCCGGCACGATGACCTCGCCATCTGA

Suppl. Fig. 10. Atlantic cod (*Gadus morhua*) connexins. Assembly: gadMor1. Genebuild: Aug 2011. Database version: 96.1. Yellow: Conserved domains as defined by Cruciani and Mikalsen (2007). Green: Conserved cysteine codons (cysteine signature). Grey: 15 nt added at the ends of the conserved domains. Other colors are explained where necessary. Only three examples of cod connexins are included here for brevity. These show the markings that indicate sites of interest and the modifications we have made to the predicted sequences.

Suppl. Figure 13. Comparisons of human '*GJA4P*' against *connexin39.2* and *GJA4*.

```
Hs-GJA4P                MSDWSFLGWLLTRVQNDSTVVGKVWLT??LVLHILLVALLGSAVC?DEHCKFICNTLRPG
Aj-NN-cx39.2            MGDWSILGRFLTEVQNHSTVIGKIWLTMLLIFRILLVTLVGDAVYSDEQSKFTCNTLQPG
Pv-NP-cx39.2            MSDWSFLGRLLTQVQNHSTVVGKVWLTVLLVFRILLVTLVGDAVYGDEQSKFTCNTLQPG
Pa-XM_006925175         MSDWSFLGRLLTQVQNHSTVVGKVWLTVLLVFRILLVTLVGDAVYGDEQSKFTCNTLQPG
Ra-XM_016138748         MSDWSFLGRLLTQVQNHSTVVGKVWLTVLLVFRILLVTMVGDAVYGDEQSKFTCNTLQPG
Wallaby-NP-cx39.2       MGDWSFLGRLLTEVQNHSTVIGKIWLTALLIFRILLVTLVGDAVYRDEQSKFTCNTLQPG
Koala-XM_020963328      MGDWSFLGRLLTEVQNHSTVIGKIWLTALLIFRILLVTLVGNAVYGDEQSKFTCNTLQPG
Md-XM_001376506         MGDWSFLGRLLNEVQNHSTVIGKIWLTALLIFRILLVTLVGDAIYGDEQSKFTCNTLQPG
                        *.***:**.:*. *** ***:**:***  *::.****:::*.*:  **:.** ****.**

Hs-GJA4P                CT???????DHFSHFR?GAFQIVLVAVPSIFFVVCVLH<MVNGnRVLAVCTAHVVLRACM
Aj-NN-cx39.2            CNNVCYDTFAPVSHLRFWVFQIVLVSTPSIFYIVYVLHKIAKDnQVLLIYIVHVVLRSIM
Pv-NP-cx39.2            CTNVCYDRFSPVSHFRFWVFQIVLVATPSIFYVIYVLHQIAREnRVLAIYIAHVVLRAFM
Pa-XM_006925175         CTNVCYDRFSPVSHFRFWVFQIVLVATPSIFYVIYVLHQIAREnRVLAIYIAHVVLRAFM
Ra-XM_016138748         CTNVCYDRFSPVSHFRFWVFQIVLVATPSIFYVIYVLHQIAREnRVLAIYIAHVVLRAFM
Wallaby-NP-cx39.2       CTNVCYNSFAPFSHLRFWIFQIVLVATPSIFYIVCLMHQVALEnRVLIIYIAHVVLRSFL
Koala-XM_020963328      CTNVCYNSFAPISHLRFWIFQIVLVATPSIFYIVCVMHQVALEnRVLVIYIAHVVLRSFL
Md-XM_001376506         CTNVCYNSFAPISHLRFWIFQIVLVATPSIFYIVCVLHQVALEnRALIIYIAHVVLRAFL
                        *.         .**:*   ******:.****::: ::* :.   ..* :  .*****: :

Hs-GJA4P                ELAFLVG???LSGCDMPWLLHCHS?PCPSSPDCFVSRAMRKKIFLNFMC?VGLGCFLLNP
Aj-NN-cx39.2            EIAFLVGQYYLFGFEVPHLFRCETYPCPNRTDCFVSRATEKTIFLNFMFSISLGCFILNI
Pv-NP-cx39.2            ELAFLVGQYYLFGFDVPYLFHCHSYPCPTSTDCFVSRATEKMIFLNFMFGVGVGCFLLNL
Pa-XM_006925175         ELAFLVGQYYLFGFDVPYLFHCHSYPCPTSTDCFVSRATEKMIFLNFMFGVGVGCFLLNL
Ra-XM_016138748         ELAFLVGQYYLFGFDVPYLFHCHSYPCPTSTDCFVSRATEKMIFLNFMFGVGVGCFLLNL
Wallaby-NP-cx39.2       ELGFLVGQYYLFGFDVPHLYRCETYPCPTKTDCFVSRATEKMIFLNFMFGVGLGCFLLSL
Koala-XM_020963328      ELGFLVGQY<LFGFNVPHLYRCETYPCPTKTDCFVSRATEKMIFLNFMFGVGLGCFLLNL
Md-XM_001376506         ELGFLVGQYYLFGFDVPHLYRCETYPCPTKTDCFVSRATEKMIFLNFMFGVGLGCFLLNL
                        *:.****    * * ::* * .* : ***. .*******  * ******  :.:***:*.

Hs-GJA4P                MELCYLGWVFPCQ
Aj-NN-39.2              VELHYLGWVYIFR
Pv-NP-cx39.2            VELHYLGWVFTYR
Pa-XM_006925175         VELHYLGWVFTYR
Ra-XM_016138748         AELHYLGWVFTCR
Wallaby-NP-cx39.2       AELHYLGWLFTFR
Koala-XM_020963328      AELHYLGWLFTFR
Md-XM_001376506         AELHYLGWLFTFR
                         ** ****::  .
```

Suppl. Fig. 13A. Alignment of conserved domains in human "*GJA4P*" (NG_026166) against *connexin39.2* ("*gjd2like*") in various species at protein level.

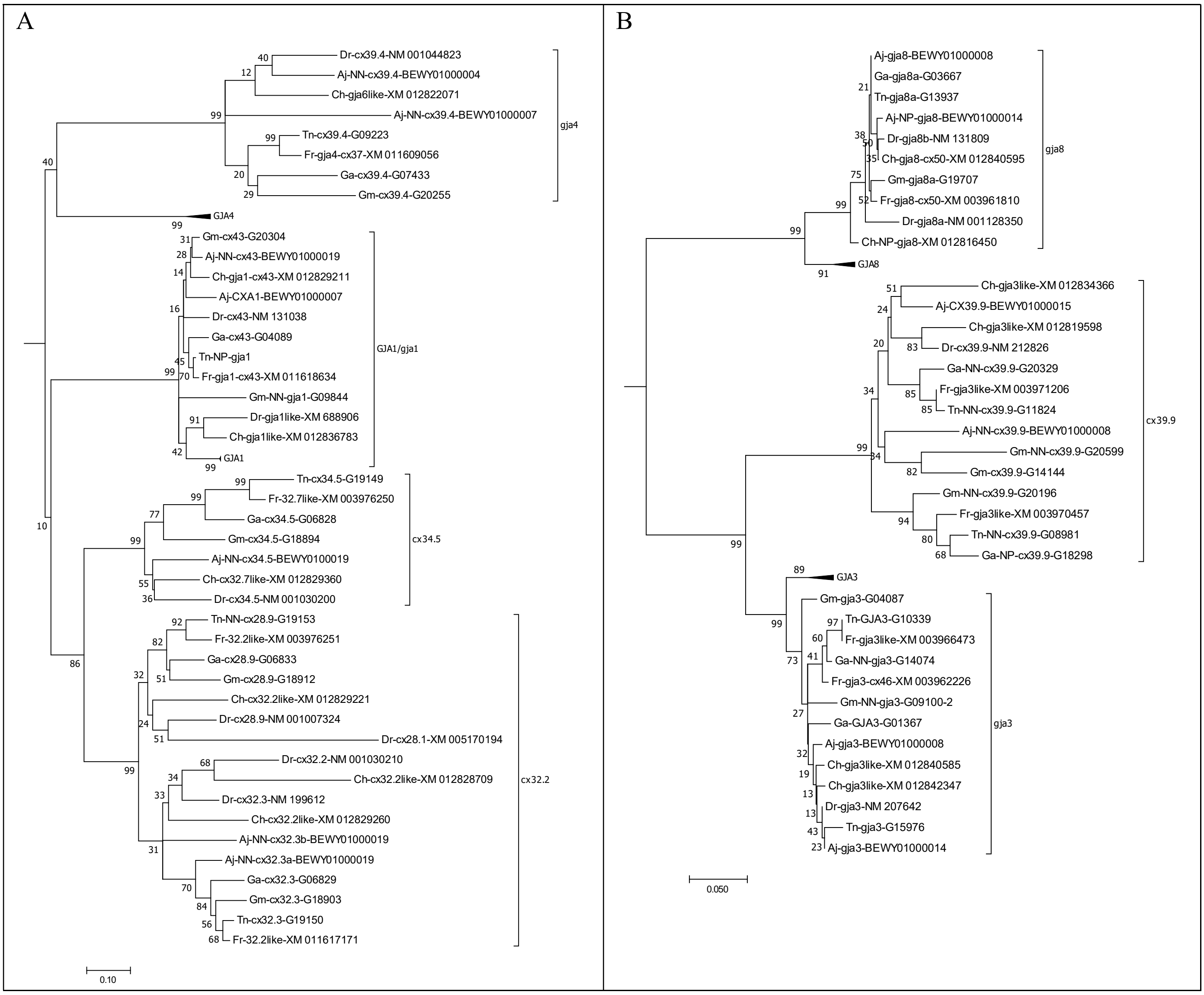
The *cx39.2* sequences given in Suppl. Fig. 12 were translated to protein and aligned. Among the pseudogenes, only the human sequence is included, as aligning several pseudogenes strongly decreases the total number of identities (*) or similarities (: or .). Also the corresponding sequence from eel (Aj-NN-cx39.2) was included. ?, corresponding to a codon that contains one or more n. <, corresponds to a stop codon. n, the first conserved domain is N-terminal to n, and the second conserved domain is C-terminal to n. The Muscle (https://www.ebi.ac.uk/Tools/msa/muscle/) identity matrix is found in Suppl. Table 7.

```
Hs-GJA4P            --MSDWSFLGWLLTRVQNDSTVVGKVWLT??LVLHILLVALLGSAVC?DEHCKFICNTLR
Aj-NN-cx39.2        --MGDWSILGRFLTEVQNHSTVIGKIWLTMLLIFRILLVTLVGDAVYSDEQSKFTCNTLQ
Hs-GJA4-Cx37        --MGDWGFLEKLLDQVQEHSTVVGKIWLTVLFIFRILILGLAGESVWGDEQSDFECNTAQ
Aj-NN-cx39.4-1      MSKSDWTFLELLLEQGQVHSTGVGKMWLTVLFLFRVLVLSTAAESVWGDEQSDFVCNTQQ
Aj-NN-cx39.4-2      MSRADWGFLERFLEEGQEYSTGIGRVWLTVLFLFRMLILGTAAESAWDDEQSDFVCNTQQ
                       .** :*  :*   *  ** :*.:***  :::.:*::   ..:.  **:..* *** .

Hs-GJA4P            PGCT???????DHFSHFR?GAFQIVLVAVPSIFFVVCVLH<MVNGnRVLAVCTAHVVLRA
Aj-NN-cx39.2        PGCNNVCYDTFAPVSHLRFWVFQIVLVSTPSIFYIVYVLHKIAKDnQVLLIYIVHVVLRS
Hs-GJA4-Cx37        PGCTNVCYDQAFPISHIRYWVLQFLFVSTPTLVYLGHVIYLSRREnALMGTYVASVLCKS
Aj-NN-cx39.4-1      PGCEAVCYDKAFPISHFRFFILQVIIVASPAIFYLSYAALHARWQnKLLRVYLCVTVLKL
Aj-NN-cx39.4-2      PGCELACYDRAFPISHFRFFVLQVIFVSTPTIFYFIYVALRMGWEnKLLCAYTLSIVLKV
                    ***          .**:*   :*.::*: *::.:.  .     .  ::       : .

Hs-GJA4P            CMELAFLVG???LSGCDMPWLLHCHS?PCPSSPDCFVSRAMRKKIFLNFMC?VGLGCFLL
Aj-NN-cx39.2        IMEIAFLVGQYYLFGFEVPHLFRCETYPCPNRTDCFVSRATEKTIFLNFMFSISLGCFIL
Hs-GJA4-Cx37        VLEAGFLYGQWRLYGWTMEPVFVCQRAPCPYLVDCFVSRPTEKTIFIIFMLVVGLISLVL
Aj-NN-cx39.4-1      LLEAAFILVLWHLYGFTVPARYVCQRWPCPHTVDCFVSRPKEKTVFTVYMQAMAGVSLLF
Aj-NN-cx39.4-2      LLEAGFILGLWFLYGFVVHAKYVCQRPPCPHTVDCFVSRPTEKTIFTVYMQAIAGVSMLL
                     :* .*:     * *  :     *   ***   ******.  *.:*  :*  :.  .:::

Hs-GJA4P            NPMELCYLGWVFPCQ
Aj-NN-cx39.2        NIVELHYLGWVYIFR
Hs-GJA4-Cx37        NLLELVHLLCRCLSR
Aj-NN-cx39.4-1      NLLEVCVLLRRYCCP
Aj-NN-cx39.4-2      NVVEFLYLAQHTVTH
                    * :*.  *
```

Suppl. Figure 13B. Alignment of conserved domains in human 'GJA4P' (NG_026166) against GJA4 (connexin37) from human and eel at the protein level.

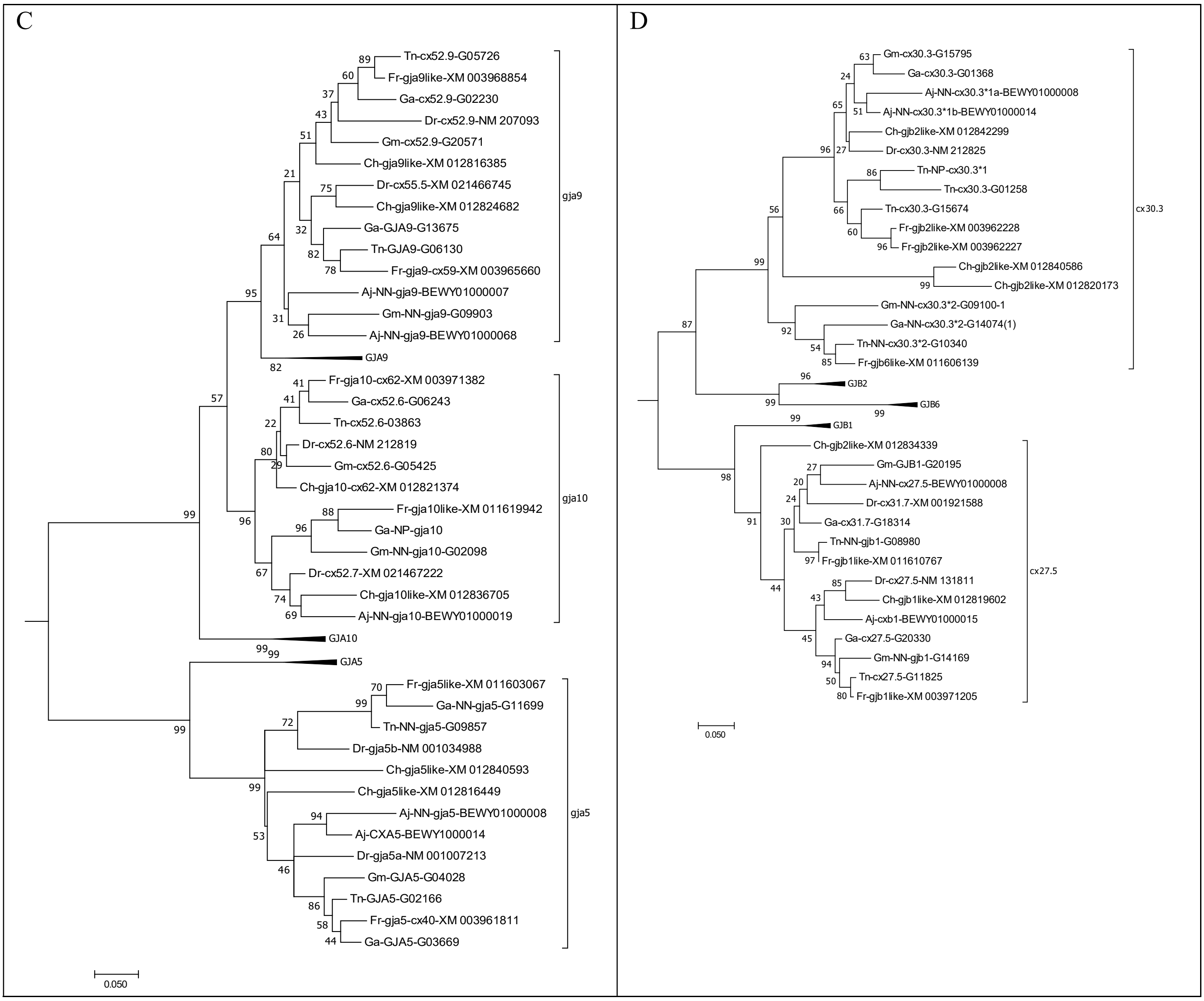
The human *GJA4P cx39.2* sequence given in Suppl. Fig. 12 were translated to protein and aligned with eel *cx39.2*, human *GJA4*, and the two eel-*gja4* (*cx39.4*) sequences. Identities (*) or similarities (: or .) are indicated below the alignment. ?, corresponding to a codon that contains one or more n. <, corresponds to a stop codon. n, the first conserved domain is N-terminal to n, and the second conserved domain is C-terminal to n. The Muscle (https://www.ebi.ac.uk/Tools/msa/muscle/) identity matrix is shown in Suppl. Table. 8.

Suppl. Figure 14. Expanded branches from the phylogenetic tree shown in Figure 1. For simplicity, in the title of the figures we often refer to both the mammalian and teleost sequences using the mammalian annotation. To save space, several orthologous groups are shown together in this abstracted version of the Supplement, and the mammalian groups have generally not been expanded, with the exception of Cx39.2 (Suppl. Figure 14G).



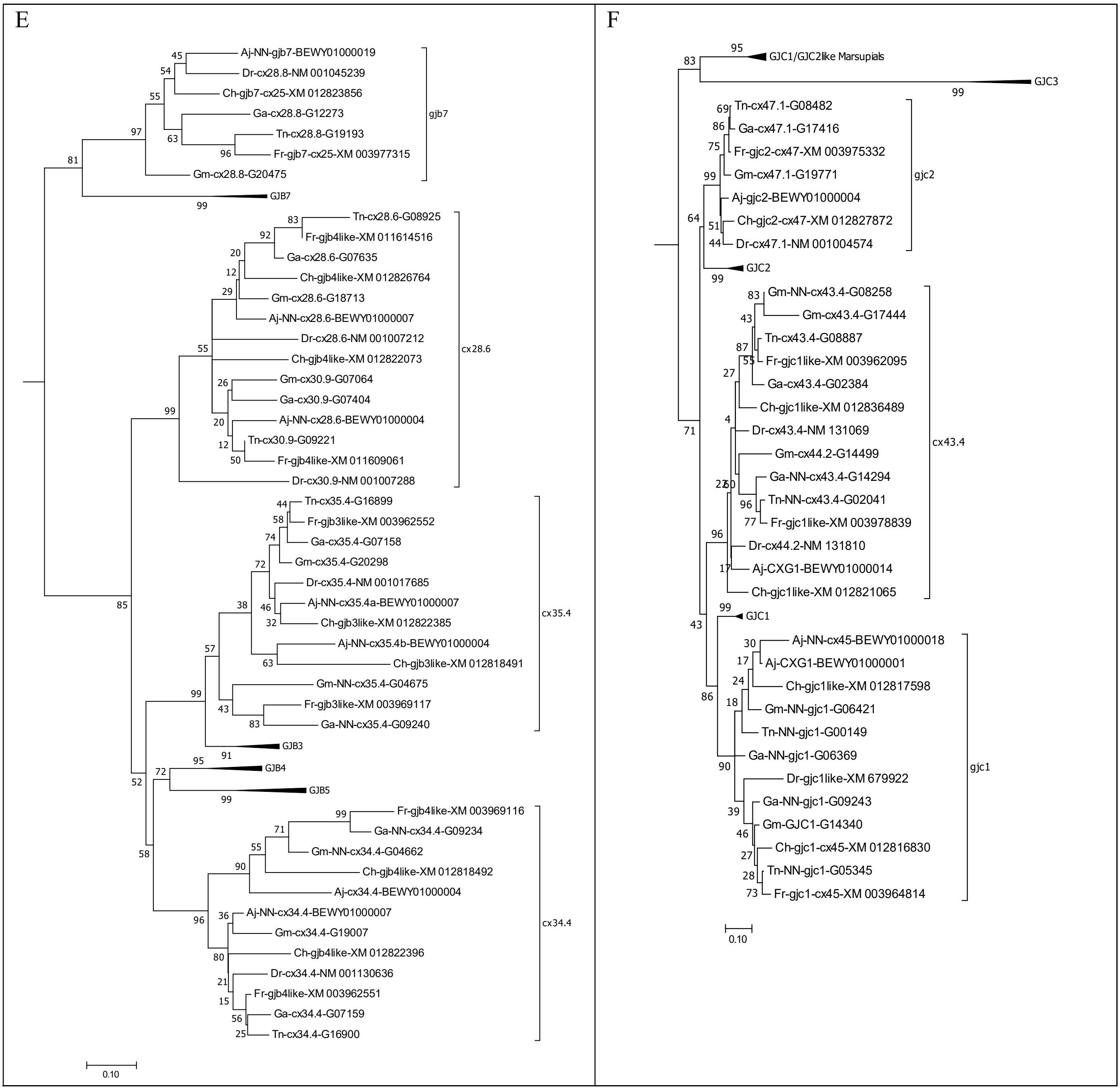
Suppl. Figure 14A. Mammalian and teleost *GJA1* and *GJA4* branches, and the teleost *cx32.2* and *cx34.5* branches.

Suppl. Figure 14B. Mammalian and teleost *GJA3* and *GJA8* branches, and the *GJA3*-related teleost *cx39.9* branch.



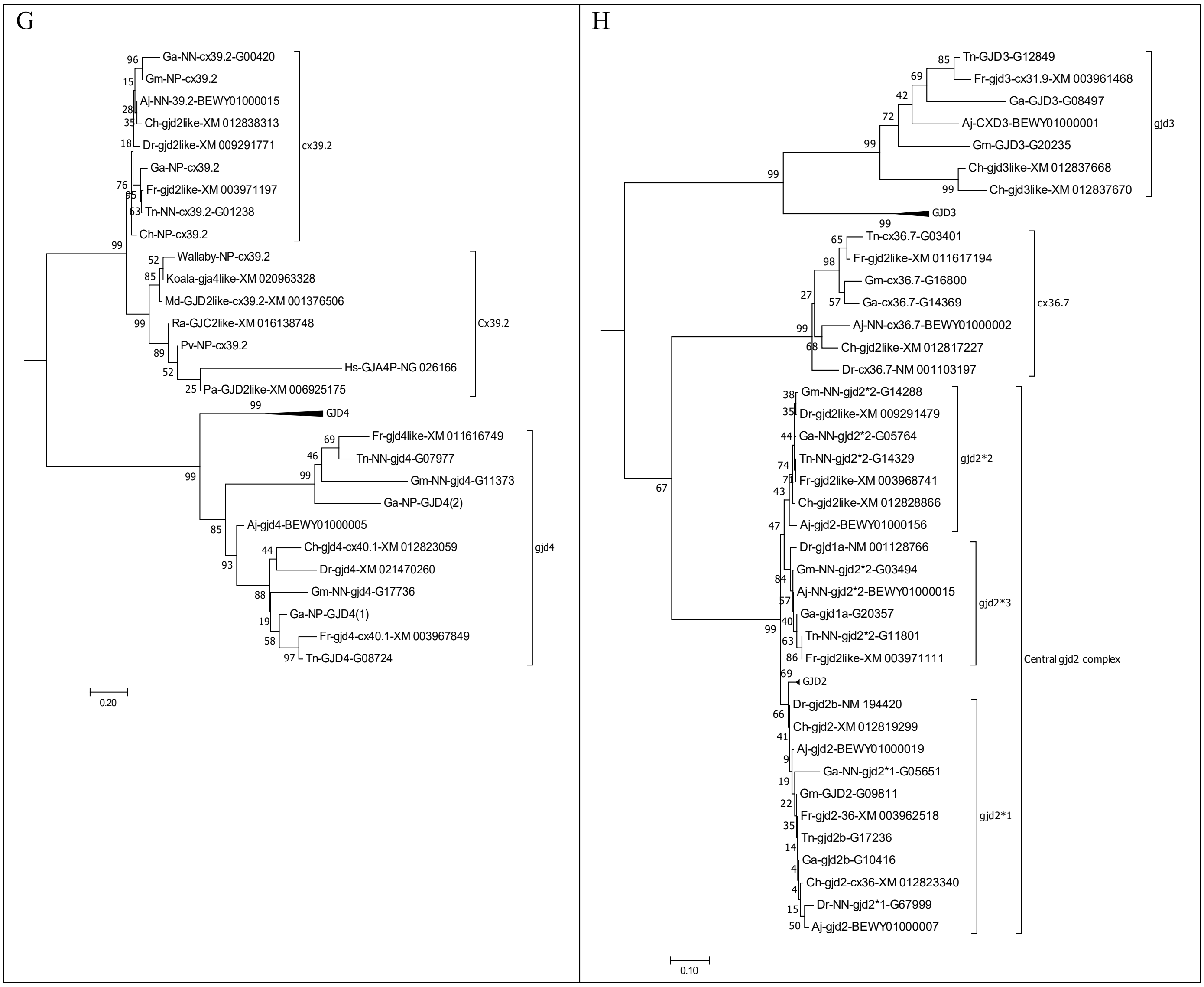
Suppl. Figure 14C. Mammalian and teleost GJA5, GJA9, and GJA10 branches. In most of the statistical analyses, *GJA10* and *gja10* switched location (*i.e., gja10* was locating outside [(*GJA9 – gja9*) - *GJA10*]).

Suppl. Figure 14D. Mammalian GJB1, GJB2, and GJB6 branches and their closest teleost homologues. Note that mammalian GJB2 and GJB6 always located together, and none of the Cx30.3 sequences ever interfered with the co-location of GJB2 and GJB6, indicating that Cx30.3 is equally distant to the splitting of the GJB2 and GJB6 groups.



Suppl. Figure 14E. Mammalian GJB7, GJB3, GJB4, and GJB5 branches and their closest teleost homologues. Mammalian GJB4 and GJB5 always located together, and cx34.4 or cx28.6 never interfered with the co-localization of GJB4 and GJB5.

Suppl. Figure 14F. Mammalian GJC1, GJC2, and GJC3 branches and their closest teleost homologues. Mammalian GJC3 and marsupial GJC1like/GJC2like tended to locate together (in 19 of 21 analyses), and thus were considered orthologues.



Suppl. Figure 14G. Mammalian and teleost *GJD4* and the associated teleost *Cx39.2*.

Suppl. Figure 14H. Mammalian and teleost *GJD3*, the central mammalian and teleost *GJD2* complex, and the *GJD2* associated teleost *cx36.7*.

Suppl. Table 1. Statistical support for clade grouping. It is referred to Fig. 1 in the paper for the naming of the different groups. To avoid some of the long-branch attraction and affected statistics, the *GJE1/gje1* group and the pseudogenes in the *Cx39.2* group, except the human pseudogene, were omitted in these statistical runs. The parameters for each run are given in Suppl. Table 2. For simplicity, the number of the analyses was counted from 1 when using the amino acid sequences, and from 20 when using nucleotide sequences. The white columns indicate bootstrap statistics (500 iterations) and the grey columns indicate interior branch statistics (500 iterations). The phylogenetic methods are abbreviated as follows: NJ, Neighbor Joining; ML, Maximum Likelihood; ME, Minimum Evolution; MP, Maximum Parsimony

Mammal or mammal-teleost	**Teleost**	**Sum>50 (Total)[E]**	**Amino acids**											**Nucleotides**									
Main model			NJ					ML		ME			MP	NJ				ML			ME		MP
Analysis # (Suppl. Table 2)			1	2	3	4	5	6	7	8	9	10	11	20	21[D]	22	23	24	25	26	27	28	29
GJA1	gja1	21/21	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
-	cx34.5-32.2	21/21	78	99	81	99	79	83	85	80	99	86	52[B]	82	75	71	99	82	88	82	83	99	68
GJA3	gja3	18/21	82	75	93	98	87	-	25[F]	74	79	99	-	92	84	99	99	78	84	81	98	99	89
Outside (GJA3-gja3)	cx39.9	19/21	98	99	98	99	99	-	91[B]	94	99	99	-	97	96	99	99	85	85	83	97	99	86
GJA4	gja4	3/21 (17/21)	17[B]	-	31	73	23[B]	19	25	27	-	40	-	41	48	15[B]	64	28	33	28[B]	41	65	-
GJA4 Not dichotomous	gja4	-	-	Tri[C]	-	-	-	-	-	-	Tri	-	-	-	-	-	-	-	-	-	-	-	-
GJA5	gja5	21/21	99	99	99	99	99	99	99	99	99	99	97	99	99	100	99	99	99	99	99	99	99
GJA8	gja8	21/21	99	99	99	99	99	99	99	99	99	99	99	99	99	100	99	99	99	99	99	99	99
GJA9	gja9	18/21	80	96	86	99	90	73	85	90	96	95	50[B]	76	80	76	98	-	-	-	67	97	52
GJA10	gja10		25[B]	-	-	-	-	20[B]	-	58[B]	-	-	31[B]	-	-	-	-	-	-	-	-	-	-
Outside (GJA10-GJA9-gja9)	gja10	11/21	99	Tri	99	Tri	99	99	99	Tri	Tri	-	-	99	99	100	Tri	99	99	99	Tri	Tri	-
GJB1	cx27.5	21/21	96	99	90	99	92	91	83	99	99	98	90	97	96	99	99	89	91	72	99	99	90
GJB2-GJB6	-	21/21	98	99	99	99	98	86	93	99	99	99	90	99	99	100	99	94	93	99	99	99	97
Outside (GJB2-GJB6)	cx30.3	18/21	58	85	66	89	68	59	63[B]	78	88	87	44	56	46	54	82	62	59[B]	62	83	82	-
GJB3	cx35.4	21/21	94	98	97	99	98	95	97	99	99	99	95	92	91	97	99	98	97	98	99[B]	99	96
GJB4-GJB5	-	21/21	92	91	76	70	76	84	80	91	92	72	67	87	83	97	79	83	79	67[B]	75	80	62
Outside (GJB4-GJB5)	cx34.4	7/21 (19/21)	64	54	56	35	47	40[B]	36	64	57	58	41	-	-	56	31	44[B]	37[B]	46[B]	45	32	46
Outside ((GJB3-cx35.4)-(GJB4-GJB5))	cx34.4	4/21	-	-	75[B]	-	73[B]	-	-	-	-	-	-	65	53	-	-	-	-	-	-	-	-
Outside ((GJB3-cx35.4)-cx28.6)	cx34.4	(5/21)	-	-	-	-	-	21	24	-	-	-	-	-	-	-	-	26	28	31	-	-	-
Outside ((GJB3-cx35.4)-(GJB4-GJB5)-cx34.4)	cx28.6	16/21	98	99	94	99	93	96[B]	-	96	99	85	-	74	69	89	99	90[B]	-	-	78[B]	99	-
Outside (GJB3-cx35.4)	cx28.6	2/21 (7/21)	-	-	-	-	-	38	39	-	-	-	-	-	-	-	-	53	47	66	20	-	39
GJB7	gjb7	17/21 (19/21)	80	79	71	91	80	99	94	88	82	81	96	46	43	74	-	88	91	80	76	-	74
GJC1	gjc1	14/21 (20/21)	62	89	80	90	80	25[B]	46	77	76[F]	86	39	54	41	-	65	67	70	56	46	68	40
GJC2	gjc2	10/21 (16/21)	69	49	73	56	57	75	70	69	52	64	55	16[B]	-	-	-	31[B]	34[B]	33[B]	-	-	49
Outside (GJC1-gjc1)	cx43.4	1/21 (4/21)	-	-	25	76	37	-	-	-	-	43	-	-	-	-	-	-	-	-	-	-	-

Outside ((GJC1-gjc1)-(GJC2-gjc2))	cx43.4	4/21 (6/21)	-	-	74[B]	-	90[B]	-	93[B]	32	-	-	-	-	-	47	-	-	88[B]	-	-	-	-
Outside (GJC2-gjc2)	cx43.4	2/21 (12/21)	33	12	-	-	-	59	42	82[B]	12	-	43	13[B]	-	-	-	19[B]	18[B]	17[B]	-	-	34
-	cx43.3-gjc2	(2/21)	-	-	-	-	-	-	-	-	-	-	-	26		-	-	-	-	-	21	Tri	-
GJC3-(GJC1like/GJC2like)	-	14/21 (19/21)	54	31	78	78	89	-	58	63	28	83	-	85	78	87	63	49	54	27[B]	88	62	39
GJD2	gjd2*1	8/21 (14/21)	30[B]	-	-	-	54	53	36[B]	59	76	66	-	-	-	-	93	47	45	45	64	92	34[B]
Outside (GJD2-gjd2*2-gjd2*3)	gjd2*1	4/21 (5/21)	-	-	-	-	-	-	-	-	-	-	24	70[B]	-	100	-	50	-	99[B]	-	-	-
GJD2 Not dichotomous	-	-	-	Tri	-	Tri	-		-	-	-	-	-				-	-	-	-	-		-
GJD2	gjd2*2	(4/21)	8	-	29	-	-	-	-	-	-	-	-	10	-	-	-	-	43	-	-	-	-
Outside (GJD2-gjd2*1)	(gjd2*2- gjd2*3)	11/21 (12/21)	99[B]	-	-	-	99	99	98[B]	99	99	99	-	20[B]	-	99	-	-	99	-	99	-	99
-	gjd2*2-gjd2*3	7/21 (11/21)	44[B]	-	-	-	52	66	60[B]	55	77[F]	47	-	29[B]		59	-	-	-	61[B]	25[B]	Tri	-
Outside (GJD2-gjd2*2-gjd2*1)	gjd2*3	3/21	-	99	99	99	-	-	-	-	-	-	-	-		-	Tri	-	-	-	-	-	-
GJD3	gjd3	21/21	96	99	99	99	99	98	99	99	99	99	98	99	99	99	99	99	99	99	99	99	98
GJD4	gjd4	21/21	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99
Outside GJD2 complex	cx36.7	14/21 (15/21)	59	95	61	97	-	-	-	58	99	67	49[B]	60	59	88	-	51	-	55	64	-	58
Outside (GJD3-gjd3)	cx36.7	(2/21)	-	-	-	-	-	22[B]	-	-	-	-	17	-	-	-	-	.-	-	-	-	-	-
Outside (GJD4-gjd4)	cx36.7	2/21 (8/21)	-	-	29[B]	-	38	-	-	-	-	-	-	39[B]	-	-	67	29[B]	32	-	31[B]	68	-
Cx39.2	cx39.2	21/21	98	99	99	99	99	97	98	96	99	99	99	99	99	100	99	99	98	99	99	99	99[B]
Hs-GJA4P within Cx39.2	-	21/21	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y

A. The mammalian sequences and the teleost sequences mix, so there is no clear dichotomy between mammalian and teleost sequences.
B. Bootstrap value from consensus tree. The consensus tree value was used if the original tree showed unexpected branching patterns.
C. Tri/Tetra: The branching pattern was trichotomous or tetratomous.
D. 53 of the bootstrap cycles failed. Thus, the bootstrap values are based on 447 replications.
E. Numbers without parentheses are summing up the number of analyses where the statistics is >50. The total number of analyses are 21. The numbers in parentheses are total number of analyses where there is some statistical support, no matter how weak.
F. One or two of the sequences split off from the remaining sequences in the group.

Suppl. Table 2. Parameter overview for statistical analyses of phylogenetic trees. The following parameters were permanent (if allowed within the phylogeny method): Rates among sites, gamma = 1.04* (implying exponential distribution of evolutionary rates among the sites); rates among lineages, different (if allowed); missing data treatment, pairwise deletion. All these statistical analyses were run in MEGA7. If the analyses were performed on nucleotide (NT) sequences, only position 1 and 2 in the codons were used. Otherwise, all substitutions are included, whether the analyses were performed on amino acid (AA) level or nucleotide (NT) level. The phylogenetic methods are abbreviated as follows: NJ, Neighbor Joining; ML, Maximum Likelihood; ME, Minimum Evolution; MP, Maximum Parsimony.

Analysis #	Phylogenetic method	Statistical test	Substitution model		Rates among lineages	# gamma categories	Gaps (deletion)	Tree interference	Initial tree	Branch swap filter	ME/MP search level
			AA/NT	Subst. matrix							
1	NJ	Bootstr 500	AA	Equal input	Different	-	Pairwise	-			
2	NJ	Int branch 500	AA	Equal input	Different	-	Pairwise				
3	NJ	Bootstr 500	AA	Dayhoff	Same	-	Pairwise	-			
4	NJ	Int branch 500	AA	Dayhoff	Same	-	Pairwise				
5	NJ	Bootstr 500	AA	JTT	Same	-	Pairwise	-			
6	ML	Bootstr 500	AA	Equal input	-	2	Partial (90%)	Nearest neighbour interchange	NJ	None	
7	ML	Bootstr500	AA	JTT	-	2	Partial (90%)	Nearest neighbour interchange	NJ	None	
8	ME Mac	Bootstr 500	AA	Equal input	Different	-	Pairwise	Close neighbour interchange	NJ		1
9	ME Mac	Int branch 500	AA	Equal input	Different	-	Pairwise	Close neighbour interchange	NJ		1
10	ME Mac	Bootstr500	AA	Dayhoff	Different	-	Pairwise	Close neighbour interchange	NJ		1
11	MP	Bootstr500	AA	-	-	-	Partial (90%)	Subtree-Pruning-Regrafting	10 **	-	1
20	NJ	Bootst 500	NT	Tamura 3 param.	Different	-	Pairwise				
21	NJ	Bootstr 500	NT	Tamura-Nei	Different	-	Pairwise				
22	NJ	Bootstr 500	NT	Max Comp likelihood	Different	-	Pairwise				
23	NJ	Intbranch500	NT	Max Comp likelihood	Different	-	Pairwise				
24	ML	Bootstr500	NT	Tamura 3 param.	-	2	Partial (90%)	Nearest neighbour interchange	NJ	None	
25	ML	Bootstr500	NT	Tamura-Nei	-	2	Partial (90%)	Nearest neighbour interchange	NJ	None	
26	ML	Bootstr500	NT	General Time reversible model	-	2	Partial (90%)	Nearest neighbour interchange	NJ	None	
27	ME	Bootstr500	NT	Max Comp likelihood	Different	-	Pairwise	Close neighbour interchange	NJ	-	1
28	ME	IntBr 500	NT	Max Comp likelihood	Different	-	Pairwise	Close neighbour interchange	NJ	-	1
29	MP	Bootstr500	NT	-	-	-	Partial (90%)	Subtree-Pruning-Regrafting	10	-	1

*An analysis of estimated gamma for the whole set of amino acid sequences was performed, indicating a gamma of approximately 1. A number of analyses were performed with different gamma values surrounding 1. Using a gamma value slightly above 1 reduced the number of instances where single or a few sequences split out of its/their group, making the branching pattern (and corresponding statistics) cleaner. Thus, gamma = 1.04 was chosen in the cases where the gamma value could be specified in the parameters.

Suppl. Table 9. Ohnology among teleost connexins. Ohnology is here functionally defined as being on different chromosomes, linkage groups or long scaffolds. For each main cell, the upper half shows the number of genes in the group for the different species (generally 1 or 2), while the lower half, which might be divided in two, shows the chromosome(s) or scaffolds (prefix, "sc")where these genes locate. If the location is given as "1/1" or "2/2/2", the genes are not ohnologs, but were generated by tandem gene duplication. Rand, scaffold/contig numbered "random". Genes that have found in other assemblies or by other groups, but are not found in the chromosomal assembly, are marked with "no hit".

Connexin group	Japanese eel		Herring			Zebrafish		Cod		Stickleback		Japanese pufferfish		Spotted pufferfish	
gja1	2[A]		2			2		2		1		1		1	
(43)	7[B]	19[B]	14	15		17	20	7	21	18		sc1725[C]		Rand	
cx34.5	1		1			1		1		1		1		1	
(32.7)	19		15			20		21		18		sc1917		14	
cx28.9	1		1			2		1		1		1		1	
(32.2)	19		15			20	20	21		18		sc1917		14	
cx32.2	1		2			2		1		1		1		1	
(32.2/32.3)	19		15	15		20	20	21		18		sc1917		14	
gja3	2		2			1		2		2		2		2	
	8	14	2	21		9		4	20	1	sc115[C]	1	8	2	3
cx39.9	2		2			1		3		2		2		2	
	8	15	8	20		5		7/7	10	4	7	14	15	1	7
gja4	2		1			1		1		1		1		1	
(39.4)	4	7	19			19		22		10		12		21	
gja5	2		2			2		1		2		2		2	
	8	14	2	21		1	9	no hit		6	16	1	4	7	17
gja8	1		2			2		1		1		1		1	
	8	14	2	21		1	9	20		16		1		2	
gja9	2		2			2		2		2		2		2	
(52.9/55.5)	7	sc68[C]	19	no hit		16	17	6	22	10	20	7	12	21	Rand
gja10	1		2			2		2		2		2		1	
(52.6/52.7)	19		14	15		17	20	5	no hit	18	sc128[C]	16	sc1843[C]	Rand	
gjb1	2		2			2		2		2		2		2	
(27.5/31.7)	8	15	20	no hit		5	14	7	10	4	7	14	15	1	7
cx30.3	2		3			1		2		2		3		4	
(33.8)	8	14	2	8	21	9		4	20	1	sc115[C]	1/1	8	2/2/2	3
cx28.6	2		2			2		2		2		2		2	
(30.9)	4	7	14	19		17	19	5	22	10	15	2	12	10	21
cx35.4	2		2			1		2		2		2		1	

	4	7	14	19	17		5	22		10	15	2	12	10	
cx34.4	2		2		1		2			2		2		1	
	4	7	14	19	17		5	22		10	15	2	12	10	
gjb7	1		1		1		1			1		1		1	
(28.8)	19		no hit		20		21			18		sc1688[C]		14	
gjc1	2		2		1		2			2		1		2	
	1	18	1	1	3		2	18		5	11	5		2	3
gjc2	1		1		1		1			1		1		1	
(47.1)	4		25		2		8			3		22		15	
cx43.4	1		2		2		3			2		2		2	
(44.2)	14		2	21	6	9	4	20	23	1	16	1	sc3571[C]	2	Rand
*gjd2*1*	2		2		2		2			2		1		1	
	7	19	14	15	17	20	5	21		15	18	2		10	
*gjd2*2/3*	2		1		2		2			2		2		2	
	15	sc156[C]	9		5	15	7	16		1	7	15	11	7	16
gjd3	1		2		-		1			1		1		1	
	1		1	1			18			5		1		Rand	
gjd4	1		1		1		2			2		2		2	
	5		17		24		2	23		3	21	10	22	15	Rand
cx39.2	1		2		1		1			2		1		1	
	15		8	9	15		16			7	sc119[C]	15		7	
cx36.7	1		1		1		1			1		1		1	
	2		6		7		14			2		sc1921[C]		Rand	
gje1	1		1		2		1			1		1		-	
	19		14		17	20	21			18		16		-	

A: The number of sequences in this group in this species.

B: The chromosomal location of the genes mentioned in the subcell above.

C: The scaffold has not been placed into a chromosome.

Supplementary material for Manuscript 3.

Supplementary File 1. List of SNPs associated with sex in Atlantic herring (*Clupea harengus),* identified via a genome wide association study.

CHROM	POS
NW_012219506.1	919996
NW_012219506.1	920006
NW_012219506.1	920082
NW_012219506.1	920107
NW_012219506.1	920132
NW_012219506.1	920168
NW_012219506.1	920404
NW_012219506.1	920419
NW_012219506.1	920438
NW_012219506.1	920472
NW_012219506.1	920501
NW_012219506.1	920525
NW_012219506.1	920534
NW_012219506.1	920540
NW_012219506.1	920693
NW_012219506.1	920725
NW_012219506.1	920844
NW_012219506.1	920913
NW_012219506.1	920928
NW_012219506.1	920940
NW_012219506.1	923420
NW_012219506.1	923442
NW_012219506.1	923463
NW_012219506.1	923497
NW_012219506.1	923508
NW_012219506.1	923553
NW_012219506.1	923565
NW_012219506.1	923619
NW_012219506.1	923702
NW_012219506.1	923781
NW_012219506.1	923786
NW_012219506.1	923837
NW_012219506.1	923927
NW_012219506.1	923945
NW_012219506.1	923950
NW_012219506.1	923967
NW_012219506.1	923980
NW_012219506.1	924050
NW_012219506.1	924064
NW_012219506.1	924146
NW_012219506.1	925029
NW_012219506.1	925043
NW_012219506.1	925075
NW_012219506.1	925090
NW_012219506.1	925223
NW_012219506.1	925241
NW_012219506.1	925307
NW_012219506.1	925333
NW_012219506.1	925533
NW_012219506.1	925662
NW_012219506.1	925753
NW_012219506.1	925784
NW_012219506.1	925817
NW_012219506.1	925865
NW_012219506.1	925867
NW_012219506.1	926126
NW_012219506.1	926572
NW_012219506.1	926618
NW_012219506.1	926625
NW_012219506.1	926673
NW_012219506.1	926746
NW_012219506.1	926800
NW_012219506.1	926879
NW_012219506.1	926915
NW_012219506.1	927001
NW_012219506.1	928063
NW_012219506.1	928085
NW_012219506.1	928179
NW_012219506.1	928352
NW_012219506.1	928531
NW_012219506.1	928552
NW_012219506.1	928562
NW_012219506.1	928664
NW_012219506.1	928754
NW_012219506.1	928804
NW_012219506.1	928862
NW_012219506.1	929533
NW_012219506.1	929596
NW_012219506.1	929843
NW_012219506.1	930168
NW_012219506.1	930194
NW_012219506.1	930214
NW_012219506.1	930259
NW_012219506.1	930469
NW_012219506.1	930483
NW_012219506.1	930607
NW_012219506.1	930658
NW_012219506.1	930941
NW_012219506.1	931065
NW_012219506.1	931101
NW_012219506.1	931327
NW_012219506.1	931618
NW_012219506.1	931639

NW_012219506.1	931665
NW_012219506.1	931692
NW_012219506.1	931761
NW_012219506.1	931928
NW_012219506.1	931960
NW_012219506.1	932072
NW_012219506.1	932089
NW_012219506.1	932195
NW_012219506.1	932460
NW_012219506.1	932495
NW_012219506.1	933220
NW_012219506.1	933245
NW_012219506.1	933440
NW_012219506.1	933630
NW_012219506.1	933642
NW_012219506.1	934107
NW_012219506.1	934464
NW_012219506.1	934505
NW_012219506.1	934591
NW_012219506.1	934636
NW_012219506.1	935373
NW_012219506.1	935716
NW_012219506.1	935946
NW_012219506.1	935956
NW_012219506.1	936262
NW_012219506.1	936355
NW_012219506.1	936649
NW_012219506.1	936718
NW_012219506.1	936941
NW_012219506.1	936988
NW_012219506.1	937019
NW_012219506.1	937247
NW_012219506.1	937275
NW_012219506.1	937591
NW_012219506.1	937603
NW_012219506.1	937608
NW_012219506.1	937761
NW_012219506.1	937876
NW_012219506.1	938202
NW_012219506.1	938247
NW_012219506.1	938413
NW_012219506.1	938777
NW_012219506.1	938827
NW_012219506.1	938936
NW_012219506.1	939339
NW_012219506.1	939882
NW_012219506.1	940076
NW_012219506.1	940294
NW_012219506.1	940681
NW_012219506.1	940760
NW_012219506.1	940926
NW_012219506.1	940966
NW_012219506.1	941652
NW_012219506.1	942005
NW_012219506.1	942336
NW_012219506.1	942896
NW_012219506.1	943227
NW_012219506.1	943336
NW_012219506.1	943815
NW_012219506.1	943836
NW_012219506.1	944399
NW_012219506.1	944922
NW_012219506.1	945092
NW_012219506.1	945185
NW_012219506.1	945206
NW_012219506.1	945891
NW_012219506.1	945898
NW_012219506.1	948796
NW_012219506.1	949601
NW_012219506.1	949713
NW_012219506.1	949823
NW_012219506.1	950674
NW_012219506.1	951602
NW_012219506.1	952391
NW_012219506.1	952412
NW_012219506.1	952568
NW_012219506.1	952759
NW_012219506.1	952807
NW_012219506.1	952838
NW_012219506.1	952861
NW_012219506.1	953092
NW_012219506.1	953104
NW_012219506.1	953597
NW_012219506.1	953687
NW_012219506.1	953918
NW_012219506.1	954197
NW_012219506.1	954294
NW_012219506.1	954416
NW_012219506.1	954756
NW_012219506.1	954778
NW_012219506.1	954810
NW_012219506.1	954838
NW_012219506.1	955098
NW_012219506.1	955257
NW_012219506.1	955453
NW_012219506.1	955482
NW_012219506.1	955558
NW_012219506.1	956092
NW_012219506.1	956234
NW_012219506.1	956252
NW_012219506.1	956867
NW_012219506.1	957043
NW_012219506.1	957057
NW_012219506.1	957075
NW_012219506.1	957131
NW_012219506.1	957230
NW_012219506.1	957899
NW_012219506.1	958112
NW_012219506.1	959676
NW_012219506.1	959943
NW_012219506.1	960119

NW_012219506.1	960687	NW_012219506.1	972354	NW_012219506.1	979635
NW_012219506.1	960740	NW_012219506.1	972965	NW_012219506.1	979733
NW_012219506.1	960774	NW_012219506.1	973180	NW_012219506.1	979795
NW_012219506.1	960883	NW_012219506.1	973655	NW_012219506.1	979915
NW_012219506.1	961335	NW_012219506.1	973897	NW_012219506.1	980026
NW_012219506.1	961452	NW_012219506.1	974001	NW_012219506.1	980033
NW_012219506.1	961511	NW_012219506.1	974009	NW_012219506.1	980137
NW_012219506.1	961547	NW_012219506.1	974069	NW_012219506.1	980650
NW_012219506.1	961841	NW_012219506.1	974550	NW_012219506.1	980753
NW_012219506.1	962082	NW_012219506.1	974882	NW_012219506.1	980821
NW_012219506.1	962153	NW_012219506.1	975093	NW_012219506.1	980914
NW_012219506.1	962789	NW_012219506.1	975264	NW_012219506.1	981153
NW_012219506.1	963621	NW_012219506.1	975343	NW_012219506.1	981175
NW_012219506.1	963774	NW_012219506.1	975734	NW_012219506.1	981353
NW_012219506.1	963804	NW_012219506.1	975836	NW_012219506.1	981800
NW_012219506.1	963809	NW_012219506.1	975988	NW_012219506.1	983454
NW_012219506.1	965261	NW_012219506.1	976049	NW_012219506.1	983599
NW_012219506.1	965274	NW_012219506.1	976332	NW_012219506.1	983944
NW_012219506.1	965322	NW_012219506.1	976879	NW_012219506.1	984053
NW_012219506.1	965358	NW_012219506.1	977029	NW_012219506.1	984150
NW_012219506.1	965964	NW_012219506.1	977096	NW_012219506.1	984489
NW_012219506.1	966040	NW_012219506.1	977125	NW_012219506.1	984665
NW_012219506.1	966227	NW_012219506.1	977666	NW_012219506.1	984772
NW_012219506.1	966291	NW_012219506.1	977675	NW_012219506.1	984917
NW_012219506.1	966333	NW_012219506.1	977743	NW_012219506.1	985614
NW_012219506.1	966377	NW_012219506.1	978230	NW_012219506.1	985886
NW_012219506.1	966454	NW_012219506.1	978293	NW_012219506.1	986143
NW_012219506.1	967185	NW_012219506.1	978369	NW_012219506.1	986598
NW_012219506.1	969077	NW_012219506.1	978384	NW_012219506.1	986720
NW_012219506.1	969157	NW_012219506.1	978501	NW_012219506.1	986736
NW_012219506.1	969981	NW_012219506.1	978563	NW_012219506.1	986762
NW_012219506.1	970971	NW_012219506.1	978622	NW_012219506.1	987194
NW_012219506.1	971050	NW_012219506.1	978873	NW_012219506.1	987413
NW_012219506.1	971349	NW_012219506.1	978885	NW_012219506.1	987609
NW_012219506.1	971891	NW_012219506.1	979102	NW_012219506.1	988666
NW_012219506.1	971908	NW_012219506.1	979218	NW_012219506.1	989019
NW_012219506.1	971918	NW_012219506.1	979420	NW_012219506.1	989117

NW_012219506.1	989556	NW_012219506.1	998320	NW_012219506.1	1008796
NW_012219506.1	989641	NW_012219506.1	998463	NW_012219506.1	1009163
NW_012219506.1	989836	NW_012219506.1	998504	NW_012219506.1	1009287
NW_012219506.1	990298	NW_012219506.1	998721	NW_012219506.1	1009394
NW_012219506.1	990305	NW_012219506.1	999001	NW_012219506.1	1009999
NW_012219506.1	990406	NW_012219506.1	999039	NW_012219506.1	1010290
NW_012219506.1	990817	NW_012219506.1	999239	NW_012219506.1	1010397
NW_012219506.1	991179	NW_012219506.1	999498	NW_012219506.1	1010427
NW_012219506.1	992000	NW_012219506.1	999548	NW_012219506.1	1010620
NW_012219506.1	992104	NW_012219506.1	999596	NW_012219506.1	1010646
NW_012219506.1	992266	NW_012219506.1	999746	NW_012219506.1	1010956
NW_012219506.1	992348	NW_012219506.1	999998	NW_012219506.1	1011769
NW_012219506.1	992494	NW_012219506.1	1000105	NW_012219506.1	1014294
NW_012219506.1	992677	NW_012219506.1	1001373	NW_012219506.1	1014301
NW_012219506.1	992733	NW_012219506.1	1002626	NW_012219506.1	1014332
NW_012219506.1	992775	NW_012219506.1	1002678	NW_012219506.1	1014573
NW_012219506.1	992912	NW_012219506.1	1002775	NW_012219506.1	1014600
NW_012219506.1	993023	NW_012219506.1	1003638	NW_012219506.1	1014629
NW_012219506.1	993035	NW_012219506.1	1003827	NW_012219506.1	1014689
NW_012219506.1	993122	NW_012219506.1	1004934	NW_012219506.1	1014713
NW_012219506.1	993340	NW_012219506.1	1005290	NW_012219506.1	1015160
NW_012219506.1	993627	NW_012219506.1	1005402	NW_012219506.1	1015245
NW_012219506.1	993672	NW_012219506.1	1005428	NW_012219506.1	1015266
NW_012219506.1	993686	NW_012219506.1	1005482	NW_012219506.1	1015586
NW_012219506.1	994318	NW_012219506.1	1005975	NW_012219506.1	1015721
NW_012219506.1	994532	NW_012219506.1	1006057	NW_012219506.1	1015778
NW_012219506.1	994770	NW_012219506.1	1006156	NW_012219506.1	1015799
NW_012219506.1	995541	NW_012219506.1	1006186	NW_012219506.1	1016316
NW_012219506.1	995659	NW_012219506.1	1006569	NW_012219506.1	1016668
NW_012219506.1	995852	NW_012219506.1	1006902	NW_012219506.1	1016676
NW_012219506.1	996071	NW_012219506.1	1007699	NW_012219506.1	1017046
NW_012219506.1	996202	NW_012219506.1	1007748	NW_012219506.1	1017293
NW_012219506.1	996497	NW_012219506.1	1007991	NW_012219506.1	1017365
NW_012219506.1	996520	NW_012219506.1	1008047	NW_012219506.1	1017621
NW_012219506.1	997163	NW_012219506.1	1008095	NW_012219506.1	1017715
NW_012219506.1	997254	NW_012219506.1	1008426	NW_012219506.1	1018836
NW_012219506.1	997881	NW_012219506.1	1008564	NW_012219506.1	1018943

NW_012219506.1	1018997
NW_012219506.1	1019219
NW_012219506.1	1019373
NW_012219506.1	1019668
NW_012219506.1	1019827
NW_012219506.1	1019850
NW_012219506.1	1020038
NW_012219506.1	1020156
NW_012219506.1	1020204
NW_012219506.1	1020321
NW_012219506.1	1020382
NW_012219506.1	1020443
NW_012219506.1	1020582
NW_012219506.1	1020731
NW_012219506.1	1021149
NW_012219506.1	1021179
NW_012219506.1	1021245
NW_012219506.1	1021314
NW_012219506.1	1021546
NW_012219506.1	1021630
NW_012219506.1	1021740
NW_012219506.1	1021972
NW_012219506.1	1022202
NW_012219506.1	1022223
NW_012219506.1	1022358
NW_012219506.1	1022376
NW_012219506.1	1022440
NW_012219506.1	1022672
NW_012219506.1	1022746
NW_012219506.1	1022765
NW_012219506.1	1023247
NW_012219506.1	1023302
NW_012219506.1	1023530
NW_012219506.1	1024238
NW_012219506.1	1026293
NW_012219506.1	1027253
NW_012219506.1	1027448
NW_012219506.1	1027737
NW_012219506.1	1028166
NW_012219506.1	1028426
NW_012219506.1	1028461
NW_012219506.1	1028474
NW_012219506.1	1029248
NW_012219506.1	1029311
NW_012219506.1	1029335
NW_012219506.1	1029545
NW_012219506.1	1029733
NW_012219506.1	1029837
NW_012219506.1	1030003
NW_012219506.1	1030091
NW_012219506.1	1030172
NW_012219506.1	1030275
NW_012219506.1	1030500
NW_012219506.1	1030774
NW_012219506.1	1031208
NW_012219506.1	1031287
NW_012219506.1	1031853
NW_012219506.1	1032336
NW_012219506.1	1032435
NW_012219506.1	1032459
NW_012219506.1	1032481
NW_012219506.1	1032935
NW_012219506.1	1033247
NW_012219506.1	1033288
NW_012219506.1	1033850
NW_012219506.1	1034217
NW_012219506.1	1034301
NW_012219506.1	1034360
NW_012219506.1	1034402
NW_012219506.1	1034661
NW_012219506.1	1034866
NW_012219506.1	1034880
NW_012219506.1	1035290
NW_012219506.1	1035605
NW_012219506.1	1035611
NW_012219506.1	1035616
NW_012219506.1	1035763
NW_012219506.1	1035993
NW_012219506.1	1036042
NW_012219506.1	1036238
NW_012219506.1	1036419
NW_012219506.1	1036800
NW_012219506.1	1036859
NW_012219506.1	1037183
NW_012219506.1	1038782
NW_012219506.1	1038983
NW_012219506.1	1039193
NW_012219506.1	1039292
NW_012219506.1	2305958
NW_012219506.1	2305994
NW_012219506.1	2306041
NW_012219506.1	2306063
NW_012219506.1	2307704
NW_012219506.1	2307750
NW_012219506.1	2307866
NW_012219506.1	2307878
NW_012219506.1	2307925
NW_012219506.1	2307932
NW_012219506.1	2308028
NW_012219506.1	2311751
NW_012219506.1	2311759
NW_012219506.1	2311901
NW_012219506.1	2312043
NW_012219506.1	2312052
NW_012219506.1	2312062
NW_012219506.1	2312076
NW_012219506.1	2312082
NW_012219506.1	2313531
NW_012219506.1	2313558
NW_012219506.1	2314060
NW_012219506.1	2314071

NW_012219506.1	2314089
NW_012219506.1	2314115
NW_012219506.1	2314119
NW_012219506.1	2314164
NW_012219506.1	2316178
NW_012219506.1	2316320
NW_012219506.1	2316440
NW_012219506.1	2316452
NW_012219506.1	2317147
NW_012219506.1	2319020
NW_012219506.1	2319254
NW_012219506.1	2319436
NW_012219506.1	2319478
NW_012219506.1	2319536
NW_012219506.1	2319648
NW_012219506.1	2320200
NW_012219506.1	2320438
NW_012219506.1	2320953
NW_012219506.1	2321039
NW_012219703.1	28023
NW_012219703.1	28069
NW_012219703.1	28094
NW_012221357.1	979071
NW_012221357.1	979136
NW_012221357.1	979141
NW_012221357.1	979160
NW_012223947.1	3178113
NW_012223947.1	3179040
NW_012223947.1	3179856
NW_012223947.1	3180530
NW_012223947.1	3181482
NW_012223947.1	3182376
NW_012223947.1	7965877
NW_012223947.1	7965890
NW_012223947.1	7965966
NW_012223947.1	7965989
NW_012223947.1	7966349
NW_012223947.1	7966367
NW_012223947.1	7966401
NW_012223947.1	7967601
NW_012223947.1	7967633
NW_012223947.1	7967673
NW_012223947.1	7967706
NW_012223947.1	7967714
NW_012223947.1	7967944
NW_012223947.1	7967955
NW_012223947.1	7968244

Supplementary File 2. Test results from the comparison of the observed proportions of homozygous female and male genotypes versus coverage to the corresponding theoretically expected probabilities.

Table S1. The experimental data for homozygous male and female Atlantic herring (*Clupea harengus*) versus coverage and test results from two-sided exact binominal tests. Number of samples is represented by n, while 95%CI low and high show the 95% confidence intervals from the binomial test. No tests were carried out for coverage higher than 21, because of low number of samples with such high coverage.

Sex	Coverage	n	Number of homozygous	Observed homozygous proportion	Expected homozygous proportion	95%CI low	95%CI high	H_0	p-value
Female	1	3870	3870	1.0000	1.0000	0.9991	1.0000	1.0000	1.0000
Female	2	4485	4459	0.9942	1.0000	0.9915	0.9962	1.0000	0.0000
Female	3	3645	3606	0.9893	1.0000	0.9854	0.9924	1.0000	0.0000
Female	4	3196	3164	0.9900	1.0000	0.9859	0.9931	1.0000	0.0000
Female	5	2424	2399	0.9897	1.0000	0.9848	0.9933	1.0000	0.0000
Female	6	1876	1863	0.9931	1.0000	0.9882	0.9963	1.0000	0.0000
Female	7	1319	1308	0.9917	1.0000	0.9851	0.9958	1.0000	0.0000
Female	8	1016	1010	0.9941	1.0000	0.9872	0.9978	1.0000	0.0000
Female	9	720	716	0.9944	1.0000	0.9858	0.9985	1.0000	0.0000
Female	10	564	562	0.9965	1.0000	0.9873	0.9996	1.0000	0.0000
Female	11	399	398	0.9975	1.0000	0.9861	0.9999	1.0000	0.0000
Female	12	301	300	0.9967	1.0000	0.9816	0.9999	1.0000	0.0000
Female	13	184	183	0.9946	1.0000	0.9701	0.9999	1.0000	0.0000
Female	14	154	150	0.9740	1.0000	0.9348	0.9929	1.0000	0.0000
Female	15	99	98	0.9899	1.0000	0.9450	0.9997	1.0000	0.0000
Female	16	50	50	1.0000	1.0000	0.9289	1.0000	1.0000	1.0000
Female	17	42	42	1.0000	1.0000	0.9159	1.0000	1.0000	1.0000
Female	18	23	22	0.9565	1.0000	0.7805	0.9989	1.0000	0.0000
Female	19	17	16	0.9412	1.0000	0.7131	0.9985	1.0000	0.0000

Female	20	8	8	1.0000	1.0000	0.6306	1.0000	1.0000	1.0000
Female	21	7	7	1.0000	1.0000	0.5904	1.0000	1.0000	1.0000
Male	1	3978	3978	1.0000	1.0000	0.9991	1.0000	1.0000	1.0000
Male	2	4855	3278	0.6752	0.5000	0.6618	0.6884	0.5000	0.0000
Male	3	4371	1776	0.4063	0.2500	0.3917	0.4211	0.2500	0.0000
Male	4	3849	1047	0.2720	0.1250	0.2580	0.2864	0.1250	0.0000
Male	5	3008	574	0.1908	0.0625	0.1769	0.2053	0.0625	0.0000
Male	6	2494	332	0.1331	0.0313	0.1200	0.1471	0.0313	0.0000
Male	7	1851	162	0.0875	0.0156	0.0750	0.1013	0.0156	0.0000
Male	8	1386	105	0.0758	0.0078	0.0624	0.0910	0.0078	0.0000
Male	9	913	55	0.0602	0.0039	0.0457	0.0777	0.0039	0.0000
Male	10	683	37	0.0542	0.0020	0.0384	0.0739	0.0020	0.0000
Male	11	531	25	0.0471	0.0010	0.0307	0.0687	0.0010	0.0000
Male	12	330	34	0.1030	0.0005	0.0724	0.1410	0.0005	0.0000
Male	13	241	11	0.0456	0.0002	0.0230	0.0802	0.0002	0.0000
Male	14	177	8	0.0452	0.0001	0.0197	0.0871	0.0001	0.0000
Male	15	116	4	0.0345	0.0001	0.0095	0.0859	0.0001	0.0000
Male	16	96	6	0.0625	0.0000	0.0233	0.1311	0.0000	0.0000
Male	17	60	2	0.0333	0.0000	0.0041	0.1153	0.0000	0.0000
Male	18	55	2	0.0364	0.0000	0.0044	0.1253	0.0000	0.0000
Male	19	28	0	0.0000	0.0000	0.0000	0.1234	0.0000	1.0000
Male	20	38	2	0.0526	0.0000	0.0064	0.1775	0.0000	0.0000
Male	21	18	0	0.0000	0.0000	0.0000	0.1853	0.0000	1.0000

Supplementary File 3. Alignment of *loc105890447* and *loc105890510* to bicaudal D-related protein-like orthologs.

loc105890510

Sequence below shows Atlantic herring *loc105890510*. The indicated parts of the sequence align with the following GenBank entries: underlined, Atlantic herring XR_001161982 ncRNA; red font, Takifugu XM_003971029 Bicaudal D-related protein 1-like positions 997-1118; italics, salmon XM_014129619 Bicaudal D-related protein 2-like positions 15-53 and 1369-1493; font size 12 (larger font), zebrafish NM_001365668 Bicaudal D-related protein-2-like, posistions 822-936. Overall, we would suggest that red part of the sequence in *loc105890510* corresponds to an exon in herring bicaudal D-related protein 2-like.

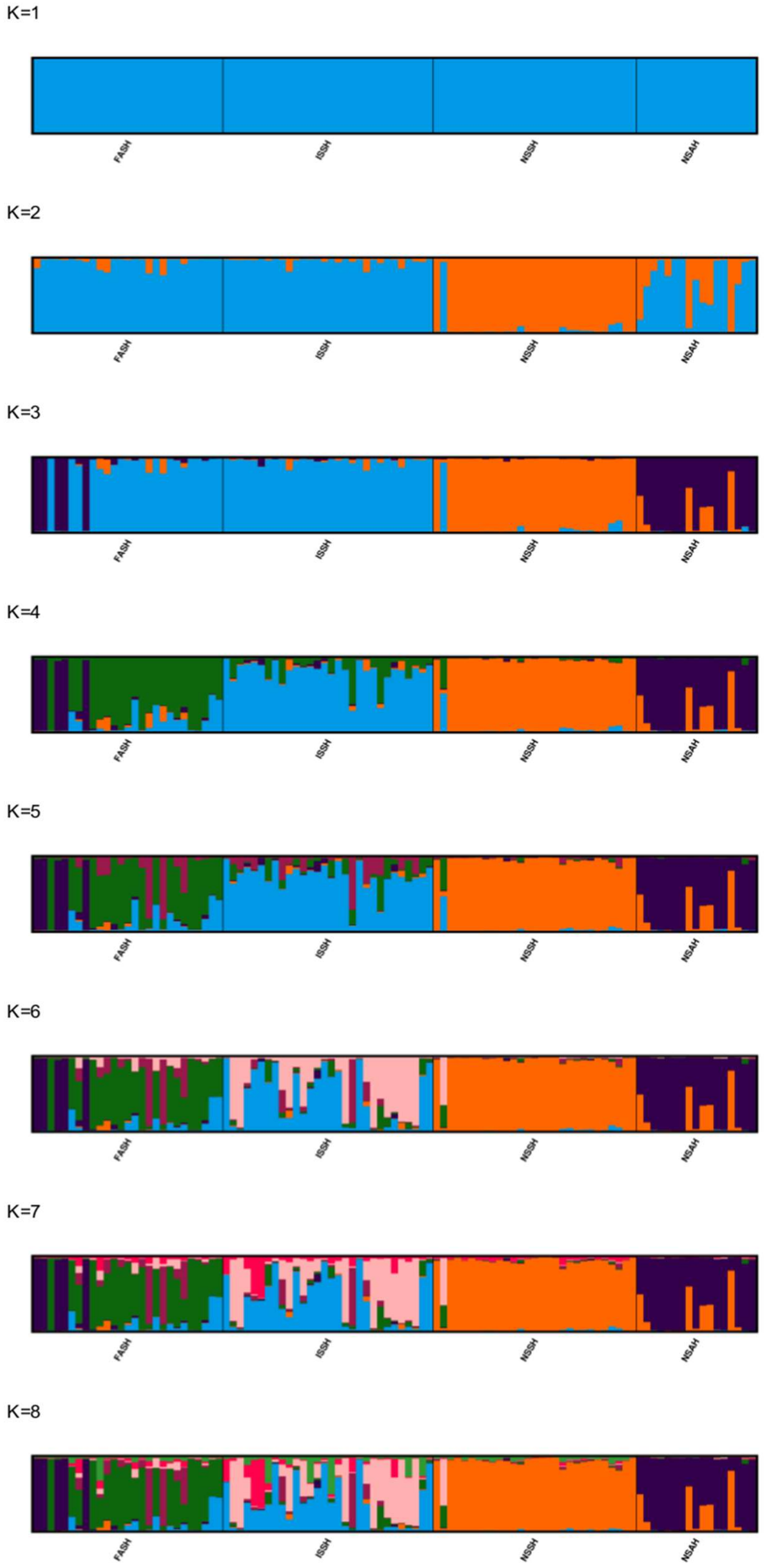
ATACTGAAGATGGG*GACAGTGCTGTTCTACAACAGGCCTTACGAGACAGAGAC*CAAGCAGTCACAAAGTGAGTGC
CAAACTGCAGAAGCACACTTACGCAGGACTTCCACCCTACCAACAACAACATGGAATTGAACTTTTAACCTTTAA
TAAAGCATTTAATGGTATTTTATTCAAACATGTTACAACATGTAACACAGTTTGCAAAGGTCAGACCTCAGTCAT
TTTGGTACTTATGACACAAATGTAACACCTGAAATACTAATTTAGTCGTCCTGATGCCTAGCTGAGCGCACAAAG
GGAAGAGACTGATTGACATTTTCAGTCATTACAAAGAGAGAGGCTAAGAGAGATAAACACTACACAGAAAGCATT
ACTTTACAGGACTCCTTTCTTTTAGTCAAATATTTCCCAAAATAGGGACCTCTGTTTTCATCTTAACCATTATCC
TCAGCCAGAATCTTCTTCAAATGCTGACCAGATTATACATTTAGTATCCCTTTTAAGAGGTCCATTGTGCTCCCC
TGTAGTATTCATTCTATAGTTTGTTTAGTGCTAATGCCTTAAATCTGGCATCATTTGCAAACTAATCTCGACTAC
TTCCCCCTGCTTGTGTTCTGTGTGG*CAGGAAGAAAGCCATGGAGAGGGAGCTGCTGAACAGCAAGAC*
GGAGATGATGAACATGAACAACCAGCTGCTGGAGGCGGTGCAGCACCGGCTGGAGCTCTCTC
*TGGAGCTAGAGGCCTGGAAGG*TGCGCTGAGACACTGAGACAGAGGCTCGCTTATAGACAGAGATGCTCTGT
GGGCGTACAGCTGTGAGGCTGTGGGGTGGAGGAGACTGGAGAGGGAAGGATGGGTATTTAGATTTAGAGAGATAT
GGGAGGTCACAGATGGTTTTCCCATTACTGAATGTTTTTGCCGTCGAAAAATGAAATATGTACCCGACACTGATT
TCACACATTCTTCAACTTAGACTCAGCAGTGGCAGGAAGTATAGGAATGTTTTAGTTAAAACAAATGCATTTTTA
AATTGTCTCAATTGTGGTCAAGAAACTTAGATGTACGTAGAGTCTAAATGTTTTTGAGTTATGCAATGTTTAGAT
TTGAGTTATGAAATGGTCAGATTAGATATAAAAATATTTCTCTCTCTCTCTCTCTCCCTCTGTGTGGCAGGAG
GACTTCCAGCTGCTCCTCCAGCAGCAGGTGTTGTCTCAGCAG

loc105890447

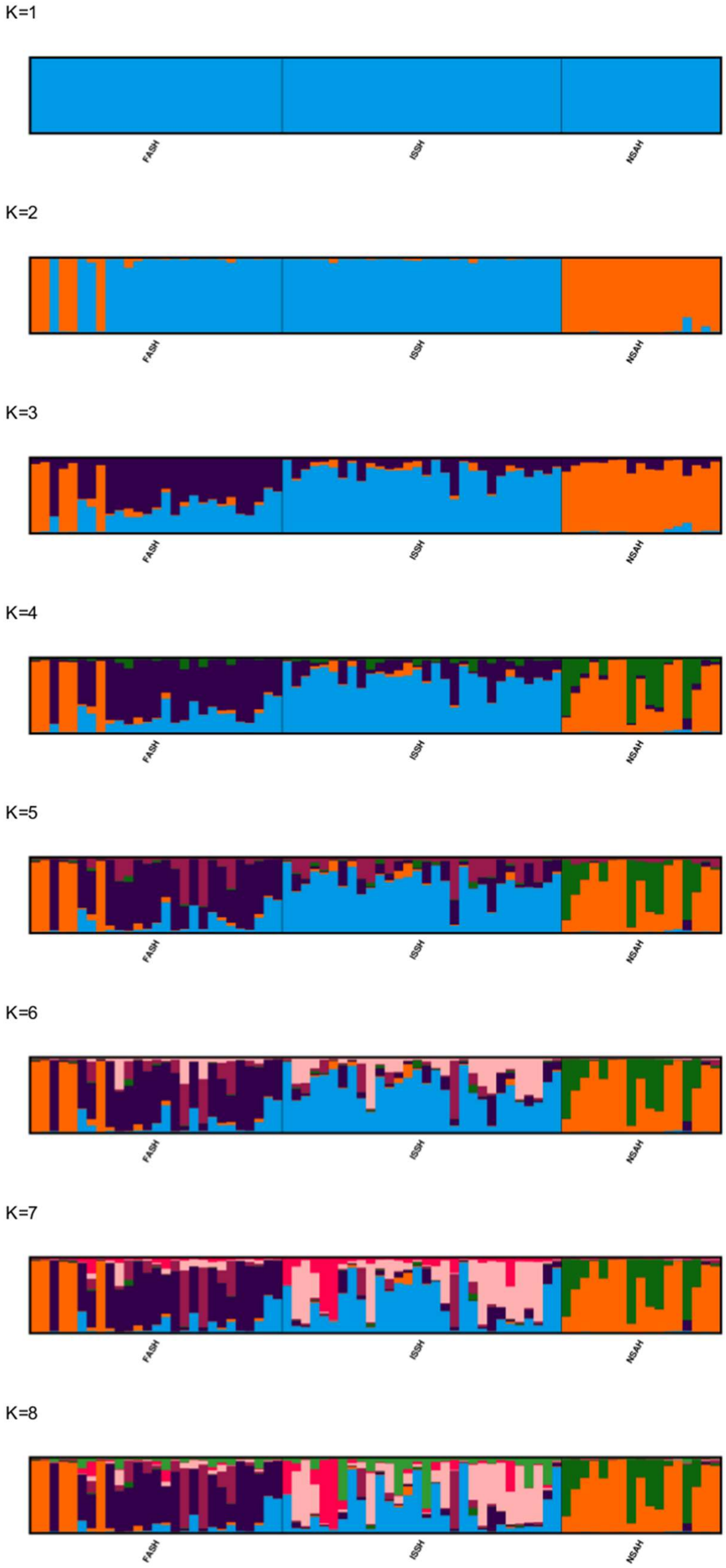
Sequence below shows Atlantic herring *loc105890447*. The indicated parts of the sequence align with the following GenBank entries: underlined, Atlantic herring uncharacterized XM_012816471; red font, Takifugu XM_003971029 bicaudal D-related protein positions 650-753; italics, salmon XM_014129619 Bicaudal D-related protein 2-like positions 933-1086; bold, Astyanax XM_007254183 Bicaudal D family like adapter 2; font size 12 (larger font), zebrafish NM_001365668 bicaudal D-related protein 2-like positions 380-454 and 584-735. Overall, we would suggest that the indicated parts the sequence in *loc105890447* corresponds to three exons in herring bicaudal D-related protein 2-like.

ATGGACTCAGTTGCTTTGCCTGAGACAGAGGACCAGCCTGAGACTGAGGTCTGTGTCGGGCAAAATGTCTGCACA
CCTCGGACCATACTAGAGGGACTGGTGGCACCAAGGCACTTTGGAAAACCCAGCCTGGCTGCTCCGGGGGAAGGA
GTGGAGATTGTTGTCCCAGACGCTGAAGGTTTGAGCTCTCTTCAGTCACCGCAGCGTGAAGAACTGAAGGAGTCA
GACTCGACAGAAAACGATGTTGGACATGTTCTCCAAGAGAACCGTGTTCTGTCAGAGGAGAGTGAGGACCCTGTG
CCACAGGCAGAGGACAGTTTCAGGCCTGCCATGAGCTCTCCACTCAGGCACTATATTGATGGGACTGTG**CC**
TGATCTGCTGAGGAGTGGGAGCCCCCTGCAAAGGAGAGTGTCCAGTCCAGTGTCTAACACCGTG
AGTAGGAAAGAAAGTCTGCAACAGCATCTTACACTGAGGGGAGAAAGATAGTCAGCTTGAACTGGACTTTAGATA
CAACAGATTTGTAAAGGCTGTCGAGGGTTGCTGAAGGAAAGGGAATGGGCTTGTTTGACTTAAGGGCATGGCTGA
AAGTAGAATATTGAAACAGCTGTGTGTTTCCAGGTTTTGCCATTCTCATTAGAGAATTTGTAACTAGTGAAAACT
GTGAAAACAAAACCAACTGTTCAAGTACTGTGTTCAGTGAGTTGTTAAATTCCGTCTTGTTTAGCCAGAAAGTCA
ACATTCCATGTGCTCTGCTGGTTCAAATAACAATGAGTCAGCGAGGCCAGAACAGGTGATATTATAATGTGGATT
AATAATGTGGAAGGTGTCAGTTTTTTACTGGACCATTCTTGGCCGTATGGGGGGCCTACTTACAACTTTTATACA
TTCACACTCCTTCCCCCGCACCACTCTCTCTCACGCACACAGACTCACAGACACACACACACACACACAGCTAGT
TGTGGATCTGGATATAAGGCTGTATTTTCCACATTTACTGCCGTTTGTCACAGCCTCTCATTATGCTGTGTGTGT
GTGTGTGTGTGTGTTCCCTGCAG**CTAAAGGTGGTCCGTCGGGAGGTGGAGCTATCTCGGCGGAGGAGCCTCAAGC**
TAAAAGCACAGGTGGAGAAACTCCATAACCGGAGTACCTCGGACTGGACCCAGCAGAGACCACAGGTACTGCAGG
AGGGACCACGTGCTGGCTTTCACTGACTAACGTTTTCAGTTCATACTACAGGACATAATGGTGTCTAAAATAATA
GATTGGCAATTATTCACTTTCATTAGGACGGGACATACTGCTACATTCACTGGAATAACACAATTGAATGAATTT
AGTTTGTTAGGGTGTGTTGAGACACATGACGCTGACCCTAGAACTGATGTGGTCTTGATCAGTGTAGGATCTGTT
CCACAATAAGCTGCTATCAGTGCAGTGAAACTCCCACAGGATGGGAGCAGCCTGTAGGTCAACTGCTTCACTGCT
CTGCAACATTTTTTTTTCTTCTGAAATGGACTTGATGACATTTAATTGGTTTTCTTCATATCAGTTTCCAAAATA
GACTATTTCTCTAGACTCCCAAAGACTCATTATTTTTCTTCTTTTCTAAGTCATTTCCCCCTTCTTTGTCATGTG
TTTTCCCAATGACAGTAATGTCTCATCTCTTT*AGGTGACAGAGGAGGTTCAGTCTCTTCTGAAGCTTC*
TGCTTCCTCTAACGGACGTGGATCCGACCCAGCCGGGCTCCTCTGGTTCTGAGGATCCTCTT
*GATGTGGCCCTGAGCCAGCTGCAGAAGGTGGCCCGCGTCCTGGCCCTGAACCACAC*CAAGGTGA
GGACACACAAGGGGGGTCTGGGTCTGTGTGCATGTGTGTTTGTGTGTGTGTTGTGGATGTGTGTGTGTGTGTGTG
TGTGTGCTGTAGCAGGTAGATTTCAGGTTGTGTAGGATTTACCCAAATATGGGAATAGGTACAGATTAGGACCTT
CTTTCATCAGTCGTCTTGGCTTTACTAAGCCTGCTACAACCCTCGTTTTAGAGATTAAAATCATAA

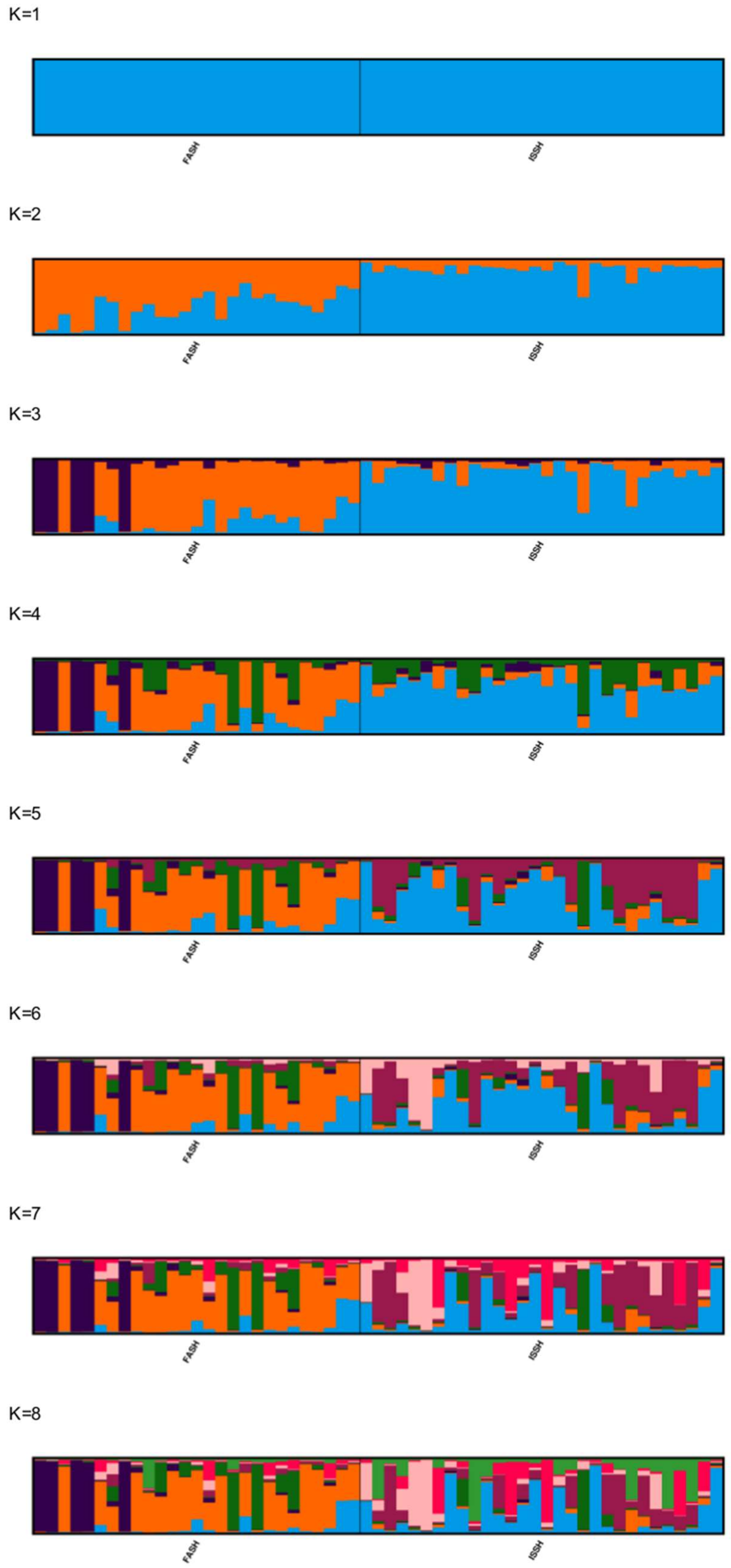
Supplementary material for Manuscript 4.



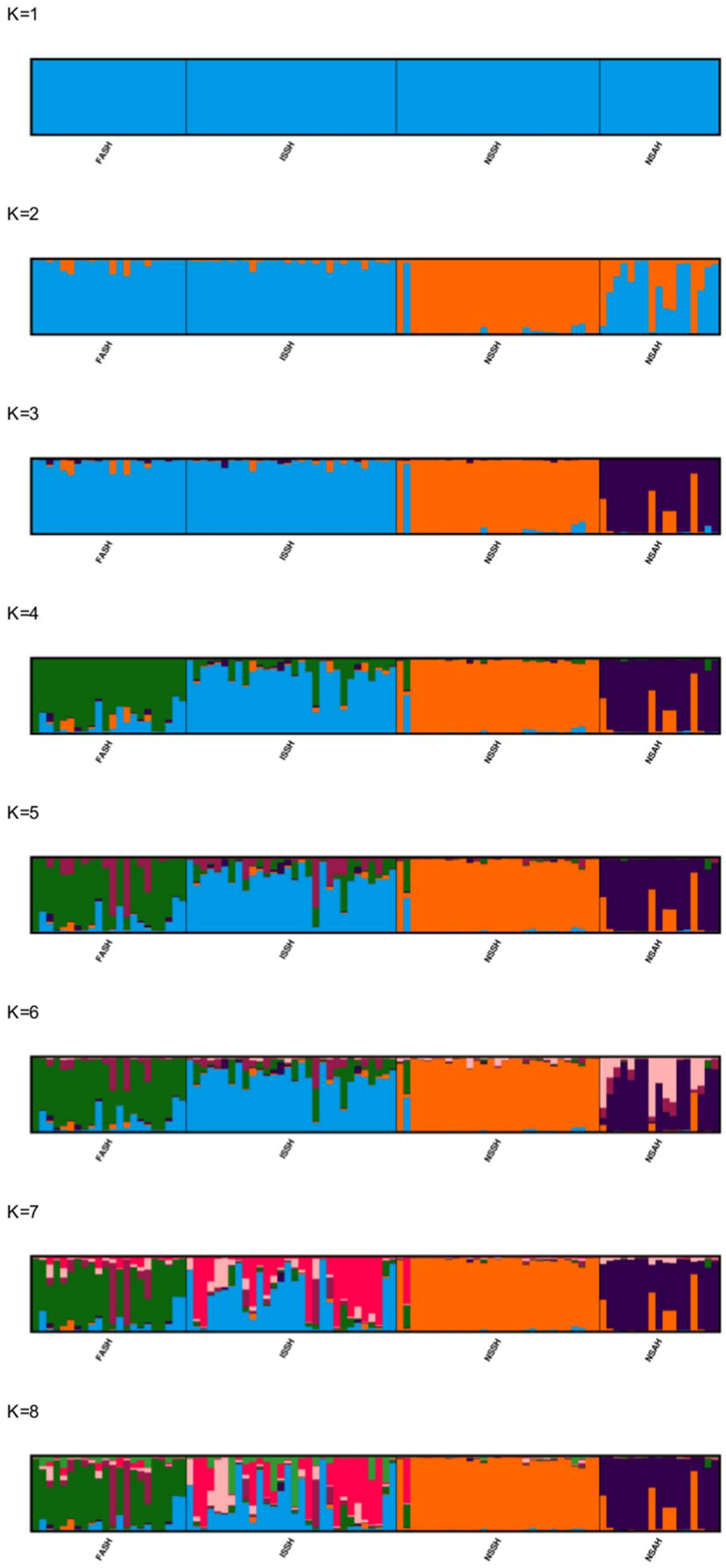
Supplementary Figure S1. Barplots showing the STRUCTURE results for all individuals from the NSSH, NSAH, FASH, and ISSH stocks and K = 1–8.



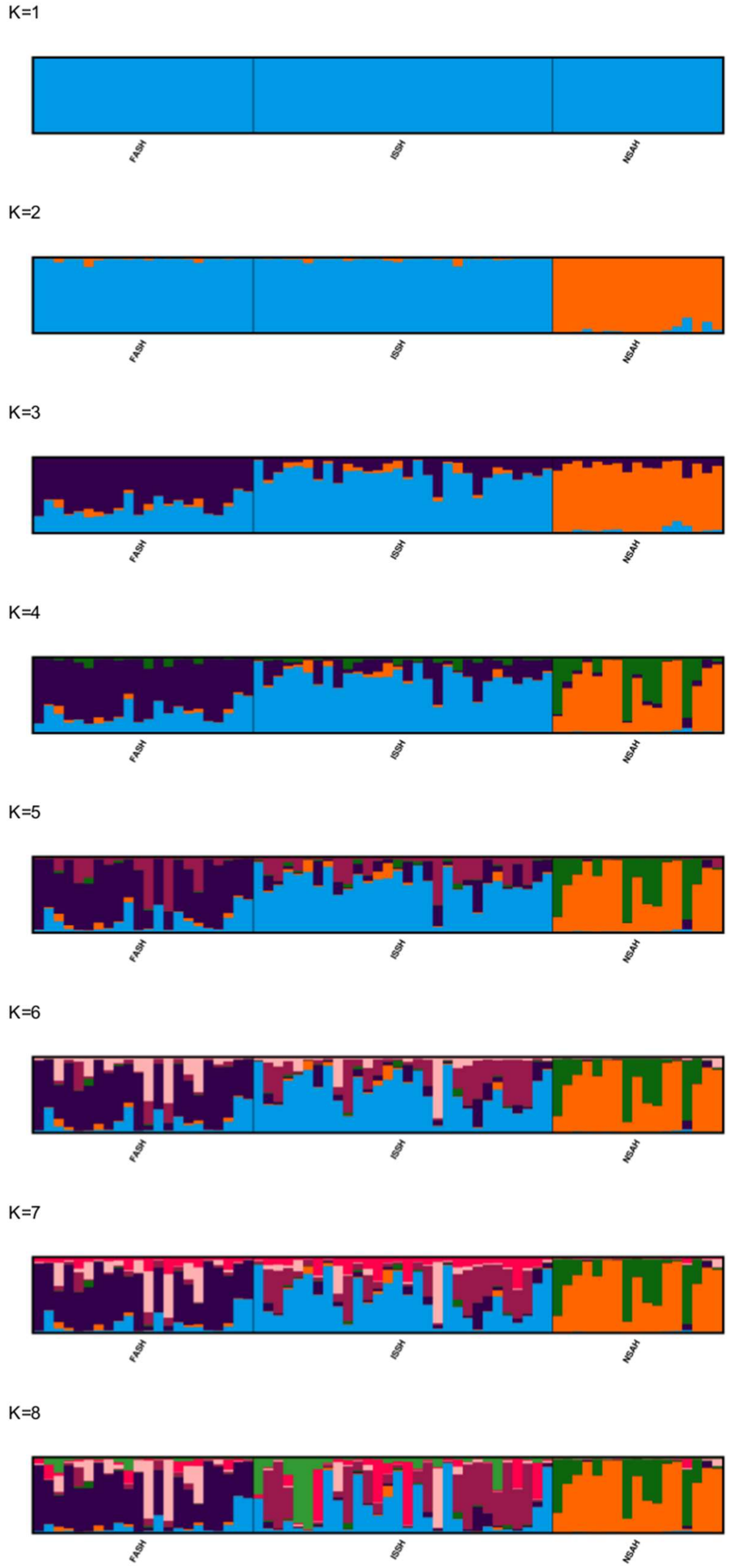
Supplementary Figure S2. Barplots showing the STRUCTURE results for all individuals from the NSAH, FASH, and ISSH stocks and K = 1–8.



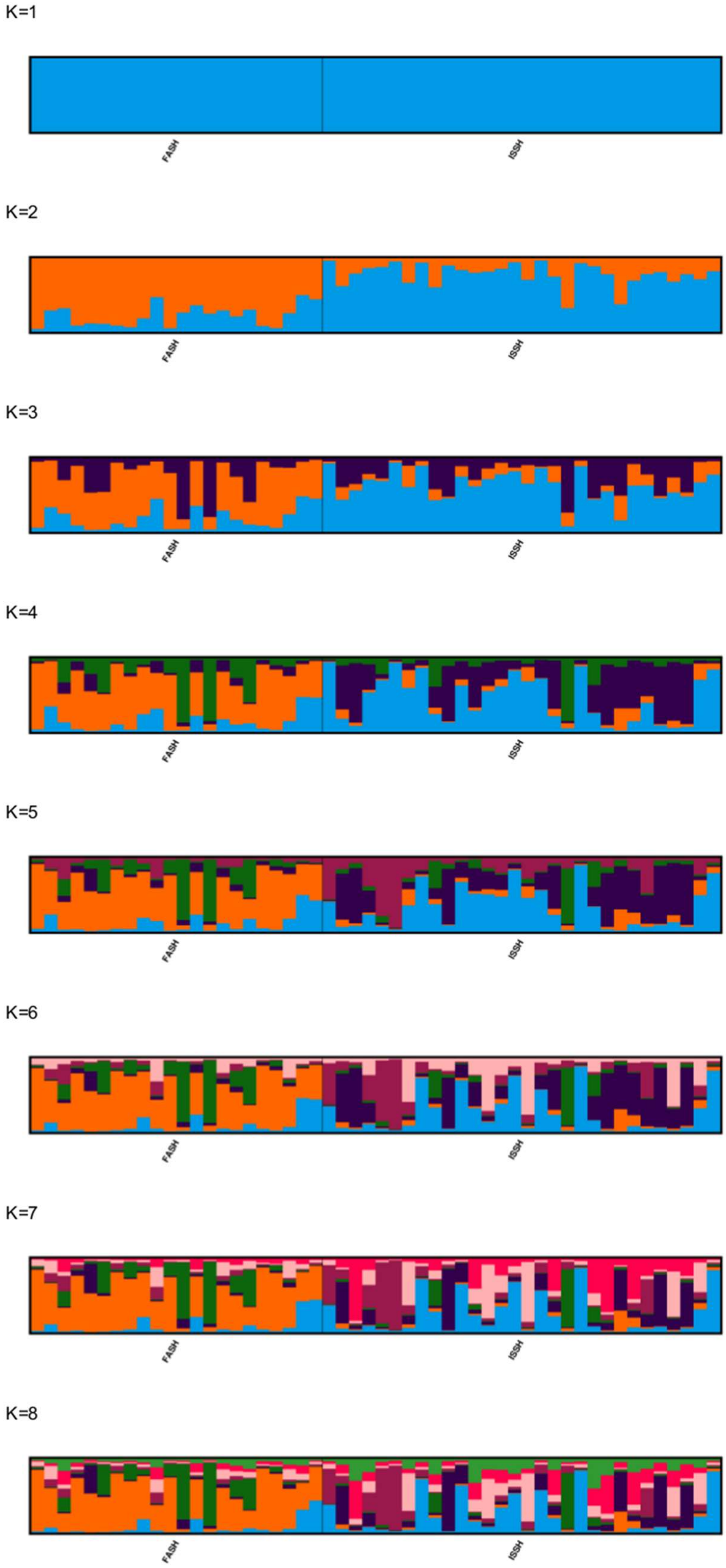
Supplementary Figure S3. Barplots showing the STRUCTURE results for all individuals from the FASH and ISSH stocks and K = 1–8.



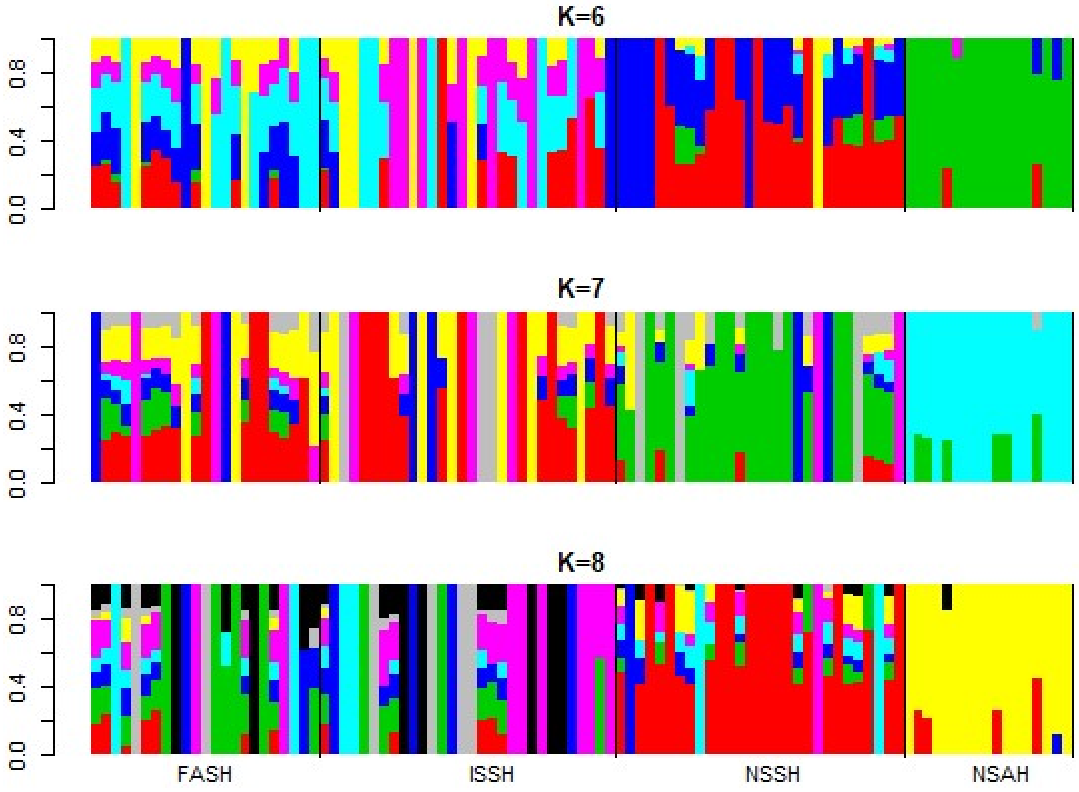
Supplementary Figure S4. Barplots showing the STRUCTURE results for individuals from the NSSH, NSAH, FASH, and ISSH stocks, excluding five suspected migrant herring, and K = 1–8.



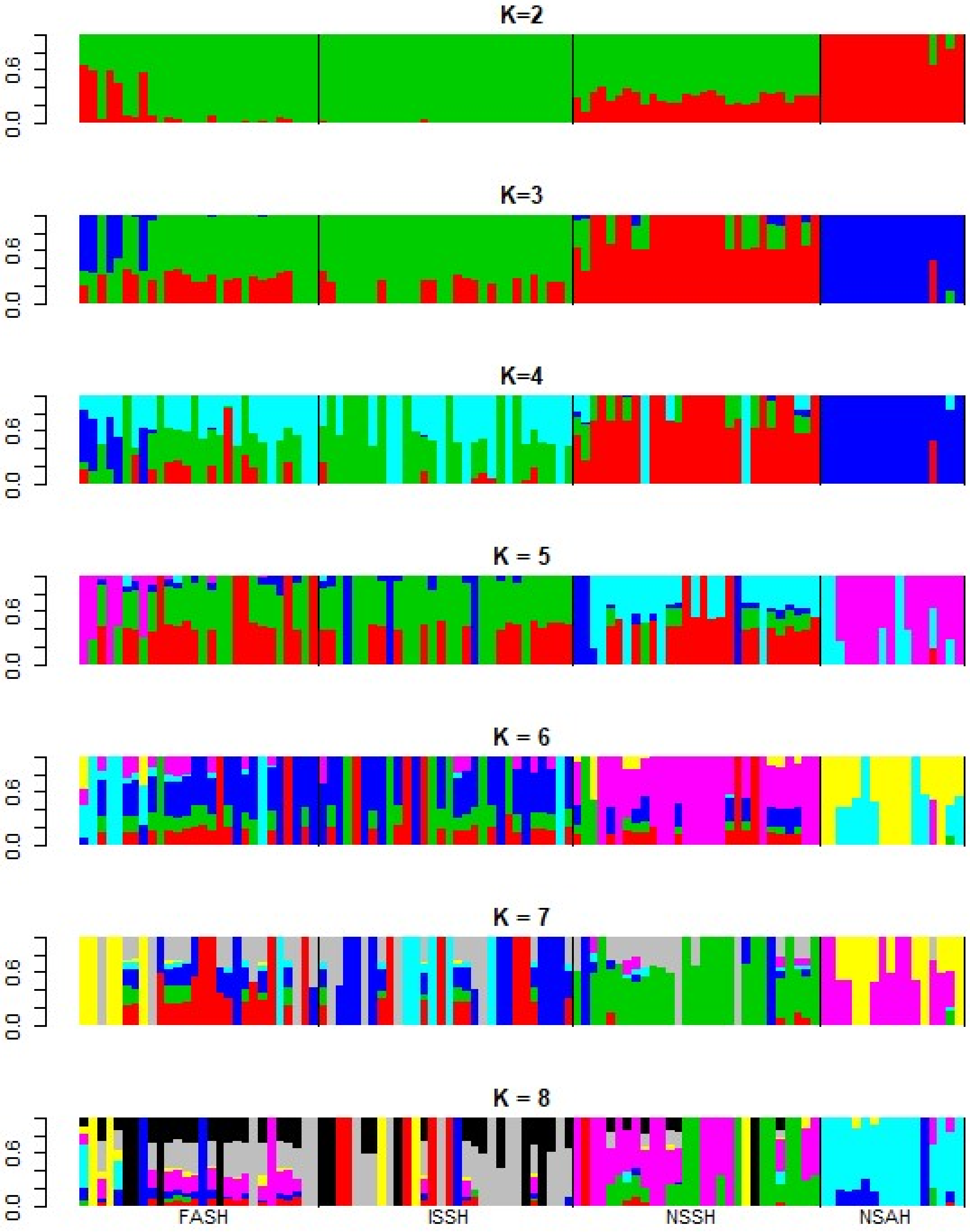
Supplementary Figure S5. Barplots showing the STRUCTURE results for individuals from the NSAH, FASH, and ISSH stocks, excluding five suspected migrant herring, and K = 1–8.



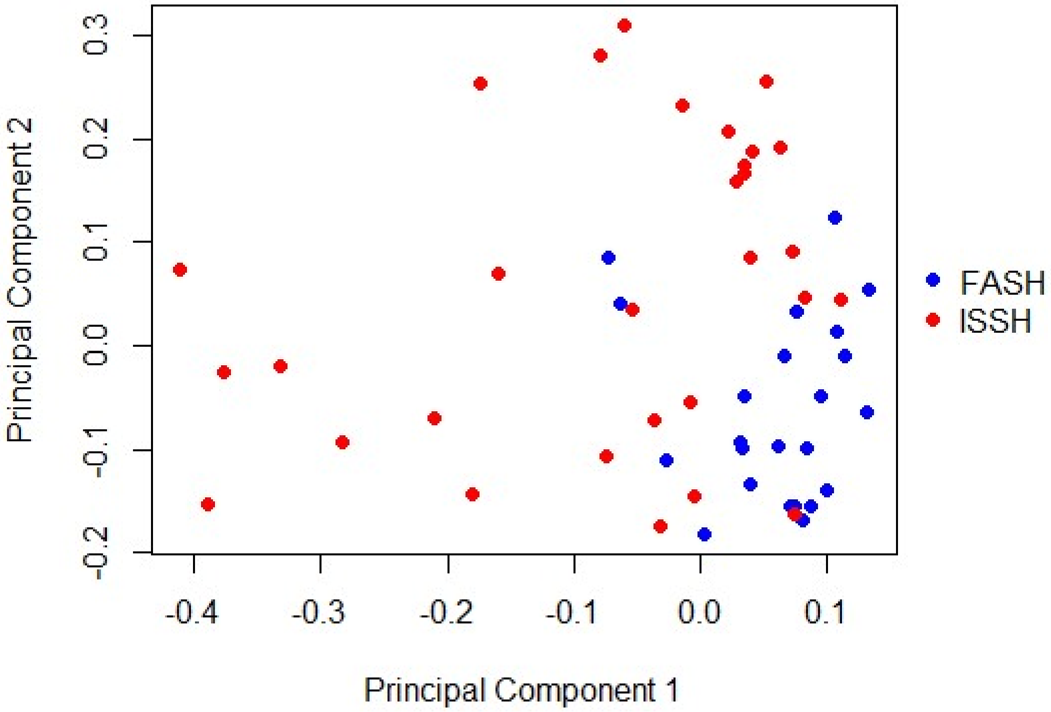
Supplementary Figure S6. Barplots showing the STRUCTURE results for individuals from the FASH and ISSH stocks, excluding five suspected migrant herring, and K = 1–8.



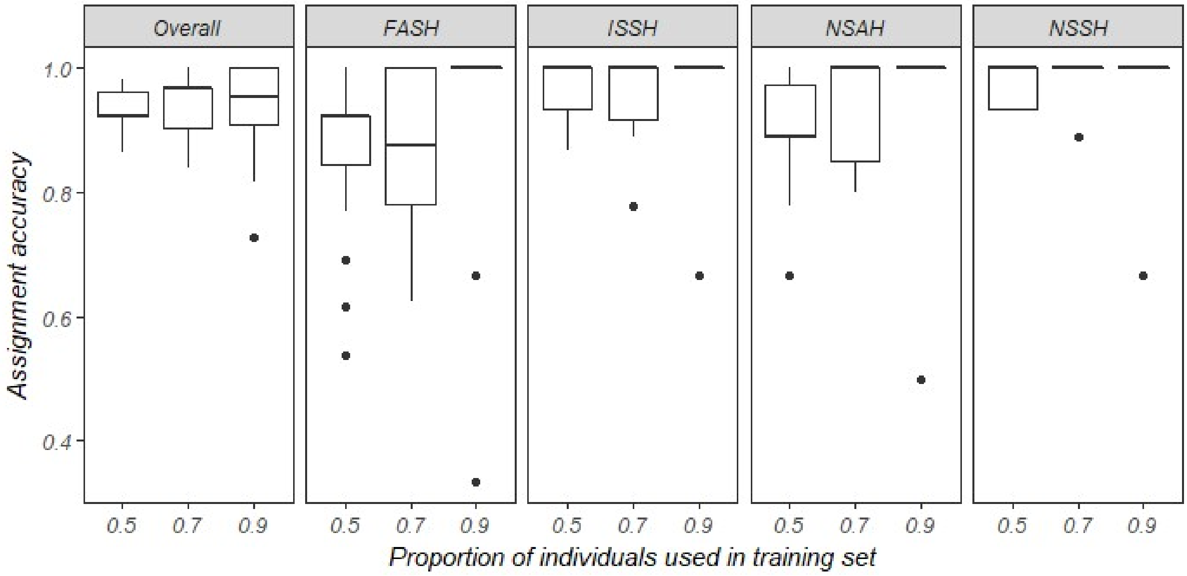
Supplementary Figure S7. Barplots showing the NGSadmix results K = 6–8 for individuals from the NSSH, NSAH, FASH, and ISSH stocks, excluding five suspected migrant herring.



Supplementary Figure S8. Barplots showing the NGSadmix results for K = 2–8 for individuals from the NSSH, NSAH, FASH, and ISSH stocks.



Supplementary Figure S9. Principal component analysis with genotype likelihoods from the FASH and ISSH stocks.



Supplementary Figure S10. Assignment accuracy of Atlantic herring to populations. Ninety tests using Monte-Carlo cross-validation were performed based on 154 SNPs. NSSH = Norwegian spring-spawning herring, NSAH = North Sea autumn-spawning herring, FASH = Faroese autumn spawning herring, and ISSH = Icelandic summer-spawning herring.