Norwegian University of Life Sciences
Faculty of Veterinary Medicine
Department of Basic Sciences and Aquatic Medicine

Philosophiae Doctor (PhD)
Thesis 2020:12

New genomics and transcriptomics tools toward improving conservation strategies for sturgeons

Nye verktøy innen genomikk og transkriptomikk for bedret produksjon av settefisk for stør

Elena Maria Santidrian Yebra-Pimentel

New genomics and transcriptomics tools toward improving conservation strategies for sturgeons

Nye verktøy innen genomikk og transkriptomikk for bedret produksjon av settefisk for stør

Philosophiae Doctor (PhD) Thesis

Elena M. Santidrián Yebra-Pimentel

Norwegian University of Life Sciences
Faculty of Veterinary Medicine
Department of Basic Sciences and Aquatic Medicine

Oslo, 2020

Thesis number 2020:12

ISSN 1894-6402

ISBN 978-82-575-1649-9

Elena M. Santidrian Yebra-Pimentel

Oslo, 2020

Acknowledgements

First of all, I would like to thank the European Union's Horizon 2020 research and innovation programme, which have funded this research under the Marie Skłodowska-Curie grant agreement no. 642893: Improved Production Strategies for Endangered Freshwater Species (IMPRESS).

There is a bunch of people without whom I would not have been able to finish or even start this thesis. I should thank in the first place my supervisors Ron Dirks, Finn-Arne Weltzien and Sylvie Dufour: Ron, thank you for believing that I was capable of performing this research and for helping me out from the beginning, making difficulties look simple and sacrificing afternoons and weekends correcting my work and giving constructive comments. Finn, thank you for providing me with feedback even with short notice and during particularly busy times, and for making me feel welcome during my short stays in Oslo. Sylvie, thank you for helping me taking the most of my (unexpectedly shortened) stay in Paris, and for keeping me fascinated when you talk about evolution with the passion you do.

Thanks to my colleagues at ZF-Screens and FG-Tech, which have always provided me with insane amounts of cookies to keep my brain working (and my butt getting fat). Thanks, Susanne, Hans, Lieke and Yunus for helping me in the lab and with the bioinformatics. Thanks, Karen and Ron for a long ride during the hottest day of the history of The Netherlands to sample that sturgeon. I am especially thankful for that cute and naïve sturgeon who died to provide me with so much data for my thesis, and I hope this thesis contributes to give something back to their conspecifics. Thank you, Peti for encouraging me to start this PhD and for all the help in the previous zebrafish work. Also, thanks to Ron, Susanne, Hans, Lieke, Yunus, Karen, Martje, Anita, Peti, Bas, Nikola and Wouter for nice conversations during lunch, tea, or beer breaks, and for making daily lab and office life more fun.

Thanks, Eirill for organizing the IMPRESS workshops that were so productive and during which the amazing ESRs team could meet. With some I have directly shared experiments or research topics, with all I have shared beers, nice chats and jokes: thanks Aurora, Ben, Carina, Carlos, Chris, Daan, Elia, German, Hannah, Ishwar, Maria, Mitch, Rasheed, Sahana and Sophia. Also, thanks to the PIs and partners of the IMPRESS Project, especially Berta, Vojta, Sven, Jörn and Avshalom, for nice discussions, arranging, facilitating and/or helping

during samplings. I would also like to thank all co-authors and reviewers for providing constructive comments that helped to present my work in a clearer manner. A special thank to the evaluation committee members for being (likely) the only people on Earth (besides myself, my supervisors and people from other planets) reading my whole thesis.

Thank you Chris Henkel for always being willing to discuss with me about the sturgeon genome in a friendly environment, and Gersende Maugars for showing interest in my project and providing me with the huge gonadotropin alignment files that I haven't used yet.

I would also like to thank all the people working at the Genetics Department of the Veterinary Faculty in Lugo for welcoming me to their great team and family back in 2012, introducing me to research and keeping me motivated. Special thanks to Laura Sánchez, Jorge Guerra, Vanessa Pérez and Pablo Cabezas for all the awesome moments in and outside the lab.

Gracias a mis padres y hermanas por su apoyo, ánimo y confianza en mis habilidades , y por siempre tratar de mantenerme optimista. Gracias a Inés por ser mi primera compañera de "laboratorio" haciendo los experimentos de *Coñecemento do medio* sugeridos al final de cada tema. También quiero agradecer a mis tías y tíos, particularmente a mi tía Eva que desde pequeña identificó y cultivó en mi a una pequeña científica y en cada problema me guió, quizás sin saberlo, hacia las preguntas adecuadas, para poder encontrar mis propias respuestas. Gracias a mis abuelas y abuelos, en especial a mi abuelo Fermín por su particular sentido del humor que, aunque diluido en las meiosis, creo haber heredado algún buen alelo, y a mi abuela Eva porque quince minutos con ella mantienen mi autoestima por las nubes por quince meses. Gracias a Pin, Santos y Paula por integrarme en su familia y preocuparse por mi.

I would also like to thank the friends that are far: Xulia, Belén, Sara, Eva, Candela, Carlos Megías, Carlos Monje, Tania, Cristina y Adrián for making my life enjoyable and for rarely have asked me how far I was with my thesis. I would also like to thank all the newer "wall" and "non-wall" friends, especially Charlotte, Lucas, Diego, Francesca, André, Jacopo and Didac, for reminding me that there is a life beyond my computer screen.

And finally, thanks Simón for taking care of each other, for filling my life with joy, for all the shared adventures and challenges along the way, and the many more to come.

Table of Contents

Summary

Sturgeons (Family *Acipenseridae*) are one of the largest and most primitive fish families on Earth. Although they have always been typically distributed throughout the Northern Hemisphere, during the last decades wild populations have declined due to anthropogenic factors such as overfishing, poaching, pollution, and habitat loss. The situation is particularly dramatic for Atlantic sturgeon, one of the most ancient species among the family, which is currently extinct in Europe. In order to reintroduce the Atlantic sturgeon in Europe, several Baltic countries have been working together for more than two decades to build an *ex-situ* broodstock locally with fish derived from Canada, where the populations are not threatened, and releasing juveniles into the rivers flowing to the Baltic Sea. However, rearing fish aimed for restocking in the same manner as aquacultured fish has been shown to impact the post-release survival of juveniles in the long term. For example, in aquaculture conditions, fish are often maintained in high densities, at constant photoperiod and temperature conditions, and feeding on commercial pellets, leaving them cognitively naïve when released into natural environments. Additionally, increased water temperatures due to global warming have a strong influence on the geographic distribution of the species, resulting in local extinctions and population shifts. Although the effect of heat and cold stress on the juvenile and embryonic development have been assessed for several sturgeon species, most studies on gene expression have looked at a very limited number of genes due to the lack of sequence information and genomic resources. Also, most studies use other sturgeon species such as Siberian, Japanese, and white sturgeons, which are widespread aquacultured species. Exposing fish to temperatures higher than the optimal can trigger phenotypic adaptations leading to increase thermotolerance and potentially improve post-release survival, however the impact of temperature-training protocols on the response to a subsequent heat shock has not yet been assessed in sturgeons.

Therefore, the main aim of this thesis was to generate genomic and transcriptomic resources for Atlantic sturgeon, which are essential to improve and promote research in many fields, such as ecology, physiology and evolutionary studies. Moreover, it provides a reference for RNAseq-mediated transcriptome mapping. Additionally, we have used these resources to develop and evaluate the impact of novel rearing techniques toward improving restoration success, focusing on temperature training. First, we have assembled a high-quality *de novo*

transcriptome, made an inventory of all the heat shock protein (HSP) gene family members and exposed a cell line derived from Atlantic sturgeon larvae to a moderate and severe heat shock in order to identify all heat-responsive genes using an RNAseq approach (**Paper I**). We found 76 HSP genes in the Atlantic sturgeon transcriptome, only 16 of which were responsive to at least one of the applied heat shock protocols, and only 5 of which were consistently upregulated after both moderate and severe heat shock at all the tested timepoints. After building the reference transcriptome and annotating all the HSP genes, we have evaluated the differences in liver transcriptome between temperature-trained and non-trained juveniles upon exposure to a new heat shock (**Paper II**). After four weeks of treatment, fish exposed to temperature training showed between 2 to 4 fold less dysregulated genes in response to a new heat shock than the non-trained group, indicating their improved ability to maintain transcriptomic homeostasis during a new heat shock. Again, like in the *in vitro* experiment, very few of the annotated HSP genes were dysregulated in response to heat shock in the liver transcriptome, namely *hspa1*, *hspc1* and *dnajb4*. Overall, the response to heat shock in the liver transcriptome was milder than the *in vitro* response, which is likely a consequence of the activation of compensatory mechanisms. These mechanisms include the neuroendocrine system and result in increased tissue protection and thermogenic capacities, especially in the trained fish. We therefore propose that temperature-training protocols like the one tested in this thesis should be included in the set of new rearing techniques for fish used for restocking; however, other protocols should be investigated.

Since the main bottleneck in the evaluation of the effect of such training is the lack of sequence information and a reference genome for RNAseq experiments, we have additionally assembled a reference genome for Atlantic sturgeon using a combination of short and long-read sequencing technologies (**Paper III**). The assembled genome provides for the first time clear evidence of a sturgeon-specific whole-genome duplication event (SR), independent from the American paddlefish (*Polyodon spathula*), which is the main representative of the sister Family (*Polyodontidae*) within the same Order (*Acipensiformes*). The presence of duplicated Hox clusters, together with synteny and phylogenetic studies of these developmental genes, and the results of microsatellite loci analysis, suggests that sturgeons have a paleotetraploid origin, and that a rediploidization process is still ongoing.

In summary, the results presented in this thesis advance the field of sturgeon research. We hypothesized that temperature training has a positive effect during the exposure to a

subsequent heat shock, but its potential to improve post-release survival in the long term should still be assessed. We therefore suggest that future work should be aimed at the optimization of rearing methods for stocking programs and that a reference genome should be used.

Sammendrag

Stør (Familie *Acipenseridae*) er blant de største og mest primitive familier av benfisk. Mens de opprinnelig var utbredt over hele den nordlige halvkule, er mange populasjoner nå kritisk truet på grunn av menneskeskapte faktorer som overfiske, forurensning og tap av habitat. Situasjonen er spesielt dramatisk for atlantisk stør, en av de eldste artene i familien, som er utdødd i Europa. For å gjeninnføre atlantisk stør, har flere baltiske land arbeidet for å etablere en *ex situ* stamfiskpopulasjon basert på fisk fra Canada (hvor bestanden ikke er truet), for produksjon av yngel til utsetting i baltiske vassdrag. Imidlertid gir produksjon av settefisk etter samme prinsipper som for oppdrettsfisk svært lav overlevelse i naturen. Eksempelvis vil høy tetthet, konstant fotoperiode og vanntemperatur, og fôring med pellets til faste tider gi en kognitivt naiv fisk som ikke klarer seg i det fri. I tillegg kommer økte vanntemperaturer som resultat av global oppvarming, og som har stor innvirkning på den geografiske fordelingen av arter, som igjen resulterer i lokal utryddelse og forflytning av populasjoner. Selv om effekter av vanntemperatur har blitt undersøkt på embryonal- og yngelutviklingen i flere størarter, har de fleste studier fokusert på et lite sett av gener fordi genomiske ressurser har manglet. I tillegg er de fleste studiene utført på andre arter som er vanlig i akvakultur, slik som sibirsk, japansk og hvit stør. Eksponering til vanntemperaturer som er høyere enn artens optimum, kan utløse fenotypiske tilpasninger som fører til økt termotoleranse og potensielt forbedre overlevelse i naturen. Men effekten av slike temperaturtreningsprotokoller på responsen på et påfølgende varmesjokk er ikke studert i stør.

Hovedmålet med denne avhandlingen var å generere genomiske og transkriptomiske ressurser for atlantisk stør, som en viktig faktor for å forbedre og fremme forskning innen økologi, fysiologi og evolusjon. Videre gir avhandlingen en referanse for RNAseq-mediert transkriptomkartlegging. Disse ressursene er så benyttet til å utvikle og evaluere virkningen av nye oppdrettsteknikker for settefisk, med fokus på temperaturtrening. Først har vi satt sammen et høykvalitets *de novo* transkriptom, deretter karakterisert genfamilien av varmesjokkproteiner (HSP), og så eksponert en cellelinje avledet fra atlantiske størlarver for et moderat og et kraftig varmesjokk for å identifisere alle varmeresponsive gener ved bruk av RNAseq (**Artikkel I**).

Vi fant 76 HSP-gener i transkriptomene fra atlantisk stør, hvorav 16 responderte på minst en av de testede varmesjokkprotokollene, og bare 5 av disse ble konsekvent oppregulert etter både moderat og kraftig varmesjokk ved alle testede tidspunkt. Etter å ha bygget referanse-transkriptomet og karakterisert alle HSP-gener, evaluerte vi forskjellene i levertranskriptom mellom temperaturtrent og ikke-trent yngel ved eksponering for et nytt varmesjokk (**Artikkel II**). Etter fire ukers behandling viste fisk som ble utsatt for temperaturtrening mellom 2-4 ganger færre dysregulerte gener som svar på et nytt varmesjokk sammenlignet med den ikke-trente gruppen. Dette antyder en forbedret evne til å opprettholde transkriptomisk homeostase under et nytt varmesjokk. I likhet med *in vitro* eksperimentet var svært få HSP-gener dysregulert som respons på varmesjokk i levertranskriptomet, nemlig *hspa1*, *hspc1* og *dnajb4*. Totalt sett var responsen på varmesjokk i levertranskriptomet mildere enn *in vitro* responsen, noe som sannsynligvis skyldes ulike kompensatoriske mekanismer. Disse inkluderer det nevroendokrine systemet og resulterer i økt vevsbeskyttelse og termogen kapasitet, spesielt i trent fisk. Selv om andre protokoller bør undersøkes nærmere, foreslår vi at protokoller for temperaturtrening lik den som ble testet i denne avhandlingen bør vurderes i nye oppdrettsprotokoller for settefiskproduksjon av stør.

Siden den viktigste flaskehalsen i evalueringen av effekten av slik trening er mangelen på sekvensinformasjon og et referansegenom for RNAseq-eksperimenter, har vi i tillegg produsert et referansegenom for atlantisk stør ved bruk av ulike sekvenseringsteknologier for korte og lange reads (**Artikkel III**). Genomet gir for første gang klare bevis for en stør-spesifikk helgenomdupliseringshendelse (SR), uavhengig av spadestør (*Polyodon spathula*), som er hovedrepresentanten for søsterfamilien (*Polyodontidae*) innenfor samme Orden (*Acipensiformes*).

Tilstedeværelsen av dupliserte klynger av hox-gener som er sentrale i tidlig utvikling, i tillegg til hox-gen synteni og fylogeni, og mikrosatellitt loci-analyser, antyder at stør har en paleotetraploid opprinnelse, og at en rediploidiseringsprosess fortsatt pågår.

Oppsummert vil de forbedrete genomiske og transkriptomiske verktøy presentert i denne avhandlingen åpne for nye muligheter innen størforskning. Videre har temperaturtrening en positiv effekt ved eksponeringen til varmesjokk, men potensialet for temperaturtrening og dermed økt overlevelse ved utsett bør undersøkes videre.

Resumo

Os esturións (Familia *Acipenseridae*) pertencen a unha das familias de peixes mais grandes e primitivas da Terra. Aínda que sempre estiveron distribuídos no hemisferio Norte, nas últimas décadas as poboacións salvaxes teñen diminuido debido a factores antropoxénicos como a sobrepesca, a caza furtiva, a contaminación e a perda do hábitat. A situación é especialmente dramática para o esturión Atlántico, unha das especies máis antigas da familia que se atopa extinta en Europa na actualidade. Co fin de reintroducir o esturión Atlántico en Europa, varios países bálticos levan traballando xuntos durante máis de dúas décadas para construír un núcleo reprodutor con peixes derivados de Canadá, onde a poboación non esta ameazada, e liberar xuvenís nos ríos que flúen ao Mar Báltico. Non obstante, cultivar peixe destinado ao repoboamento utilizando as mesmas técnicas típicamente utilizadas en acuicultura afecta negativamente á supervivencia a longo prazo. Por exemplo, en condicións de acuicultura os peixes adoitan producirse en altas densidades, baixo condicións abióticas constantes (fotoperíodo e temperatura), e aliméntanse de *pellets* comerciais, producindo animais congitivamente inxenuos cuando son libertados no ambiente natural. Ademais, o aumento das temperaturas da auga debido ao quentamento global ten unha forte influencia na distribución xeográfica das especies, dando lugar a extincións locais. Aínda que os efectos do estrés térmico no desenvolvemento embrionario e etapas xuvenís xa teñen sido avaliados en varias especies de esturións, a maioría dos estudos de expresión xénica céntranse en un número moi limitado de xenes debido á falta de recursos xenómicos. Ademáis, a maioría dos estudos utilizan outras especies de esturións como o branco, siberiano ou xaponés, mais comúns en acuicultura. A exposición dos peixes a temperaturas superiores ás óptimas pode desecandear adaptacións fenotípicas resultando nun incremento da tolerancia térmica e potencialmente unha mellora da supervivencia no hábitat onde son libertados. Sen embargo, os efectos do réxime de temperatura utilizado durante a cría na resposta a un choque de calor posterior non teñen sido avaliados.

Por todo isto, o obxectivo principal desta tese foi xerar recursos xenómicos e transcriptómicos para o esturión Atlántico, esenciais para mellorar e promover a investigación en moitos campos da ciencia como a ecoloxía, fisioloxía e a xenómica evolutiva, ademais de fornecer unha referencia para o mapeado do transcriptoma. Ademais, empregamos estes recursos para desenvolver e avaliar o impacto de novas técnicas de

cultivo para mellorar o proceso de repoboación, centrándose na xeración de individuos termotolerantes.

En primeiro lugar, temos ensamblado un transcriptoma de alta calidade, fixemos un inventario de todos os membros da familia das proteínas de choque térmico (HSP) e expuxemos unha liña celular isolada a partir de larvas disgregadas de esturión Atlántico a un choque de calor moderado e severo para identificar xenes sensibles ó calor (**Artigo I**). Atopamos 76 HSP no transcriptoma de esturión Atlántico, dos cuáis só 16 foron sensibles a polo menos un dos protocolos de choque de calor avalidado, e 5 foron sensibles a ambos choques térmicos idenpendentemene do momento da amostraxe.

Utilizando as secuencias dos xens HSP e o trancriptoma de esturión ensamblado no **Artigo I** como referencia, temos avaliado as diferenzas no transcriptoma hepático en resposta a un choque térmico entre xuvenís criados baixo un réxime de temperatura constante e en réxime fluctuante (**Artigo II**). Despois de catro semanas de tratamento, os peixes criados en réxime de temperatura fluctuante mostraron entre 2 e 4 veces menos xens diferencialmente expresados en resposta a un novo choque térmico que os peixes criados en réxime de temperatura constante, indicando a súa mellor capacidade para manter a homeostase transcriptómica durante un novo choque térmico. Como xa indicaron os resultados *in vitro*, moi poucos HSP foron diferencialmente expresados en resposta ó choque de calor no transcriptoma hepático, concretamente *hspa1*, *hspc1* e *dnajb4*. En xeral, a resposta ao choque térmico no transcriptoma hepático foi máis leve que a resposta *in vitro*, o que é probablemente consecuencia da activación de mecanismos compensatorios. Estes mecanismos inclúen o sistema neuroendocrino e teñen como consecuencia un aumento da protección dos tecidos e das capacidades termoxénicas, especialmente no peixe criado a temperaturas fluctuantes. Polo tanto, propoñemos que a cría de peixes a temperaturas fluctuantes debe ser incluida no conxunto de novas técnicas empegadas en peixes criados con fins de repoboamento, non obstante, outros protocolos de temperatura deben ser inestigados.

Dado que unha importante limitación para a avaliación de novas técnicas de cría é a falta dun xenoma de referencia para experimentos de mapeado do transcriptoma, no **Artigo III** temos ensamblado un xenoma de referencia para o esturión Atlántico, combinando tecnoloxías de secuenciación de curta e longa lectura. O xenoma do esturión Atlántico evidencia por primeira vez a presenza dun evento de duplicación específico de esturión (SR)

e independente do peixe-espátula (*Polyodon spathula*). A presenza de xenes Hox duplicados, xunto con estudos filoxenéticos e de sintenia e os resultados da análise de *loci* microsatélite suxire que os esturións teñen unha orixe paleotetraploide e que a rediploidización é aínda un proceso activo.

En resumo, os resultados presentados nesta tese avanzan no campo da investigación con esturións. Os nosos resultados suxiren que a cría de peixes baixo un réxime de temperatura fluctuante ten un efecto positivo durante un choque térmico subsecuente, pero a influencia de esta nova técnica de cría na supervivencia dos xuvenís tras a súa liberación no habitat a repoboar aínda debe ser avaliada. Contudo, suxerimos que o traballo no futuro ten que estar centrando na optimización dos métodos de cría en programas de repoboación e que o xenoma de referencia debe ser usado.

List of publications

Paper I

Deep transcriptome analysis of the heat shock response of Atlantic sturgeon (*Acipenser oxyrinchus)* cell line.

Yebra-Pimentel, E.S.; Gebert, M.; Jansen, J.H.; Jong-Raadsen, S.A.; Dirks, R.P.H.

Fish and Shellfish Immunology 88 (2019) 508-517.

https://doi.org/10.1016/j.fsi.2019.03.014

Paper II

Temperature training improves transcriptional homeostasis after heat shock in juvenile Atlantic sturgeon (*Acipenser oxyrinchus*)

Yebra-Pimentel, E.S.; Reis, B; Gessner, J.; Wuertz, S; Dirks, R.P.H.

Submitted to Fish and Shellfish Immunology.

Paper III

The genome of the Atlantic sturgeon (*Acipenser oxyrinchus*) provides insights into genome evolution in tetraploids

Yebra-Pimentel, E.S.; Campo, A.; Shivaramu, S.; Henkel, C.V.; Jong-Raadsen, S.A.; Jansen, H.J.; Dirks, R.P.H.

Manuscript

Abbreviations

ASIC	**Application specific integrated system**
bp	**Base pair/basepairs**
bps	**Bases per second**
ccdc7	**Coiled-coil domain containing protein 17**
cDNA	**Complementary DNA**
CI	**Confidence interval**
clu	**Clusterin**
DBG	**De Bruijn graphs**
DNA	**Deoxyribonucleic acid**
dsDNA	**Double stranded deoxyribonucleic acid**
EDCs	**Endocrine disrupting chemicals**
EDTA	**Ethylenediaminetetraacetic acid**
FACS	**Fluorescence-activated cell sorting**
FAO	**Food and Agriculture Organization**
FCS	**Fetal calf serum**
gadd45	**Growth arrest and dna damage inducible protein 45g**
Gb	**Gigabase**
gck	**Glucokinase**
gDNA	**Genomic DNA**
GO	**Gene ontology**
grb2	**Growth factor receptor-bound protein 2**
HS	**Heat shock**
HSF	**Heat shock factor**
HSP	**Heat shock protein**
HSR	**Heat shock response**
IPCC	**Intergovernmental panel on climate change**
IUCN	**International Union for Conservation of Nature**
kb	**Kilo base/kilobases**
Mb	**Megabase**
mRNA	**Messenger ribonucleic acid**
mtDNA	**Mitochondrial DNA**
Mya	**Million years ago**
nt	**Nucleotide**
N_{50}	**Median assembly length**
NCBI	**National Center for Biotechnology Information**
NGS	**Next generation sequencing**

NOR	**Nuclear organizer regions**
npr2	**Atrial natriuretic peptide receptor 2**
OLC	**Overlap-layout-consensus**
ONT	**Oxford Nanopore Technologies**
P22	**Passage 22**
PBS	**Phosphate-buffered saline**
pck1	**Phosphoenolpyruvate carboxykinase**
PCR	**Polymerase chain reaction**
PFGE	**Pulsed-field gel electrophoresis**
Q	**Read quality**
qPCR	**Quantitative polymerase chain reaction**
QTL	**Quantitative trait loci**
RNA	**Ribonucleic acid**
RNAseq	**Ribonucleic acid sequencing**
RPKM	**Reads per kilobase and million**
SAM	**Sequence alignment Map**
SR	**Sturgeon-specific genome duplication**
PR	**Paddlefish-specific genome duplication**
t	**Metric ton**
TGS/ 3R	**Teleost-specific genome duplication**
TULIP	**The Uncorrected Long read Integration Process**
WGD	**Whole genome duplication**
WGR	**Whole genome resequencing**

Introduction

General background

Based on fossil records dated from the Lower Jurassic (~ 200 Mya), sturgeons are one of the most primitive fish on earth, which were depicted in Egyptian temples and frequently appeared in coins and postage stamps. They are considered very curious fish that occasionally leap out of the water, and native North Americans used to wait for them on the water surface with a torch to catch them. In addition to their meat and roe, their swim bladder was used as an adhesive and paint-binding agent, and leather used to be made from their skin. Sturgeons are among the largest freshwater fish and, like sharks, they have a heterocercal tail, and their skeleton consists almost exclusively of cartilage. Additionally, instead of scales, their body is covered with bony plates called scutes. Due to their high morphological stasis throughout evolution, they have been largely considered "living fossils" (Bemis et al., 1997; Krieger and Fuerst, 2002).

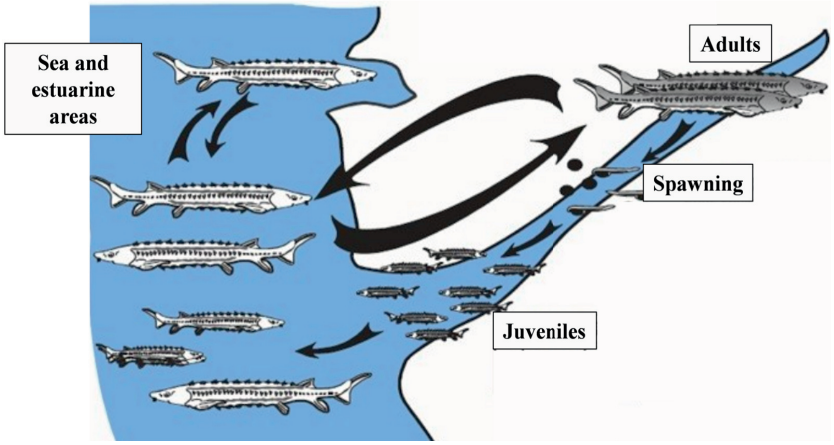


Figure 1. Generalized life cycle of the Atlantic sturgeon (*Acipenser oxyrinchus*). Illustration modified from Government of Canada.

Sturgeons feed as opportunistic benthivores and have therefore developed some morphological adaptations: they have a flat belly and a siphon-like protractile mouth with barbels, which contain taste buds to search for prey hidden in the mud. They are currently distributed throughout the Northern Hemisphere and most of them show an anadromous life cycle: larvae hatch and develop into juveniles in rapid-moving freshwaters, where they

remain up to 5 years until they initiate their migration to the sea and estuarine areas and become sexually mature. During the reproductive season, which can happen as frequently as once per year in some species like sterlet and as rarely as once per decade in white sturgeon, they show homing behavior, migrating upriver to spawn in the same rivers where they hatched (**Fig. 1**).

Their generation time is generally quite long, and their life expectancy can be up to 100 years. In addition, sexual maturation occurs late in life, taking 4-35 years in females and 2-24 years in males, depending highly on the species (Billard and Lecointre, 2000). With the exception of some species, adults lack sexual dimorphism (Keyvanshokooh and Gharaei, 2010). They are categorized as bony fish (*Osteichtyes*), the largest vertebrate group, which diverged 420 million years ago (Mya) into two classes: *Actinopterygii* (ray-finned fish) and *Sarcopterygii* (lobe-finned fish) (Betancur et al., 2013). Sarcopterygians have an important evolutionary role as ancestors of amphibians and tetrapods (Swartz, 2012); however, Actinopterygians are the dominant class of vertebrates and comprise nearly 99% of the extant fish species. Actinopterygians are divided into two subclasses: *Chondrostei* and *Neopterygii*, and in turn, *Neopterygii* are divided into *Holostei* and *Teleostei*. Sturgeons are chondrosteans of the order *Acipensiformes* that diverged from teleosteans 250 Mya (Betancur et al., 2013), occupying a leading position in evolution (**Fig. 2**).

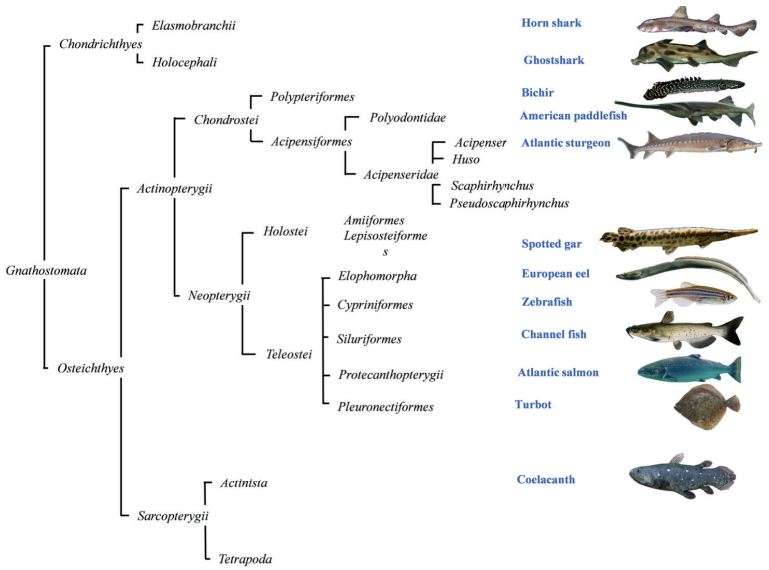


Figure 2. Phylogenetic relationship of sturgeons among gnathostomata (jawed vertebrates).

It was for long debated whether they were part of the sister or basal clade to teleosts (Inoue et al., 2003; Kikugawa et al., 2004), while they have been recently considered as a sister clade to Neopterygians (Near et al., 2012). Either way, because of their phylogenetic position, they have been used as a key outgroup taxon for studies investigating the evolution of teleosts (Metscher and Ahlberg, 1999). The Order *Acipensiformes* contains 27 species divided into two families that diverged between 184.4 and 204.1 Mya (Peng et al., 2007; Luo et al., 2019): *Acipenseridae* and *Polyodontidae*. While the *Polyodontidae* family contains only two extant species (*Polyodon spathula* and *Psephurus gladius*), the *Acipenseridae* family includes four genera (*Acipenser, Huso, Scaphirhynchus,* and *Pseudoscaphirhynchus),* and 25 species with adult weights ranging from ~16 kg (sterlet, *Acipenser ruthenus*) up to 1,500 kg for beluga sturgeon (*H. huso*), the largest freshwater fish described to date. Due to the absence of nuclear markers, the evolutionary relationship among sturgeons has been studied for many years using only mitochondrial DNA (mtDNA), but recent investigations have developed phylogenetic studies based on nuclear protein markers derived from transcriptome data (Luo et al., 2019) (**Fig. 3**).

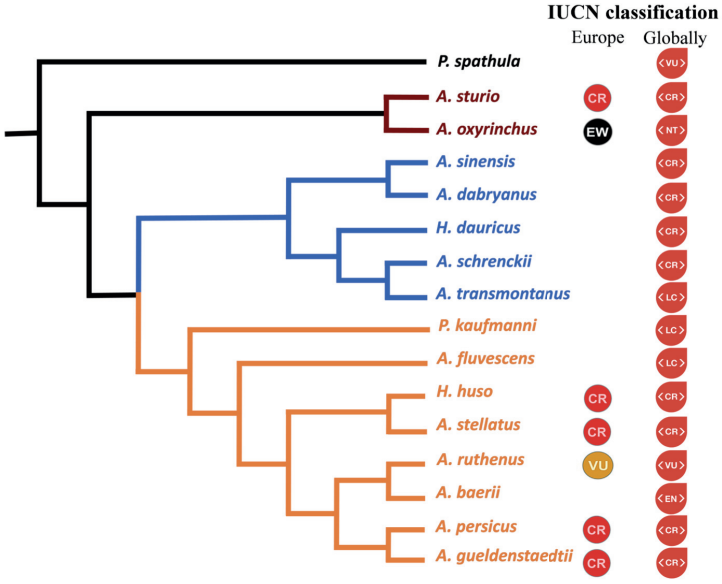


Figure 3. Phylogenetic classification of some sturgeon species. Brown lines indicate the Sea clade, blue indicate the Pacific clade, and orange indicate the Atlantic clade. Columns indicate the IUCN classification in Europe (left) and globally (right). LC: Least concern; NT: Near threatened; VU: Vulnerable; EN: Endangered; CR: Critically endangered; EW: Extinct in the wild. Figure adapted from Luo et al., 2019.

According to these new studies, sturgeons can be divided into the Sea clade, the Pacific clade and the Atlantic clade. The Sea clade diverged from the Atlantic-Pacific clade 144.9 Mya and contains only two species that diverged 68.3 Mya: *A. sturio* and *A. oxyrinchus*. The remaining sturgeon species belong to either the Pacific or the Atlantic clade, which diverged around 117.6 Mya (Luo et al., 2019).

As reported by the International Union for Conservation of Nature (IUCN) in its European Red List of Freshwater Fishes (Freyhof and Brooks, 2011), 37% of Europe's freshwater fish species are threatened, but the status of sturgeons is particularly worrying, with seven out of eight typically European species being critically endangered (*A. gueldenstaedtii, A. nacarii, A. nudiventris, A. stellatus, A. sturio, A. persicus* and *H. huso*) and one (*A. ruthenus*) classified as vulnerable. Natural occurrence of Atlantic sturgeon (*A. oxyrinchus,* Mitchill, 1815) in the Baltic Sea has been reported since the Middle Ages, but this species is now extinct in the wild in Europe (Ludwig et al., 2002, 2008; Elvira et al., 2015;). However *A. oxyrinchus* is the only sturgeon species that has a different classification on the European list than on the global list (www.iucnredlist.org), as wild populations are still present from the North of Quebec (Canada) to the Gulf of Mexico (United States), where the population tends to increase and is classified as near threatened (Pierre and Parauka, 2006).

Due to their long life span and late maturation, sturgeons are especially susceptible to anthropogenic factors like overfishing, poaching, pollution, river damming and climate change. Despite the fishing bans, recent studies on critically endangered sturgeon stocks in the Danube river basin, one of the last places in Europe where wild populations reproduce, reported illegal fishing activities that are threatening the future of these local populations (Smederevac-Lalic et al., 2012). In addition, the presence of pollutants like endocrine disrupting chemicals (EDCs) in the aquatic environment can interfere with the functioning of the fish endocrine system (Scholz and Klüver, 2009) and result in hermaphroditism and feminization (Wang et al., 2011). River dams and hydropower turbines can delay or even block migratory routes and affect the distribution of anadromous species, including sturgeons. Also, changes in water flow and river modernization have reduced the distribution of spawning grounds, limiting spawning opportunities and reducing the survival of the offspring. Fish, like all the living organisms, have an optimal temperature range in which they can live and reproduce. Therefore, increasing water temperatures caused by climate change have a modifying role in fish maturation and spawning, forcing them to migrate to cooler waters or, if they are unable to relocate, become less productive.

Sturgeon aquaculture

During the last 20 years the worldwide demand for caviar continued to grow, despite the decline of wild sturgeon populations. Sturgeons are currently protected and fishery is prohibited in most countries including Russia, one of the most important caviar producers worldwide (GAIN Report, 2014). Despite the taken measures, wild sturgeon populations continued to decline (Bronzi and Rosenthal, 2014; IUCN, 2018) and aquaculture has emerged as an alternative to supply the market, especially in the areas where sturgeons are or were naturally prevalent (Bronzi et al., 2011).

In 2017, up to 2,329 sturgeon farming plants were registered worldwide, most of them located in China (54%) and Russia (24%), and only 140 in Europe (Bronzi et al., 2019). The global sturgeon aquaculture biomass has increased more than four times in the last decade, peaking in 2015 with 129,608 tons (t), and decreasing ~ 20% in 2017 (**Fig. 4a).** With 79,638 t, the Chinese production represents approximately 78% of the global biomass production, followed by Russia (6,800 t), Armenia (6,000 t) and Iran (2,514 t) (FAO FishstatJ).

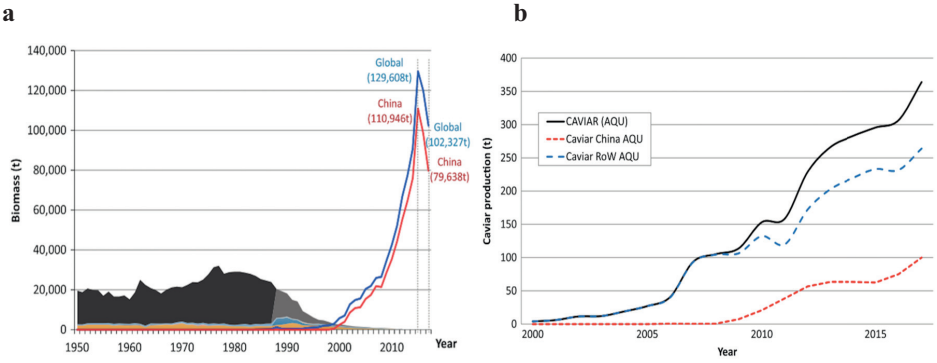


Figure 4. a: Global production of sturgeon in tons (t), between 1950 and 2017. Grey areas represent fisheries, while solid lines correspond to aquaculture, both globally (blue line) and in China (red line). **b:** Caviar production in tons (t) between 2000 and 2017. Solid black line represents the total amount produced in aquaculture (AQU), dashed red line represents only China and dashed blue line represents the rest of the world (RoW). Data and Illustration derived from Bronzi et al., 2019.

Sturgeon aquaculture pursues the production of two products: meat and caviar. The word "caviar" is restricted by the Food and Agriculture Organization (FAO) to eggs (roe) derived from sturgeons. While in most of the countries sturgeon meat is considered a sub-product of

the caviar industry, in countries such as China, Russia and Iran, sturgeon meat is an appreciated product. This explains why China is the leading sturgeon biomass producer worldwide while, when considering caviar only, its contribution to the market is much lower (approximately 27%) (**Fig. 4b**).

Caviar production from aquaculture has increased during the last 15 years, reaching 364 t in 2017. The time to reach gonad maturity varies among the different sturgeon species, and therefore has a direct effect on the length of the farming before harvesting and influences the choice of the species to farm.

In farms, rearing conditions are generally more favorable than natural conditions and puberty occurs considerably earlier. Siberian sturgeon (*A. baerii*) females reach sexual maturation after ~ 7 years in aquaculture conditions and therefore it is the most popular species for both caviar (31%) and meat (39.5%) production. Russian (*A. gueldenstaedtii*) (20.4%) and White sturgeons (*A. transmontanus*) (12.1%) are also popular species for caviar production, and Japanese sturgeon (*A. schrenckii*) (10.2%) for meat production. Hybrids (e.g. *A. schrenckii* x *Huso dauricus*, *A. baerii* x A. schrenckii) are also widely used in both industries, especially for caviar (35.6%). Overall, Atlantic sturgeon (*A. oxyrinchus*) represents only 0.1% of the caviar production worldwide.

Although caviar production has increased, it is still considered a luxury and pricey product and the market demand has not increased sufficiently suggesting that, if the trends remain the same, the demand will continue to be lower than the offer during the coming years (Bronzi et al., 2019; Sicuro, 2019).

Atlantic sturgeon restoration program

Aquaculture can be used as a tool to restore or enhance endangered species in the wild and compensate for losses in fish diversity. This practice is called stocking and, provided it is performed under well-designed guidelines, it can result in an effective increase in fish populations. This practice is very extended in salmonids (Finstad and Jonsson, 2001) and has also been used in tropical marine species (Ziemann, 2001); however, when it is used to create a self-sustaining fish stock in a location where it is extinct, it is called restoration. Atlantic salmon populations were also restored in several water bodies, like the Rhine, the Elbe and the Odra rivers, during the 1990s (Bartel, 1997).

As previously mentioned, the two sturgeon species living at either shore of the Atlantic Ocean, *A. sturio* in Europe and *A. oxyrinchus* in North America, diverged around 68.3 Mya (52.2-82.1 Mya, 95% CI) (Luo et al., 2019). However, archaeological evidence supports that *A. oxyrinchus* colonized the Baltic and replaced the native population (Ludwig et al., 2002, 2008) before getting extinct. *A. sturio* is currently restricted to a small population in the Gironde basin in France (Rochard et al., 1990) where is under a restoration program (Williot et al., 2005). On the other hand, *A. oxyrinchus*, currently restricted to North America, has been under restoration programs in several Baltic countries since the 1990s. Poland and Germany have been working together since 1997 to restore the Baltic population (Gessner, 2000). The reintroduction program focused initially on raising an *ex-situ* broodstock locally with which to produce restocking material that can be released in different regions of the rivers that flow into the Baltic Sea.

The introduction of nonnative species could result in increased stocking mortality, and even in the event of survival, the nonnative species could negatively influence the structure of native populations or other native species (Ludwig, 2006). Therefore, it seems logical that the same population that swam to the Baltic back in the Middle Ages is the one used to restore it, especially considering that Atlantic sturgeon in North America (also known as Gulf sturgeon) is not considered threatened. Since 2006, broodstock from the St John River in Canada is being transported to Europe in cooperation with two Canadian research facilities, while spawn and hatch is in parallel being collected and sent to Germany and Poland. All the imported material is genetically analyzed using microsatellite DNA markers to ensure genetic diversity, and a fragment of cytochrome b, a mitochondrial gene, is sequenced in order get species confirmation (Fopp-Bayat et al., 2015).

Successful spawning requires certain photoperiod and temperature conditions and the presence of spawning grounds, which should consist of shallow rock or gravel substrate to adhere the eggs, and an adequate water flow for egg oxygen supply. Therefore, rivers flowing to the Baltic and their tributaries have been searched for possible spawning grounds. The Odra River is 865 km long and hosts historical spawning ground for sturgeons. Although the construction of dams along the upper parts was one of the most important factors that led to extinction of the Baltic population, the lower sections remain quite conserved (Gessner and Bartel, 2000). Oxygen concentrations, presence of sediments and temperature fluctuations have also been assessed in order to select the best potential locations for the release of stocking material. It is important to remember that released

sturgeons will only go to their spawning ground to reproduce when they have matured (at least a decade later), therefore the conditions of their upriver migration and spawning grounds should be ensured at least until then, and this can only be achieved by the establishment of habitat preservation policies and international cooperation (Kirschbaum et al., 2015). As of 2018, more than 3 million individuals have been released in the Vistula and Odra Rivers (**Fig. 5**), among others (Gessner et al., 2019), although this has not yet resulted in any known reproduction. The age upon release is also an important factor to take into consideration: younger fish have higher post-release mortality, but they are also more likely to keep natal river imprinting and subsequent homing fidelity. Short-term (30 days) telemetry studies showed that 9-month old Atlantic sturgeon juveniles released into the Odra perform downriver migration (Fredrich et al., 2008), in contrast with observations of juvenile pallid sturgeon (*S. albus*) released in the Missouri River that showed upstream water dispersal (Eder et al., 2015). In both cases, the survival rate was close to 100%. On the other hand, very few studies focus on the long-term survival from release to adulthood.

Figure 5. Location of the Odra and Vistula Rivers in Germany and Poland, and estuary in the Baltic Sea.

Aquaculture-reared *A. oxyrinchus* juveniles (220 days) from wild parents were released into the Suwanee River and monitored during 19 years, showing more than twice as much (26.7%) mortality (or lack of homing behavior) compared with the wild cohort (11.2%), especially at the onset of marine life (Sulak et al., 2014). They also showed smaller size and lower weight, suggesting lower fitness. Other studies with shortnose sturgeon (*A. brevirostrum*) reported imprinting failure indicated by recaptures in other river systems and

a 0.3% survival after 10 years post-release (Quattro et al., 2002; Smith et al., 2002). In addition, millions of hatchery-reared fry (30-50 days) of Russian, Beluga and stellate sturgeons stocked annually since the 1950s in the Volga have reported survival to sexual maturation rates of 0.1-1% (Khodorevskaya et al., 1995; Levin, 1995).

Aquaculture methods are focused on producing fish fast and uniformly, rather than adapting the production to the requirements of releasing vigorous fish, leading to high post-release mortality or disrupted homing behavior. Sturgeon distribution and abundance in the wild is determined by several factors, such as temperature, photoperiod and salinity fluctuations along the migration route (Beitinger et al., 2000; Gradil et al., 2014), however fish reared in farms are raised in stimulus-deprived tanks, with constant abiotic factors such as temperature and photoperiod, and feeding on commercial pellets (Mohler et al., 2004). For all these reasons, it seems logical to conclude that innovative techniques should be designed and implemented in order to wild-condition the fish produced for stocking. A recent study described that Atlantic sturgeon juveniles raised in an artificial river section exposed to natural temperature and photoperiod showed increased brain plasticity markers compared with those raised in aquaculture conditions (Camara-Ruiz et al., 2018a). In addition, fish fed for two weeks with chironomids buried in sand were able to feed significantly faster than the untrained cohort upon exposure of both groups to the same feeding regime (Camara-Ruiz et al., 2018b). Other authors focused on studying the swimming performance and the response to swimming exercise in sturgeons (Du et al., 2014; Katopodis et al., 2019).

In this thesis, we have focused on temperature training. According to the intergovernmental panel on climate change (IPCC, 2013), the upper 75 m of the ocean warmed constantly from 1971 to 2010 at a rate of 0.11ºC per decade; however, rivers can be affected more strongly, e.g. the Loire River in France has seen an increase of 2ºC in the past two decades. Temperature constitutes an important factor for the distribution of anadromous species and an increase, especially in summer, might alter environmental conditions during migrations and reproduction and result in a shift of the geographic distribution of species and local extinctions (Fields et al., 1993; Hofmann and Todgham, 2010). To date, there is no evidence of decreased water temperatures negatively influencing sturgeon survival. Fish perceive seasonal changes through changes in temperature and photoperiod; therefore, thermal adaptation is crucial for survival. Considerable work has been conduced on upper temperature tolerances of sturgeon. Siberian and Green sturgeon larvae exposed to

temperatures up to 26ºC showed notochord and muscle alterations and increased heat shock protein (HSP) levels (Werner et al., 2007; Aidos et al., 2017), while 28ºC was shown to be lethal (Linares-Casenave et al., 2013) and temperatures above 24ºC reduced the growth rate and feeding efficiency and increased mortality in shovelnose sturgeon juveniles (Kappenman et al., 2009). With the exception of Ziegeweid et al. (2008), who showed an increase in the critical thermal maxima in shortnose sturgeon juveniles acclimated to high temperatures (Ziegeweid et al., 2008), very little research has been done on how temperature training affects the response to a subsequent heat stress in sturgeons.

Genetic resources for sturgeons

The genome contains all the genetic material of an organism, including protein-coding genes, non-protein-coding functional DNA (e.g. structural RNA genes, promoters, silencers), and DNA for which the function is still poorly understood (e.g. simple sequence repeats and transposable elements). In eukaryotes, the typical protein-coding gene structure contains coding regions alternated with non-coding regions, which are transcribed but not translated. After the spliceosome has spliced the introns, the mature messenger RNA (mRNA) is transported to the cytoplasm and translated into proteins in the ribosomes. Alternative splicing results in different mRNA isoforms and therefore several protein isoforms, which are often tissue-specific.

In eukaryotes, genomic DNA (gDNA) is organized in chromosomes in the cell nucleus. The number of sets of chromosomes is called the monoploid number or ploidy level (x), while the number of chromosomes in a gamete, which has undergone meiosis, is known as the haploid number (n). Therefore, in diploid organisms the monoploid (x) and haploid (n) numbers are interchangeable concepts. Chromosome numbers can vary from as little as a single chromosome in males of the species *Myrmecia pilulosa* (Crosland and Crozier, 1985) to up to 1,260 in some ferns of the genus *Ophioglossum* (Khandelwal, 1990). Among vertebrates, the cyprinid *Ptychobarbus dipogon* contains the highest number of chromosomes (~446) reported to date, while the second highest chromosome count (~437) was recently reported in a cultured heptaploid Siberian sturgeon (*Acipenser baerii)* (Havelka et al., 2016). Despite the above-mentioned exceptions, the haploid number of chromosomes in fish remains remarkably constant between 24 and 29, even among species with high DNA content, like the Atlantic salmon (Klinkhardt et al., 1995). Previous studies found that

Atlantic sturgeon's haploid chromosome number is 60 (2n=120), with a high presence of microchromosomes (Fontana et al., 2008; Ludwig et al., 2001).

The C-value is the amount of DNA contained in a germinal cell and it's measured in picograms (pg) or base pairs (bp) per nucleus (1pg= 978 Mbp). Genome size and C-value are interchangeable concepts in diploids, but in polyploids the C-value represents two or more genomes. Among vertebrates, genome sizes are highly variable from as little as 378.6 Mbp in Golden puffer (*Arothron meleagris*) (Ojima and Yamamoto, 1990) to up to 129.9 Gbp in Marbled lungfish (*Protopterus aethiopicus*) (Pedersen, 1971). The lack of correlation between the genome size and the organismal complexity remains still unexplained and is known as the C-value paradox. The American paddlefish (*Polyodon spathula*), sturgeon's closest relative, has a haploid chromosome number of 60 (like Atlantic sturgeon) and a genome size of 1.9 Gbp, based on flow cytometry (Gregory, 2019). Similarly, Atlantic sturgeon's genome size is estimated to be 2.14 Gbp for diploid individuals, based on Feulgen Image Analysis densitometry (FIA) in red blood cells (Gregory, 2019).

Evidence of two rounds (2R) of whole genome duplication (WGD) events affecting vertebrates 590 Mya (Meyer and Schartl, 1999) was in large part initially revealed by the discovery of duplicated Hox genes, whose structure and function are conserved across wide taxonomic distances (McClintock et al., 2001; Wagner et al., 2003; Soshnikova et al., 2013). Today it is known that teleosts have undergone an additional teleost-specific genome duplication event (TSGD) about 334 Mya, and salmonids and cyprinids have undergone a fourth round each. As non-teleost actinopterygians, Acipensiformes have not undergone the TSGD, but they have arisen by a polyploidization event from a diploid ancestor, and functional diploidy was reestablished before the radiation of the order (Krieger et al., 2008). Based on chromosome numbers and microsatellite markers, it is believed that sturgeons have undergone three subsequent linage-specific duplication events that lead to diploid, tetraploid and hexaploid species that are particularly prone to hybridize (Ludwig et al., 2001; Miloš et al., 2011; Havelka et al., 2013). Polyploidy is very common in plants but rare in fish, with the exception of chondrichthyans (Stingo and Rocco, 2001) and sturgeons (Blacklidge and Bidwell, 1993), in which polyploidization might still be ongoing as triploids can occasionally be found within diploid species, like the Atlantic sturgeon (Blacklidge and Bidwell, 1993). Previous studies considered species with 60 chromosomes (2n=120) as tetraploids, based on the high nuclear content and chromosome numbers, but

following studies agreed to classify them as diploids, based on microsatellite disomic allelic patterns (Havelka et al., 2013).

Overall, chromosome number, genome size and ploidy level are important features to consider before initiating a genome sequencing project. In addition, it is also important to know the percentage of repetitive content and heterozygosity. The bacterium *Haemophilus influenza* had the first published genome (Fleischmann et al., 1995), followed by the unicellular eukaryote *S. cerevisiae* (Goffeau et al., 1996), the multicellular eukaryote *C. elegans* (C elegans Sequencing Consortium, 1998), the model plant *A. thaliana* (Arabidopsis Genome Initiative, 2000) and the first fish, the Japanese pufferfish (Aparicio et al., 2002). All these genomes were quite small and it wasn't until 2004, after 15 years of work, that the human genome was published (Human Genome Sequencing Consortium, 2004) (haploid genome size 3.23 Gbp). Today more than 25,000 eukaryotic genome projects are deposited at the National Center for Biotechnology Information (NCBI), of which more than 11,000 correspond to plants, and almost 5,000 to animals, of which 318 are fish.

Table 1 shows a selection of fish genomes belonging to different fish lineages:

Species	C-value (Gbp)	Chromosome number (haploid)	Genome assembly (size; n. scaffolds; N50)	References
Elephant shark (*Callorhinchus milii*)	1.89	Unknown	0.93 Gbp; 21,204; 4.52 Mb	(Hardie and Hebert, 2004; Venkatesh et al., 2014)
Coelacanth (*Latimeria chalumnae*)	2.68	24	2.86 Gbp; 22,819; 0.92 Mb	(Amemiya et al., 2013)
Zebrafish (*Danio rerio*)	1.71	25	1.37 Gbp; 2,848; 7.37 Mb	(Ciudad et al., 2002; Howe et al., 2013)
Spotted gar (*Lepisosteus oculatus*)	1.40	29	0.945 Gbp; 2,106; 6.92 Mb	(Braasch et al., 2016; Ojima and Yamamoto, 1990)
Atlantic salmon (*Salmo salar*)	3.03	29	2.97 Gbp; 241,573; 1.36 Mb	(Lien et al., 2016; Vinogradov, 1998)
European eel (*Anguilla anguilla*)	1.19	19	**A**: 1 Gbp; 501,148; 59 kb **B**: 0.86 Gbp; 2,338; 1.2 Mbp	(**A:** Henkel et al., 2012; **B**: Jansen et al., 2017a)
American paddlefish (*Polyodon spathula*)	1.9	60	Not sequenced	
Sterlet (*Acipenser ruthenus*)	1.82	60	1.83 Gbp; 215,913; 191.06 kb	Gene Bank Accession GCA_004119895.1
Atlantic sturgeon (*Acipenser oxyrinchus*)	1.28	60	**This thesis**	(Hardie and Hebert, 2004; Yebra-Pimentel et al., 2019)

Table 1. List of selected fish genomes, indicating the species, genome size in Mega base pairs (Mbp), haploid number of chromosomes and genome assembly statistics.

chondrichthyans (elephant shark), sarcopterygians (coelacanth), teleosts (zebrafish, Atlantic salmon, European eel) and non teleost-actinopterygians: the holostean spotted gar, and the chondrosteans American paddlefish, sterlet and Atlantic sturgeon, the species under study in this thesis.

Sequencing technologies

DNA sequencing began with Sanger sequencing in 1977 (Sanger et al., 1977) and has rapidly advanced since then with the introduction of shotgun sequencing (Gut, 2013; Verma et al., 2017). High-throughput or next generation sequencing (NGS) is relatively new and rapidly evolving and can be grouped into three categories: 1) Sequencing by synthesis; 2) Sequencing by ligation and 3) Single molecule sequencing (third generation sequencing, TGS).

These three categories employ different gDNA preparation methods and nucleotides recognition systems, resulting in differences in throughput, read length and sequencing errors. Illumina sequencing (sequencing by synthesis) consists of shearing gDNA and ligating it to a short nucleotide stretch with known sequence, called adapter, resulting in modified gDNA fragments that are subsequently PCR amplified and poured into a flow cell. The flow cell surface is covered with reverse complementary copies of the adapters that bind covalently to the poured fragments. The fragments are locally amplified several times forming bridges (bridge PCR), resulting in the formation of so-called clusters that are read by the system using a replication machinery that includes modified nucleotides. Each of the four nucleotides is attached to a fluorescent group and the 3' OH group is modified in order to temporarily interrupt the replication after the incorporation, making possible to control and monitor the replication process (**Fig. 6a**).

On the other hand, Oxford nanopore technologies (ONT) provide platforms for TGS based on nanopore sensing (Division et al., 1996; Meller et al., 2001; Venkatesan and Bashir, 2011). The flow cell consists of a polymeric membrane with an array of wells with nanopores embedded, and an application specific integrated circuit (ASIC) underneath, which contains the signal amplifiers. In this case, gDNA does not need to be sheared or amplified, but adapters are still needed. Tethering oligonucleotides are used to keep the DNA molecules on the membrane and increase the concentration near the pores. During the run, a voltage is applied across the membrane, generating a current through the nanopores. A motor protein unwinds the double stranded DNA (dsDNA) allowing it to translocate

through the pores at a controlled speed (450 bases per second, bps). This translocation results in a disruption of the current through the pore, generating what are called events that, using base-calling software, are translated into nucleotides with a probability value or a quality score attached (**Fig. 6b**).

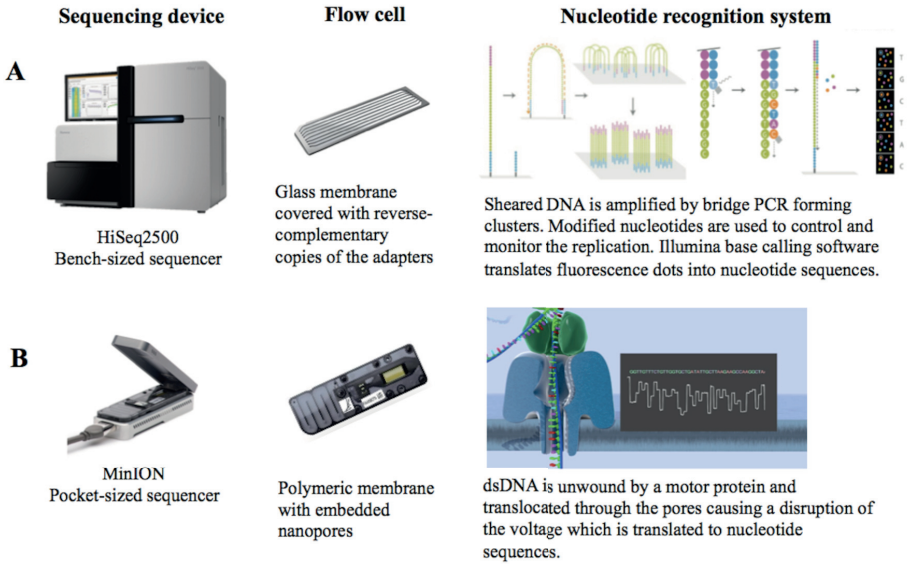


Figure 6. Comparison between Illumina (**A**) and nanopore (**B**) sequencing platforms in terms of sequencing devices, flow cells and nucleotide recognition systems.

Nanopore sequencing was introduced in the market in 2014 (Mikheyev and Tin, 2014) and currently offers five sequencing devices: the pocket-sized SmidgION, Flongle and MinION, and the bench devices GridION and PromethION. The procedure to prepare the gDNA to be sequenced is the same for all the devices, the only difference being the amount of pores in the flow cells and, consequently, the throughput. Both MinION and GridION devices run with flow cells containing 512 sensors connected to 2,048 pores, but the GridION offers the possibility to run up to 5 flow cells simultaneously. The PromethION flow cells contain 3,000 sensors connected to 12,000 pores, and depending on the system they can run up to 24 or 48 flow cells at a time.

In order to reconstruct the original genome or transcriptome, sequencing reads need to be assembled. The quality of the assembly depends on the quality and length of the sequencing reads and the sequencing coverage. The quality (Q) of the sequencing reads is measured

using Phred scores (Ewing et al., 1998), which is the negative ratio of the error probability ($P= 10^{-Q/10}$). In this way, a Q of 10, 20, 30 or 40 means that the error rate for that base call is 10, 1, 0.1% and 0.01% respectively. When sequencing by synthesis several thousands of molecules at the same time, at the end of the fragments some molecules are one or more nucleotides ahead and some lack behind, producing signal noise and resulting in a lower Q of the called bases. This feature limits the Illumina sequencing read length to up to 300 bp, nevertheless, ensuring a Q≥30 (Nakamura et al., 2011). In nanopore sequencing read length is not limited by the system, however, the average Q is ~10 (Goodwin et al., 2015) and needs to be addressed by post-sequencing bioinformatics. In addition, nanopore sequencing reads are also characterized by the under-representation of homopolymers in the raw reads (Lu et al., 2016). The sequencing coverage or sequencing depth refers to the total amount of sequencing data relative to the genome size. When the reads are of a fixed length (e.g. Illumina) the coverage can be calculated by multiplying the number of reads (N) by their length (L) and dividing it by the genome size (G).

Highly repetitive regions of the genome represent a challenge for genome assemblers and can only be resolved if the reads are long enough to cover the repetitive elements (Treangen and Salzberg, 2012). Therefore, long reads make the assembly computationally easier, however high quality reads result in a more accurate assembly. Illumina reads solely are poorly suited for *de novo* assembly of complex genomes due to their short length and in general platforms generating long reads (e.g. ONT) are better for building a contiguous assembly, besides the high error rate.

Genome assembly contiguity is usually evaluated by the number of contigs (contiguous sequences assembled from reads) or scaffolds (succession of contigs separated by gaps) and the median length of the assembly (N50), which is the minimum contig or scaffold size in the assembly after which half the genome is covered (Hirakawa et al., 2019). However, these metrics don't consider misassembly that can result from high heterozygosity and ploidy. Genome completeness can be assessed based on the presence of gene content using a selection of near-universal single-copy orthologues (Waterhouse et al., 2018). Finally, based on the quality standards established by the Human Genome Organization, the error rate of the final assembly should be ≤0.01%. Once the genome assembly has been established to a satisfactory level, gene finding and annotation can be performed.

Knowledge gap

Restocking protocols need to be developed in order to improve post-release survival of fish used for restoration. Recent studies focused on swimming and foraging training; however, due to the global increase of water temperatures and its impact on sturgeon reproduction and distribution, we have focused on temperature training.

Many studies have assessed the temperature upper limits in sturgeons, but none of them focused specifically on Atlantic sturgeon. Very few studies have focused on how temperature training and adaptation affect the response to a subsequent heat shock event. Instead, most studies focused on survival, growth rate and phenotypic malformations. Moreover, due to the lack of genomic resources, gene expression studies only focused on a limited number of genes.

Good genomic and transcriptomic resources have a strong impact on research approaches in molecular biology, ecology, physiology and evolution, and are essential for improving artificial reproduction, broodstock management and restocking strategies. Additionally, it provides a reference for RNAseq-mediated transcriptome mapping for the analysis of maturation, immune response, stress tolerance and local adaptations.

At the beginning of this thesis, no transcriptomic data was available for Atlantic sturgeon and genomic data was absent for any of the sturgeon species. During the course of this thesis, transcriptomic data for Atlantic sturgeon were deposited by others in NCBI, including only embryonic stages. In addition, the genome sequence of sterlet (*A. ruthenus*) has recently become available at NCBI; however, this assembly was done using short reads only and is still fragmented (contig and scaffold N50 18.88 and 191.06 Kb, respectively), and additionally sterlet and Atlantic sturgeon diverged ~145 Mya, which highlights the need to create a genome reference specifically for Atlantic sturgeon.

In order to fill the knowledge gap in sturgeon research, we have developed a reference genome and transcriptome for Atlantic sturgeon and used them to find all the heat shock proteins (HSP) genes and measure their response to heat-shock in an *in vitro* culture sturgeon cell line. Additionally, we have analyzed the transcriptomic response to heat shock in temperature-trained fish compared to naïve fish in order to determine the effect of the training.

Aims of the study

The general aim of this thesis was to generate genomic and transcriptomic resources for sturgeons that can be used to develop novel techniques and applications in sturgeon aquaculture and restoration programs. The following sub-aims were identified:

1. Generate complete transcriptomic resources for Atlantic sturgeon originating from different organs (**Paper I**).
2. Develop a set of gene markers that can be used for improving restocking programs using *in vitro* and *in vivo* models (**Papers I and II**).
3. Generate genomic resources for Atlantic sturgeon and determine the evolutionary relationships between Atlantic sturgeon and other vertebrates (**Paper III**).

Methodological considerations

Experimental animals

For studies described in **papers I and III**, a juvenile Atlantic sturgeon female was sampled at the Fischzucht Rhönforelle GmbH in Gersfeld (Germany) on July 3rd, 2015. The specimen was initially anesthetized with clove oil, followed by anesthetic overdose and decapitation. In order to generate transcriptomic (**Paper I**) and genomic resources (**Paper III**) for Atlantic sturgeons, 21 samples were stored (**Fig. 7**) in RNAlater and subsequently transferred to liquid nitrogen (after removing RNAlater excess). Blood was taken with a heparin-flushed needle, transferred to a heparin-flushed cryovial tube and stored in liquid nitrogen.

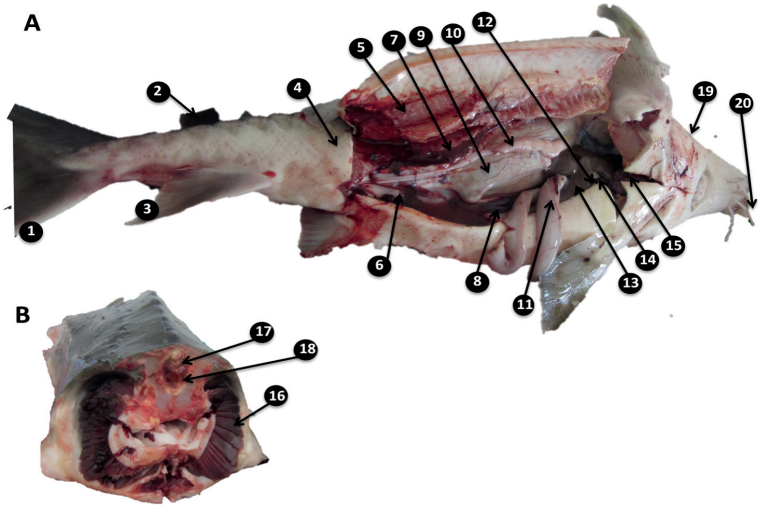


Figure 7. List of tissue samples taken by dissection of juvenile *Acipenser oxyrinchus*. **A:** Longitudinal view. **B:** Transversal view of the skull. **1:** Caudal fin. **2:** Dorsal fin. **3:** Pelvic fin. **4:** Skin. **5:** Muscle. **6:** Intestine (posterior). **7:** Kidney. **8:** Spleen. **9:** Swimbladder. **10:** Gonad. **11:** Intestine (anterior). **12:** Stomach. **13:** Liver. **14:** Gallbladder. **15:** Heart. **16:** Gill. **17:** Brain (dorsal). **18:** Brain (ventral). **19:** Eye. **20:** Barble.

For the experiment described in **paper II,** 24 Atlantic sturgeon juveniles were sampled at the Leibniz Institute of Freshwater ecology. Three-month-old juveniles were distributed to

two raceway units at a natural photoperiod and acclimated for 7 days. Two experimental groups were established: during the 30 days prior to the heat shock experiment, the control group (C; n=12) was constantly kept at 20ºC, while the trained group (T; n=12) was subjected to 7 rounds of a temperature challenge effectuated by an increase of the water temperature from 20ºC to 24ºC over 12 hours and a subsequent decline to 20ºC over the following 12 hours. The seven rounds of temperature challenge were randomly distributed over the 30-day period to avoid habituation. The heat shock challenge comprised the simultaneous exposure of both groups to the temperature challenge. At the beginning of the temperature increase (t=0h), at the peak of the heat treatment (t=12h), at the return to basal temperature (t=24h) and after 24 hours of recovery (t=48h), three animals per group (n=3) were euthanized and their livers were sampled in RNAlater for subsequent transcriptome studies.

Cell cultures

In **paper I**, a commercially available Atlantic sturgeon embryo-derived cell line (Grunow et al., 2011) was used in an *in vitro* heat shock experiment. Frozen cells from passage 22 (P22) were sent by the Fraunhofer Research Institution of Marine Biotechnology and, upon arrival, seeded in a 25 cm^2 cell culture flask (Corning Life Sciences, Tewksbury, MA, USA) at 25°C (the optimal growing temperature) in Leibovitz-15 medium supplemented with 15% FCS (fetal calf serum), 100U/mL penicillin and 0.1mg/mL streptomycin (**Fig. 8**).

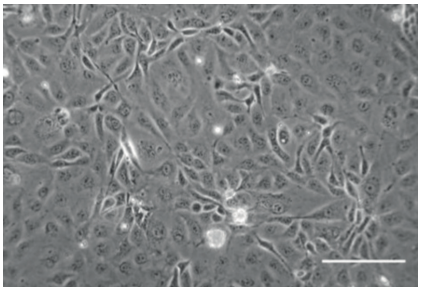


Figure 8. AOXlar7y cell line (**A**cipenser **ox**yrinchus **lar**vae nº**7**, tr**y**psin-digestion), at passage P9. Scale bar: 100µm.

Cells were propagated up to P29 and exposed to a 1-hour heat shock at either 30ºC or 35ºC. After the heat shock all the flasks were transferred back to 25°C to recover. After 4, 8 and

24 hours of recovery, cells were lysed in QIAzol lysis reagent (Qiagen GmbH, Hilden, Germany) and stored at -80°C until further processing.

RNA sequencing

Gene expression studies identify genes and pathways that play important roles in physiological or pathological processes, and can be performed either at the transcript level (e.g. Northern blot, microarrays, qPCR, RNAseq) or at the protein level (e.g. Western blot, mass spectrometry).

In this thesis we have measured gene expression at the transcriptome level using RNA sequencing (RNAseq), which relies on the reverse transcription of mRNA to complementary DNA (cDNA) followed by adapter ligation, PCR amplification and sequencing of the fragments. Highly expressed transcripts will be overrepresented in the sample and in the sequencing reads, therefore the number of reads mapping to each transcript correlates with the expression level (Mortazavi et al., 2008). Depending on the aim and budget, one or both ends of the fragments can be sequenced, resulting in either single reads (SR) or paired-end reads (PE). In addition, the length of the reads can also vary: longer reads from both ends (e.g. PE150) are better if there is no reference, as the reads themselves can be assembled to build a reference, against which the reads can be aligned again. If there is already a good reference available, shorter reads (e.g. SR50) are enough to quantify gene expression. Quantitative PCR (qPCR) is frequently used to study the expression of a limited number of genes, while microarrays can be used if the aim is to study thousands of genes at a time. However, these techniques rely either on primer amplification of the transcripts or on the hybridization of the mRNA with a DNA template immobilized on an array. In both cases, the transcript sequences need to be known, while RNAseq provides the possibility to discover new transcripts, detect splicing variants and, as mentioned before, allows for *de novo* assembly, which will ideally contain the full-length transcripts of the expressed genes (Grabherr et al., 2011). RNAseq has been widely used to study expression since its inception more than a decade ago, and several studies have found good correlation between RNAseq and qPCR, provided they are properly performed (Asmann et al., 2009; Griffith et al., 2010; Shi and He, 2014; Wu et al., 2014).

The sequencing reads need to be aligned to the reference using one of the many RNAseq alignment tools (Langmead et al., 2009; Li and Durbin, 2010; Langmead and Salzberg, 2012; Li and Durbin, 2010; Bray et al., 2016), resulting in a file in Sequence Alignment Map (SAM) format describing where (coordinates) and how (alignment quality) the reads align to the reference. Afterwards, the reads need to be counted in order to get a measure of the gene expression. Once a list of transcripts with the raw number of reads mapping to each of them is generated, these counts need to be normalized in order to perform any expression comparison. This is due to the fact that some samples might have been sequenced more deeply and therefore a higher number of reads were aligned to the reference in the first place. In the same manner and given the same level of expression, longer transcripts will have a higher number of reads aligning to them than those that are short. Therefore, the raw counts need to be divided by the number of sequencing reads (in millions) and by the transcript length (in Kb), resulting in normalized counts expressed in reads per kilobase and million (RPKM). Thereafter it is possible to perform differential expression analysis using one of the several statistical software packages designed for analyzing RNAseq data, which follows a negative binomial distribution (Robinson et al., 2009; Anders and Huber, 2010). In addition, it is possible to group the differentially expressed genes under a set of biological terms, called gene ontology (GO) terms (Carbon et al., 2017) and link them to pathways (Huang et al., 2009; Fabregat et al., 2016) (**Fig. 9**).

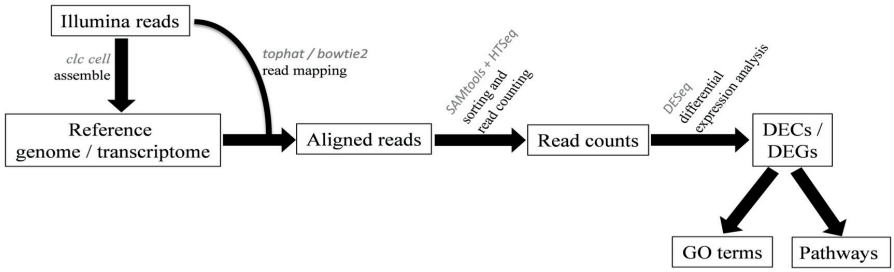


Figure 9. RNAseq analysis workflow. Sequencing reads are either assembled to create a reference transcriptome or are mapped to the reference genome or transcriptome. The alignments are counted and a differential expression analysis is performed in order to identify differentially expressed contigs (DECs) or genes (DEGs) that can be grouped in gene ontology (GO) terms or pathways.

In this thesis, the *de novo* assembler implemented in the CLC bio Genomics Workbench version 3.6.5 (CLC bio, Aarhus, Denmark) was used for the assembly of Illumina reads (**Papers I and III**). Bowtie2 (version 2.2.5), SAMtools (version 1.2) and the HTSeq python package (version 0.5.3p9) were used to align the reads to the reference transcriptome, exclude secondary aligned reads and count the reads, respectively (**Papers I, II and III**). DESeq package, running in R (version 3.5.2.), was used to perform differential expression analysis (**Papers I and II**).

DNA sequencing

For Illumina sequencing, DNA was isolated from blood using the DNeasy Blood and Tissue kit (Qiagen), which allows rapid purification of DNA. Commercial kits are generally designed for isolating DNA from mammals and therefore, in the case of blood, it is assumed that only the white blood cells contain DNA. Contrary to mammals, fish erythrocytes are nucleated; therefore very little starting material is needed (10µL) as overloading the column will decrease the yield of the recovered DNA. Blood cells were lysed, proteins were digested with proteinase K and the lysate was then applied to the column and spun down. During this process DNA adsorbed to the silico-based membrane of the column due to the presence of high concentration of chaotropic salt in the buffer. Two washing steps were performed to ensure that contaminants, divalent cations and enzyme inhibitors were washed away and afterwards the DNA was eluted, producing DNA of ~30 kb in length, assuming proper sampling and storage conditions of the starting material. For the Illumina sequencing library preparation (**Paper III**), DNA needs to be sheared mechanically (e.g. nebulizer) or chemically (e.g. transposase) and size selected. In this thesis, DNA was run on 1% agarose gel and the 400nt band was cut from the gel and the DNA fragments were purified from the gel. Thereafter, Illumina adapters were attached to each end of the fragments, amplified by 10 PCR cycles with random primers, poured into the Illumina flow cell and run at paired end 150 (PE150) on a HiSeq2500 Illumina sequencing device. Sequencing reads were provided in FASTQ format, which contains the nucleotide sequences and the quality scores associated with each base. The quality of the raw reads can be inspected using a variety of software packages like FASTQC (Andrews, 2010). For the *de novo* assembly, it is important to remove adapter sequences that might lead to misassemblies.

In nanopore sequencing, the length of the sequencing reads depends on the length of the DNA used to prepare the library, therefore the DNA isolation method plays an essential

role. Many isolation protocols have been developed to extract high molecular weight (HMW) DNA from several types of tissues or cells (Gong et al., 2019); however, we have used the commercial kit Genomic-Tip 100/G (Qiagen) which yields HMW DNA (≥ 60 kb) with high reproducibility (**Paper III**). For this protocol, a proteinase K step is also used; however, the column is drained by gravity. The length of the DNA can be estimated accurately up to 60 kb using regular gel electrophoresis (e.g. 4200 TapeStation System) and lambda phage DNA (48 kb) as marker. For DNA longer than 60 kb pulsed-field gel electrophoresis (PFGE) is needed to estimate the size accurately, using lambda phage DNA concatemers as marker. The purity of the DNA is also crucial in nanopore sequencing, as impurities will block the pores and inhibit the sequencing. The absorbance at 230, 260 and 280 nm can be measured by spectrophotometry (e.g. NanoDrop) and the optimal A_{260}/A_{280} and A_{260}/A_{230} ratios for pure DNA are ~1.8 and ~2 respectively. Deviations from these ratios are indicating the presence of contaminants such as ethylenediaminetetraacetic acid (EDTA), phenol or chaotropic salts like guanidine hydrochloride. Several library preparation protocols are available for ONT platforms, but for whole genome sequencing without PCR amplification, there are mainly two: a faster "rapid" method (RAD004) and the "1D ligation" method (LSK109). The first is specifically used to obtain ultra long sequencing reads (>100 Kb), while the latter produces comparatively shorter reads but at a higher yield. Like other ONT users, we have experienced an inverse correlation between the length of the DNA fragments in the library and the yield of the run. Therefore, it is very challenging to obtain high coverage in ultra long reads within a single sequencing run. The majority of the Atlantic sturgeon nanopore reads were obtained using the LSK109 library preparation protocol. In brief, this involves an initial DNA repair step to repair nicks in the DNA that would interrupt the sequencing, and a step to prepare the ends for the adapter ligation. Thereafter, sequencing adapters, which serve as binding points for the motor protein, are ligated in order to produce the final library. After every enzymatic step there is a purification step with magnetic AMPure beads (Beckman Coulter) in order to remove impurities and remaining enzymes that would negatively influence the sequencing. A small proportion of the Atlantic sturgeon nanopore reads was obtained using the RAD004 library protocol, which involves enzymatic shearing using a transposase, followed by adapter ligation and loading into the flow cell without any purification step. The transposase concentration and incubation time are correlated with the length of the DNA fragments in the library: the higher the transposase and the longer the incubation, the shorter the DNA fragments in the library. Once the library is ready, it can be loaded onto the flow cell. Then,

the sequencing run can be started, the duration of which can be set by the user but is by default 48 hours for the MinION and GridION systems and 60 hours for the PromethION. The amount of loaded library (expressed in femtomoles, fmol) is also highly correlated with the yield and ONT recommends between 5 and 50 fmol. When less library is loaded this will result in low pore occupancy (number of pores sequencing divided by the number of pores available for sequencing), while overloading the flow cell will initially result in a high pore occupancy, but a fast decrease of pores available for sequencing (usually after 10-15 hours). As for Illumina, a series of resources is freely available to analyze the quality and length of the nanopore reads after the run (De Coster et al., 2018; Leger and Leonardi, 2019).

Genome size and ploidy level estimation

Ploidy level can be estimated through a variety of methods, such as direct visualization of chromosome numbers (karyotyping), detection of nuclear organizer regions (NORs), and *in vitro* amplification of polymorphic markers such as microsatellite loci (Ludwig et al., 2001). Flow cytometry (Allen and Stanley, 1978) can also be used to determine both ploidy level and genome size. In addition, *k*-mer analysis estimates using bioinformatics can be used, provided the availability of highly accurate sequencing reads. Although this approach has been used in other species (Guo et al., 2015; Bian et al., 2016; Austin et al., 2017), *k*-mer analysis has not been performed in sturgeons due to the absence of genomic Illumina reads. In this thesis, we have used three different approaches to estimate the ploidy level and genome size of the Atlantic sturgeon specimen used for building the reference genome (**Paper III**).

K-mer counting

Genome size can be estimated from sequencing reads by counting the occurrence of all *k*-mers of size *k* in a given sequence of length *L*. The total number of possible *k*-mers will be given by [(*L*-*k*) +1], but not all *k*-mers will have the same frequency: a large fraction will occur *c* times, which corresponds to the sequencing coverage. Therefore, knowing the sequencing coverage and the total amount of sequencing reads, it is possible to get a good approximation of the actual genome size. In this thesis we have used Jellyfish (version 2.2.6.) (Marçais and Kingsford, 2011) in order to count *k*-mers and examine their profiles

using the Illumina reads (filtered for Q≥30 and after adapter removal) and subsequently have used GenomeScope (Vurture et al., 2017) to produce a *k*-mer frequency graph that provides an estimation of the haploid genome size, heterozygosity and percentage of repetitive fraction. The genome size is estimated by normalizing the *k*-mer frequency to the average coverage for homozygous sequences. Homozygous genomes will show a typical Poisson profile with one peak while highly heterozygous genomes will show a bimodal profile with two peaks (Kajitani et al., 2014).

Flow cytometry

Flow cytometry is currently the most effective, rapid and accurate method to identify the ploidy status and determine the genome size in different fish species (Allen, 1983; Zhou et al., 2008) and has previously been used in sturgeons (Lebeda et al., 2018). In this thesis, sturgeon erythrocytes were used, and European eel erythrocytes were prepared in parallel and used as a standard (genome size 1.1 Gbp) (Henkel et al., 2012). Blood was washed with phosphate-buffered saline (PBS) and centrifuged in order to pellet the erythrocytes, which were subsequently fixed in 70% cold ethanol. After RNase A treatment to digest the RNA, propidium iodide (PI), an intercalating agent that binds to both DNA and RNA, was used to stain the erythrocyte nucleus (Hare and Johnston, 2011). Thereafter, the fluorescence of the cells was measured by FACS in up to 10,000 events per sample and the Atlantic sturgeon genome size was estimated based on the relative fluorescence of the standard (European eel).

Microsatellites markers

Microsatellites loci are the most commonly used molecular markers in population genetics studies and have been used to improve management of endangered species. A large set of microsatellites has been identified in several sturgeon species (King et al., 2001; Boscari et al., 2015; Li et al., 2015; Que et al., 2015; Shin et al., 2019), which have been used to infer ploidy level, as the maximum number of alleles per individual at each locus should reflect the minimal ploidy level. As mentioned in the introduction section, previous studies have observed 3 ploidy levels among sturgeons: species with 2n=120 chromosomes (like Atlantic sturgeon) typically show disomic band patterns, while species with 2n=250 or 2n=500 show tetrasomic and octosomic bands respectively (Havelka et al., 2017, 2013). In order to

confirm the ploidy level, we have tested 13 microsatellite loci in our sequenced Atlantic sturgeon, 11 of which were successfully amplified and 8 were chosen for subsequent analysis, based on their level of polymorphism.

Assembly methods

Once sequencing is finished to the desired coverage, sequencing reads are stacked together using assembly software in order to reconstruct the original genome in the form of contigs and scaffolds. To this end, several assembly algorithms have been developed but, irrespective of the chosen software, if the assembler finds a repeat that cannot be resolved with the provided sequencing reads, the contigs can't be extended. Generally, assemblers are based on several different types of algorithms, such as overlap-layout-consensus (OLC), de Bruijn graphs (DBG) or string graphs (Henson et al., 2012). The DBG approach is generally used to assemble short reads, but OLC-based algorithms perform better for longer reads with higher error rates. If NGS reads are available, one of the several approaches to hybrid assembly may be chosen: short accurate reads can be used to correct long error prone reads either before or after the assembly, which is referred to as "polishing".

In order to build up the transcriptome assembly (**Paper I**) and the Illumina-only genome assembly for Atlantic sturgeon (**Paper III**), the DBG-based *de novo* assembler implemented in the CLC Genomics Workbench version 4.4.1 (CLC bio, Aarhus, Denmark) was used. This approach divides the sequencing reads in smaller fragments of length *k*, called *k*-mers, and uses them to make an assembly graph, reducing significantly the search time for the optimal path (Pevzner et al., 2001). When sequencing non-haploid genomes, graph bubbles will form around heterozygous variants resulting in two possible paths that merge back together a short distance later. The optimal *k*-mer and bubble size to perform the assembly varies from genome to genome and can be set by the user, allowing testing of several combinations until the best balance between genome completeness and contiguity is achieved. Long *k*-mers are more unique in the genome and should result in longer contigs (contiguity), while small *k*-mers provide more connections in low coverage regions (completeness). The bubble size, on the other hand, corresponds to the length of the graph bubble that the assembler will try to resolve based on read coverage: typically, bubbles resulting from heterozygosity will have similar coverage in both paths, while in bubbles generated by sequencing errors the erroneous path will have lower coverage.

Due to the length of the reads used, the resulting assembly was very fragmented and further sequencing using ONT was performed in order to improve the contiguity of the genome. In addition to the many applications that nanopore sequencing has (Beato et al., 2018; Gallagher et al., 2018), it has also been used in order to improve the assembly contiguity of previously published genomes (Jain, 2018; Michael et al., 2018; Elbers et al., 2019; Shin et al., 2019) and to assemble genomes *de novo* (Austin et al., 2017; Jansen et al., 2017a; Jansen et al., 2017b).

As mentioned before, OLC-based algorithms are traditionally used to find overlaps among reads to create a read layout that is subsequently corrected based on read consensus. This is the strategy employed by SMARTdenovo (Schmidt et al., 2017) and Canu (Koren et al., 2017) assemblers, the latter being an extension of the Celera assembler used to assemble the first human draft genome. This approach implies aligning all the sequencing reads against each other and for Canu, also an initial correction of the long reads, resulting in high quality assemblies. However, the computational demand to complete this task is positively correlated with the genome size in a quadratic manner: doubling the genome size and therefore the coverage will quadruplicate the amount of read alignments and therefore the computing hours, making this approach suitable for small genomes like bacteria, but not for bigger genomes: Jain et al (2018) reported that Canu took 62,000 CPU hours to assemble the human genome (3.1 Gb) with 30x coverage, which would mean that the Atlantic sturgeon genome (1.25 Gb) would take ~10,000 CPU hours with the same coverage (Jain et al., 2018).

Therefore in this thesis we have used a heuristic approach to the OLC-based algorithms based on constructing an assembly graph by aligning the sequencing reads to a set of self-made short and unique sequences of the genome, called seeds, using the assembly software Tulipa-Julia. The first version of this software, called TULIP (The Uncorrected Long reads Integration Process), was successfully used to improve the European eel genome sequence (genome assembly statistics included in **Table 1** of the Introduction section) (Jansen et al., 2017b). The assembly approach implies building an assembly graph leading to an uncorrected assembly that needs to be corrected based on a nanopore consensus and polished with Illumina reads (**Fig. 10**). The assembly graph is built based on alignments of the long sequencing reads (links) to the selected seeds (nodes), therefore reads that don't align to seeds will not be included in the assembly, while seeds spanning repeats will have too many reads aligning to them and produce tangles in the graph. Therefore, the seed

selection step is crucial for the production of a complete and accurate reference genome. The seeds must span the entirety of the genome and be evenly distributed throughout it. The origin of the seeds can be either Illumina reads or contigs (if available) or nanopore reads. The seed amount, length and average quality can be set by the user: optimal results were obtained with seeds of 1-1.5 kb in length, an average quality of 9 to 10 Phred, and between 0.5-0.6 folds of the genome to assemble. Alternatively, Illumina contigs can also be split into fragments of equal sizes (1-1.5 kb) and used as seeds. Once we have a final seed set, the long nanopore read are aligned to them using Minimap2 (Li, 2018) producing an alignment file in SAM format. Based on this alignment, Tulia-julia builds an assembly graph resulting in contigs. Finally, the contigs need to be corrected, which is the most CPU-intensive stage of the assembly process, however quite standardized. Two to four iterations of nanopore consensus correction are needed, which can be done using Racon software (Vaser et al., 2017), and thereafter the Racon-corrected contigs can be polished with Illumina reads using one of the many software packages available for this purpose (e.g. Pilon, ntEdit) (Walker et al., 2014; Warren et al., 2019).

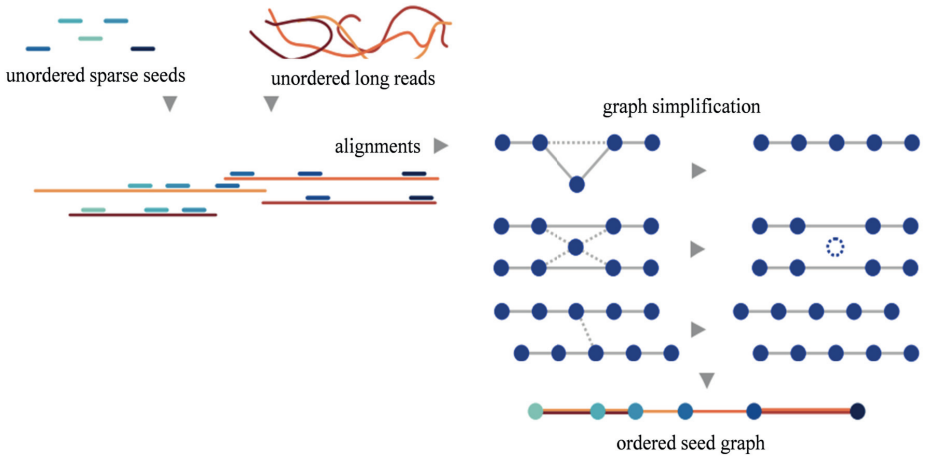


Figure 10. Overview of the genome assembly steps. Seeds are initially selected from nanopore or illumina reads or Illumina contigs, and nanopore reads are aligned to them. Tulipa-julia uses the alignments to produce a seed graph which is subsequently untangled and linearized producing uncorrected scaffolds that are corrected based on nanopore consensus and Illumina reads with correction software.

Evolutionary studies

Comparative genomics is used to compare genomic features of different organisms (de Crécy-Lagard and Hanson, 2013) and maximizes the added value of genome sequencing projects. Exploring the relationship between the Atlantic sturgeon genome and other vertebrate genomes in terms of extent of synteny conservation and protein sequence similarities (phylogeny) could give insights into the evolution of vertebrate genomes.

In this thesis (**Paper III**) we have performed synteny and phylogenetic studies of the Hox clusters using Atlantic sturgeon and other representative species in vertebrates.

For the synteny studies, we identified the Hox clusters and neighboring genes in the Atlantic sturgeon genome and the synteny was compared with the genome of key species, including: two sarcopterygians (coelacanth and Human), two chondrichthyans (Elephant shark and Hornshark), two teleosteans (Zebrafish and European eel) and three non-teleost actinopterygians (Spotted gar, Sterlet and Atlantic sturgeon).

The same species were used for the construction of phylogenetic trees, including also the American paddlefish, for which a reference genome is lacking, but HoxA and HoxD sequences are available. Alignments were performed using CLUSTAL, included in SeaView, followed by manual curation. Phylogenic analysis of the alignments was achieved by using a maximum likelihood method in CIPRESS (www.phylo.org) using RAxML black-box with 1,000 bootstrap replicates and JTT substitution matrix.

Results: Summary of papers

Paper I: Deep transcriptome analysis of the heat shock response in an Atlantic sturgeon (Acipenser oxyrinchus) cell line

Yebra-Pimentel, E.S.; Gebert, M; Jansen,H.J.; Jong-Raadsen, S.A.; Dirks, R.P.H.

Fish and Shellfish Immunology 88 (2019) 508-517.

https://doi.org/10.1016/j.fsi.2019.03.014

Despite efforts to restore Atlantic sturgeon in Europe, aquaculture techniques result in domesticated animals with high post-release mortality due to, among other reasons, their low tolerance to increasing water temperature. Marker genes to monitor heat stress are needed in order to identify heat-resistant fish. Therefore, an Atlantic sturgeon cell line growing at an optimal temperature of 25ºC was exposed to different heat shock protocols (30ºC and 35ºC) and differences in gene expression were investigated. In total 3,020 contigs (~1.5%) were differentially expressed. As the core of the upregulated contigs corresponded to heat shock proteins (HSP), the heat shock factor (HSF) and the HSP gene families were annotated in Atlantic sturgeon and mapped via Illumina RNA sequencing to identify heat-inducible family members. Up to 6 *hsf* and 76 *hsp* genes were identified in the Atlantic sturgeon transcriptome resources, 16 of which were significantly responsive to the applied heat shock. The *hspa1* (*hsp70*) gene, previously used as a heat-shock marker, was only significantly upregulated at the highest heat shock (35º), while a set of five genes (*hspc1, hsph3a*, *hspb1b*, *hspb11a* and *hspb11b)* was upregulated at all conditions. Although the *hspc1 (hsp90a)* gene was also previously used as heat-shock marker in sturgeons, *hspb11a* is the most heat-inducible gene, with up to 3,296-fold higher expression in the treated cells, constituting the best potential marker for in vivo trials.

Paper II: Temperature training improves transcriptional homeostasis after heat shock in juvenile Atlantic sturgeon (*Acipenser oxyrinchus*).

Yebra-Pimentel, E.S.; Reis, B; Gessner, J.; Wuertz, S; Dirks, R.P.H.

Submitted to Fish and Shellfish Immunology

Exposure to high temperatures can lead to thermotolerance in fish, which is hypothesized to potentially improve post-release survival in species under restocking programs, like Atlantic sturgeon. The aim of this study was to determine whether Atlantic sturgeon juveniles exposed to a four-week temperature treatment at 24ºC respond differently to a subsequent heat shock than juveniles reared at aquaculture temperatures (20ºC) (naive fish). Response to heat shock was assessed by mapping the liver transcriptome. In total 838 unique contigs were differentially expressed between the trained and the control group (592 downregulated, 261 upregulated and 15 down- or upregulated, depending on the condition), corresponding to genes involved in the response to heat, tissue damage, proteolysis and metabolism. Temperature-trained fish showed 2-4-fold fewer dysregulated contigs than naive fish, indicating their ability to maintain and recover homeostasis faster. During heat shock, *hspc1* was upregulated in both experimental groups, while *hspa1* and *dnaja4* were exclusively upregulated in the control. Heat shock markers previously proposed *in vitro,* like *hspb11a*, were not differentially expressed. Compensatory mechanisms were observed in addition to the heat shock response. Only two genes, *fgg* and *apnl*, were upregulated at nearly all timepoints in both groups. Peptidases were more strongly downregulated in control fish, which also showed a reduction in lipid metabolism during recovery. Keratins, *pck1, gadd45ga* and *gadd45gb* were differentially expressed between trained and control fish, and due to their roles in tissue protection and ER stress reduction, they might be responsible for the maintenance of the transcriptional homeostasis observed in trained fish.

Paper III: The genome of Atlantic sturgeon (*Acipenser oxyrinchus*) provides insights into genome evolution in tetraploids

Yebra-Pimentel, E.S.; Campo, A.; Shivaramu, S.; Henkel, C.V.; Jong-Raadsen, S.A.; Jansen, H.J.; Dirks, R.P.H.

Manuscript

The availability of a reference genome improves research on many fields; however, genomic resources are lacking for Atlantic sturgeon, which is under a restoration program in Europe. Sturgeons have high chromosome counts resulting from several lineage-specific whole-genome duplication events (WGD), but their ploidy status is still controversial. We have *de novo* assembled the Atlantic sturgeon genome sequence using Illumina and nanopore sequencing technologies. Ploidy level, genome size and heterozygosity were estimated using microsatellite loci analysis, flow cytometry and k-mer counting. Additionally, Hox gene clusters were analyzed to infer the evolutionary position of Atlantic sturgeon among gnathostomata and their expression was assessed in 4 stages of embryonic development. Microsatellite analysis showed disomic patterns in 5 loci and tetrasomy in 3 loci. Flow cytometry results suggested a genome size of 1.1 Gb, while k-mer counting inferred 652 Mb and high heterozygosity (5.36%). However, our nanopore assembly size was 1.28 Gb (N_{50} 0.97 Mb). We found 76 Hox genes and 1 pseudogene distributed over 7 clusters, with one of the HoxC clusters being lost. Hox paralogs shared an average identity of 92% at the protein level, while synteny was not fully conserved between the HoxA paralogs. Phylogenetic reconstruction indicated that they were resulting from a sturgeon-specific WGD (SR), independent from paddlefish and common for the two analyzed sturgeon species. The developed genomic resources serve as a basis for new research in Atlantic sturgeon, which is likely a paleotetraploid species in the process of rediploidization, based on the found paralogs, microsatellite results and discrepancy between estimated genome size and total assembly size.

General discussion

The results of the work presented in this thesis provide transcriptomic and genomic resources for Atlantic sturgeon research and evaluate the potential of temperature-training programs as a new rearing technique to improve post-release survival during the application of restoration programs. First, we have *de novo* assembled an Atlantic sturgeon transcriptome, identified all the heat shock proteins (HSP) genes, and studied their response to heat using a cell line derived from Atlantic sturgeon larvae. Second, we have studied the effects of temperature training of Atlantic sturgeon juveniles in response to a subsequent heat shock (HS) by using RNAseq to map liver gene expression. Lastly, we have generated genomic resources for Atlantic sturgeon, which provide a reference for future RNAseq experiments and has a great value for sturgeon research in all the fields, including physiology, conservation, and evolution.

Innovative rearing techniques in Atlantic sturgeon

In order to restore Atlantic sturgeon populations in the Baltic, it is important that we identify the main challenges juvenile fish face upon release and design innovative rearing methods to address them in order to improve fitness and survival of the fish used for restocking. However, the majority of the research directed to the development of such methods was performed in teleost fish, especially salmonids. These studies researched the effects of environmental enrichment (Salvanes et al., 2013; Solås et al., 2019), swimming exercise (Palstra et al., 2015) and foraging (Moberg et al., 2011), and temperature trials (Alexandre and Palstra, 2017; Prystay et al., 2017) on the post-release survival. Few studies have been conducted with sturgeons, and the very few that have done so, have focused on sturgeon species other than Atlantic sturgeon. These studies included swimming exercise (Du et al., 2014; Katopodis et al., 2019) and foraging behavior trials (Camara-Ruiz et al., 2018b); however, the vast majority of the research was directed towards the effects of temperature on the growth and survival rates (Eder et al., 2015; Kappenman et al., 2009), and the incidence of notochord and muscle deformities (Aidos et al., 2017; Linares-Casenave et al., 2013; Werner et al., 2007). Few of the available studies additionally included an evaluation of HSP expression in response to heat shock, and the ones that did,

studied a limited number of HSP family members, as sequence information of these genes is lacking for sturgeons.

Therefore, in **Paper I** we set out to *de novo* assemble a transcriptome reference for Atlantic sturgeon, annotate the complete HSP and HSF family members and study their response to heat shock in a cell line derived from Atlantic sturgeon larvae using an RNAseq approach.

The first main finding of this study was that the cell line gene expression profile was only mildly affected in response to heat shock at 30 ºC, while at 35 ºC a strong response was observed. This suggests a quite narrow temperature tolerance window in the cell line, which is in line with previous *in vivo* results (Linares-Casenave et al., 2013; Peng et al., 2016). Moreover, the vast majority of the dysregulated genes (3,020 contigs) were upregulated (2,302 contigs), which indicates a high transcriptional cellular demand in order to preserve homeostasis during and after heat shock.

The second main finding of **Paper I** is that the HSP family in Atlantic sturgeon is quite extensive and includes at least 76 genes, of which only 16 were responsive to either of the applied heat shock protocols. Furthermore, only 5 HSP genes (*hspb1, hspb11a, hsph3a, hsph2* and *hspc1*) and additional genes such as *clu*, *grb2*, *npr2* and *ccdc17* were upregulated in response to both of the applied heat shock protocols, of which *hspb11a* showed more than 3,000-fold upregulation.

In summary, this work provides transcriptomic resources for Atlantic sturgeon and an inventory of the HSP gene family members, which constitute a baseline for further research on temperature adaptation in sturgeons (**Paper II**). Additionally, we provided a general overview of heat-inducible HSP genes and proposed the best HSP gene markers for *in vivo* temperature trials. Although this cell line potentially represents an alternative for reducing experimentation with endangered animals, the proposed markers need validation during *in vivo* trials, as *in vitro* results might not correlate to the results at the organismal level.

Very few studies have focused on how temperature training affects the response to a subsequent heat shock event. Therefore, in **Paper II** we set out to describe the effects of a 4-week temperature-training period on the Atlantic sturgeon liver transcriptome in response to an ecologically relevant heat shock (+5 ºC), comparable to the highest temperatures measured in the restocking ecosystems (Odra River). Temperature-trained fish showed between 2 to 4 times fewer dysregulated contigs than naïve (untrained) fish, indicating their ability to maintain and recover homeostasis faster than the untrained counterparts, which

activate a more complex response to the same stimuli. Interestingly, three times fewer genes were dysregulated (838 contigs) when compared to the *in vitro* results (**Paper I**), most of them (592) downregulated. Only 3 HSP genes were upregulated (*hspc1, hspa1* and *dnaja4*), of which *hspa1* and *dnaja4* were restricted to the naïve fish group. Furthermore, the remaining proposed markers based on *in vitro* results, namely *hspb1, hspb11a, hsph3a* and *hsph2*, where not differentially expressed *in vivo*. In fact, a wide variety of genes involving various compensatory mechanisms were responsive to heat shock *in vivo*. For example, *gadd45g* and keratin genes, which are tissue-protective genes regulating thermogenic capacities and known to confer heat shock resistance in *Drosophila* (Gantner et al., 2014; Moskalev et al., 2012), were differentially expressed between trained and control fish. Additionally, *pck1*, involved in carbohydrate metabolism during gluconeogenesis, was strongly upregulated in temperature-trained fish before and during the final heat shock experiment, while naïve fish showed lower levels of *pck1* upregulation and restricted to the recovery period. Both groups showed downregulation of g*ck,* involved in glycolysis, and genes involved in lipid metabolism. Overall, only two genes, *hspc1* and *ccdc17*, were upregulated at all the analyzed timepoints both *in vivo* and in *vitro.* Compared to the epithelial cell line, the liver transcriptome represents a response at the organismal level, where both the neuroendocrine-mediated stress response and cellular responses might play a pivotal role during heat shock. The neuroendocrine system responds to stress by a rapid synthesis of stress hormones such as catecholamines and cortisol and their release to the bloodstream (Gamperl et al., 1994). In fact, cortisol regulates the synthesis of many enzymes related to carbohydrate and lipid metabolism like *pck1* (Choi et al., 2007; Qu and Ajuwon, 2018), and are known to suppress *hspa1* expression in the liver of rainbow trout and gill of rainbow trout and tilapia (Basu et al., 2001). Therefore, a higher level of cortisol would explain the reduced levels of *hspa1* and increased levels of *pck1* in temperature-trained fish compared to their non-adapted conspecifics. The different heat shock protocols applied in the cell line and juveniles, including temperature and heating rate, duration of exposure and sampling timepoint, could have also impacted the results; however, optimal temperature in the entire animal is lower than for the cell line, and temperatures higher than the applied in the *in vivo* experiment (24 ºC) have already been reported to reduce growth and increase mortality in sturgeon juveniles (Kappenman et al., 2009). Therefore, a heat shock performed at a higher temperature, more comparable to the *in vitro* experiment, would have had compromised the survival of Atlantic sturgeon juveniles. Previous studies

have reported tissue-specific expression patterns in response to heat shock, which adds an extra source of variation in the results of **Papers I and II.**

In summary, at the organismal level, there might be an important cortisol-mediated response that triggers different compensatory mechanisms, although cortisol measurements would be needed to confirm this hypothesis. Due to the low level of *hspa1* and high level of *pck1*, we also hypothesize that cortisol might be higher in trained fish and overall result in a reduced gene dysregulation and phenotypic advantage toward heat shock resistance. Therefore, *hspa1* might be the only good HSP marker gene to predict thermotolerance *in vivo* in the liver of Atlantic sturgeon juveniles, contradicting previous results in doctor fish (*Garra rufa*) and Arctic charr (Oksala et al., 2014; Quinn et al., 2011), which showed higher upregulation of several HSP genes in response to heat shock in warm-adapted populations.

Genomic resources for Atlantic sturgeon

The interest of genomic studies in sturgeon arises as a consequence of their high commercial value and the need for conservation measures. In order to study genomic variability in a given population, whole-genome re-sequencing (WGR) is a common practice in conservation genetics, often limited by the absence of a reference genome for read mapping and downstream analysis. The availability of a reference genome also has important applications in the development of gene markers and quantitative trait loci (QTL) analysis, which provide information about loci linked to traits of interest like local adaptations or resistance to disease and abiotic factors, such as temperature.

Therefore, in **Paper III** we have assembled the first Atlantic sturgeon genome using short and long read sequencing technologies, resulting in a quite contiguous (N_{50} ~1Mb) and complete genome (92.8%, according to BUSCO results).

The main finding of **Paper III** was the presence of 7 Hox clusters in the Atlantic sturgeon genome (HoxAα, HoxAβ, HoxBα, HoxBβ, HoxC, HoxDα, HoxDβ), presumably derived from a WGD event. These genomic clusters conserved many more Hox-coding (76) and neighboring genes than teleosts, which have undergone massive gene loss after the 3R. Two copies of each of the HoxA and HoxD clusters were previously described in the American Paddlefish, which were attributed to a paddlefish-specific whole-genome duplication (PR) (Crow et al., 2012). Furthermore, sturgeons (like paddlefish) retain one of the HoxD14

paralogs, which is absent in the genome of all the remaining actinopterygians studied to date.

We have additionally observed an unequal read coverage distribution of the final genome, with some contigs (e.g., the one containing the HoxC cluster) having double the coverage than others (e.g., contigs containing duplicated Hox clusters). Therefore, we tested whether the found Hox duplicated clusters in the Atlantic sturgeon genome were the result of haplotype assembly or a WGD event. Assuming that the heterozygosity (presumably ~5%) is equally distributed throughout the genome, the average amino acid sequence divergence between the Hox duplicates (~8%) is too high to be derived from heterozygosity, although slightly lower than paddlefish (9.76%) and much lower than teleosts (Crow et al., 2012). Moreover, all the annotated Hox-coding genes were transcriptionally active. Phylogenetic studies attributed these duplicates to a sturgeon-specific WGD event (SR), independent from the paddlefish PR, and taking place after the split between sturgeons and paddlefish (~204 Mya), and before the split between the basal Acipenseridae Sea clade and Atlantic and Pacific clades (~145 Mya). The paralog nature of the duplicated HoxA clusters in paddlefish has already been confirmed by molecular cytogenetic techniques, which located each paralog in two separate metacentric chromosomes (Symonová et al., 2017). In order to confirm the SR origin of the Hox duplicates, the same technique should be applied in Atlantic sturgeon; however, this could be challenging if some of the Hox clusters is located in one of the many sturgeon microchromosomes, which have often been described as "indistinguishable" (Billard and Lecointre, 2000).

A WGD event is the result from either the duplication of the species' own genome (autopolyploidization) or hybridization between two species (allopolyploidization) (Kihara and Ono, 1926). Genes duplicated by allopolyploidization (e.g. Common carp, dated from ~12 Mya) follow disomic patterns during meiosis, opposite to those resulting from autopolyploidization (e.g. Atlantic salmon, dated from ~80 Mya), which show full or partial tetrasomic patterns during meiosis and usually share higher sequence similarity (David et al., 2003; Lien et al., 2016). A process called rediploidization commonly follows WGD, switching from tetrasomic allelic patterns observed in paleotetraploids (4 alleles at 1 locus) to disomic patterns observed in functional diploids (2 alleles at 2 loci). Although teleost species that have only undergone the 3R duplication (~250 Mya) are today considered functional diploids (De La Serrana et al., 2014; Martin and Holland, 2014), while Atlantic salmon, which has suffered an extra (recent) duplication, shows a delayed rediploidization

and authors suggest that 10% of their genome may still retain residual tetrasomy (Lien et al., 2016). Our microsatellite results showed a mixture of disomic and tetrasomic allelic patterns in Atlantic sturgeon, suggesting that the WGD is the result of autopolyploidization. This pattern was already described by chromosome painting in sterlet, where authors stated that the transition from paleotetraploidy to diploidy might still be an ongoing process (Romanenko et al., 2015).

Finally, Hox gene complement is highly conserved in both paralog clusters, but upon inspection of syntenic genes in the HoxA paralog clusters, we identified asymmetric gene losses in 4 of the 18 scrutinized neighboring genes. Unlike medaka, which has lost all Hox-coding genes in cluster HoxCb but preserves some syntenic genes in two genomic locations (Kurosawa et al., 2006), in Atlantic sturgeon both HoxC-coding and neighboring genes are single copy, indicating that one of the HoxC paralog clusters has been either lost or not affected by the SR event.

Overall, this work presents for the first time a quite contiguous draft genome for Atlantic sturgeon using a combination of two sequencing technologies, which is especially useful in this case considering the presence of some genomic regions with high sequence similarity that would be impossible to locate properly relying only on short Illumina reads. We can conclude that both Atlantic and sterlet sturgeon have undergone an extra round of WGD (besides the 2R affecting all vertebrates), common to both species, and independent from paddlefish, that we have called SR. The SR likely affected the (currently extinct) common ancestor of all the 25 sturgeon species, but previous work suggests that a second round took place in the Atlantic and Pacific clades separately in addition to a third round affecting only the shortnose sturgeon (*A. brevirostrum*) (Ludwig et al., 2001; Peng et al., 2007; Fontana et al., 2007; Vasilev, 2009). More sequence information is needed in order to build a genome reference with chromosome resolution for Atlantic sturgeon. Molecular chromosome characterization should follow, in order to resolve a series of hidden unknowns in sturgeon biology such as the sex determination mechanisms and interspecific hybridization, which have great relevance on sturgeon aquaculture and conservation.

General conclusions

The main aim of this thesis was to generate genomic and transcriptomic resources for sturgeon that can be used to improve research on the species and application in aquaculture and restoration programs.

In this thesis, we have firstly generated a reference transcriptome for Atlantic sturgeon and annotated all the identified HSP gene family members. Both the transcriptome and the HSP annotations represent an important tool for sturgeon research, as it allows the use of a reference transcriptome for RNAseq experiments as well as sequence information for qPCR primer design in larger-scale studies.

The results show that the epithelial cell line has higher transcriptional activity in response to heat shock than the Atlantic sturgeon juvenile livers, while in both experiments only a small fraction of the 76 annotated HSP genes was responsive to heat shock. We propose that the milder response observed during *in vivo* experiments results from the activation of compensatory mechanisms, especially the neuroendocrine system, resulting in tissue protection and thermogenic capacities. The exposure of fish to a 4-week temperature-training protocol results in a stronger activation of those mechanisms resulting overall in an increase in transcriptomic homeostasis in response to a subsequent heat shock compared to untrained fish. We therefore propose that temperature-training protocols should be considered as a new rearing technique for fish used for restocking.

Additionally, using a combination of short and long-read sequencing technologies we have assembled the first reference genome for Atlantic sturgeon and, based on Hox cluster identification followed by synteny and phylogenetic studies, confirmed the presence of an independent sturgeons-specific genome duplication (SR). Based on previous dating information we dated the duplication from 205 to 145 Mya, but further analysis should be performed in order to provide a narrower dating. Based on microsatellite analysis showing residual tetrasomy and previous studies in other members of the Order, we conclude that the Atlantic sturgeon is a paleotetraploid, in which the rediploidization process is still ongoing.

Future perspectives

In this thesis we have filled important knowledge gaps in sturgeon research, which will contribute to increasing the value of all sturgeon research. However, we have also opened new knowledge gaps, which should be addressed in future studies. The main issues that require further investigations concern the resolution of the ploidy level. Molecular cytogenetics should be performed to finally conclude that the duplicated Hox clusters found are indeed a result of SR. In order to build an assembly to the chromosome level, more long read data and possibly Hi-C are required, in order to create a genome assembly that is finalized and for which each scaffold represents a chromosome. In order to further support the independence of the SR and PR duplication events in sturgeon and paddlefish, phylogeny analysis on more genes of different molecular families and whole genome comparisons should be performed when the paddlefish genome is available.

Additionally, it is widely known that the response to heat shock varies greatly among different tissues, and we have only studied the liver. Fin clips or blood markers should also be studied in order to generate minimally invasive gene markers for the assessment of thermotolerance. It would also be important to compare the results of the cell line study to the results in the *in vivo* epithelium, which would more conclusively state the relevance of this cell line in sturgeon research. Additionally, survival experiments in trained and untrained juveniles should also be carried out. Empirical studies should aim to elucidate how characteristics such as age of the fish and treatment duration should be implemented in the temperature training towards increased survival rates. Also, other temperature-treatment protocols might be performed during other stages of development, such as larval stages or sexually mature stages.

References

Aidos, L., Pinheiro-Valente, L.M., Sousa, V., Lanfranchi, M., Domeneghini, C., Di Giancamillo, A. 2017. Effects of different rearing temperatures on muscle development and stress response in the early larval stages of *Acipenser baerii*. *Eur. J. Histochem*. 61, 287–294. doi:10.4081/ejh.2017.2850

Alexandre, C.M., Palstra, A.P. 2017. Effect of short-term regulated temperature variations on the swimming economy of Atlantic salmon smolts. *Conserv. Physiol.* 5, doi:10.1093/conphys/cox025

Allen, S.K. 1983. Flow cytometry: Assaying experimental polyploid fish and shellfish. *Aquaculture.* doi:10.1016/0044-8486(83)90412-X

Allen, S.K., Stanley, J. 1978. Reproductive sterility in polyploid Brook trout, *Salvelinus fontinalis*. *Trans. Am. Fish. Soc*. 107, 473–478. doi:10.1577/1548-8659(1978)107<473

Amemiya, C.T., Alföldi, J., Lee, A.P., Fan, S., Philippe, H., MacCallum, I., Braasch, I. 2013. The African coelacanth genome provides insights into tetrapod evolution. *Nature* 496, 311–316. doi:10.1038/nature12027

Anders, S., Huber, W. 2010. Differential expression analysis for sequence count data. *Genome Bio*l. doi:10.1186/gb-2010-11-10-r106

Andrews, S. 2010. FASTQC. A quality control tool for high throughput sequence data. 2010, Http://Www.Bioinformatics.Babraham.Ac.Uk/Projects/Fastqc/. doi:v.0.11.5

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.M., Dehal, P., et al. 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science*. 297, 1301–1310. doi:10.1126/science.1072104

Arabidopsis Genome Initiative, T., 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815.

Asmann, Y.W., Klee, E.W., Thompson, E.A., Perez, E.A., Middha, S., Oberg, A.L. et. al. 2009. 3’ tag digital gene expression profiling of human brain and universal reference RNA using Illumina Genome Analyzer. *BMC Genomics*. doi:10.1186/1471-2164-10-531

Austin, C.M., Tan, M.H., Harrisson, K.A., Lee, Y.P., Croft, L.J., Sunnucks, P. et al. 2017. De novo genome assembly and annotation of Australia’s largest freshwater fish, the Murray cod (*Macchullochella peelii*) from Illumina and Nanopore sequencing reads. *Gigascience* 1–19. doi:10.1093/gigascience/gix063

Bartel, R., 1997. Preliminary results on restoration of Atlantic salmon (*Salmo salar* L) in Poland. OT: Wstepne wyniki restyucii w Polsce poglowia lososia atlantyckiego (*Salmo salar* L.). *Arch. Polish Fish*. 5, 201–207.

Basu, N., Nakano, T., Grau, E.G., Iwama, G.K. 2001. The effects of cortisol on heat shock protein 70 levels in two fish species. *Gen. Comp. Endocrinol.* doi:10.1006/gcen.2001.7688

Beato, M.S., Marcacci, M., Schiavon, E., Bertocchi, L., Di Domenico, M., Peserico, A. et al. 2018. Identification and genetic characterization of bovine enterovirus by combination of two next generation sequencing platforms. *J. Virol. Methods.* doi:10.1016/j.jviromet.2018.07.002

Beitinger, T., Bennett, W., McCauley, R. 2000. Temperature tolerances of North American freshwater fishes exposed to dynamic changes in temperature. *Environ. Biol. Fishes.* 58, 237–275. doi:10.1023/A:1007676325825

Bemis, W.E., Findeis, E.K., Grande, L.1997. An overview of Acipenseriformes. *Environ. Biol. Fishes.* 48, 25–71. doi:10.1007/0-306-46854-9_4

Betancur, R., Broughton, R.E., Wiley, E.O., Carpenter, K., López, J.A., Li, C. et al. 2013.

The Tree of Life and a New Classification of Bony Fishes. PLOS Curr. 5, 1–45. doi:10.1371/currents.tol.53ba26640df0ccaee75bb165c8c26288.
Bian, C., Hu, Y., Ravi, V., Kuznetsova, I.S., Shen, X., Mu, X. et al. 2016. The Asian arowana (*Scleropages formosus*) genome provides new insights into the evolution of an early lineage of teleosts. *Sci. Rep.* 6, 1–17. doi:10.1038/srep24501
Billard, R., Lecointre, G. 2000. Biology and conservation of sturgeon and paddlefish. *Rev. Fish Biol. Fish*. 10, 355–392. doi:10.1023/A:1012231526151
Blacklidge, K.H., Bidwell, C.A. 1993. Three ploidy levels indicated by genome quantification in Acipenseriformes of North America. *J. Heredity*. 84, 427–430.
Boscari, E., Vidotto, M., Martini, D., Papetti, C., Ogden, R., Congiu, L. 2015. Microsatellites from the genome and the transcriptome of the tetraploid Adriatic sturgeon, *Acipenser naccarii* (Bonaparte, 1836) and cross-species applicability to the diploid beluga sturgeon, *Huso huso* (Linnaeus, 1758). *J. Appl. Ichthyol.* 31, 977–983. doi:10.1111/jai.12906
Braasch, I., Gehrke, A.R., Smith, J.J., Kawasaki, K., Manousaki, T., Pasquier, J. et al. 2016. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nat. Genet.* 48, 427–437. doi:10.1038/ng.3526
Bray, N.L., Pimentel, H., Melsted, P., Pachter, L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* doi:10.1038/nbt.3519
Bronzi, P., Chebanov, M., Michaels, J.T., Wei, Q., Rosenthal, H., Gessner, J. 2019. Sturgeon meat and caviar production: Global update 2017. *J. Appl. Ichthyol*. 35, 257–266. doi:10.1111/jai.13870
Bronzi, P., Rosenthal, H. 2014. Present and future sturgeon and caviar production and marketing: A global market overview. *J. Appl. Ichthyol*. 30, 1536–1546. doi:10.1111/jai.12628
Bronzi, P., Rosenthal, H., Gessner, J. 2011. Global sturgeon aquaculture production: An overview. *J. Appl. Ichthyol*. 27, 169–175. doi:10.1111/j.1439-0426.2011.01757.x
C elegans Sequencing Consortium, T. 1998. Genome Sequence of the Nematode *C . elegans* : A Platform for Investigating Biology. *Science*. 282, 2012–2018. doi:10.1126/science.282.5396.2012
Camara-Ruiz, M., Espirito Santo, C., Gessner, J., Wuertz, S. 2018a. Environmental enrichment on neural plasticity in juvenile Baltic sturgeon (*Acipenser oxyrichus*). http://dx.doi.org/10.1101/491910. The.
Camara-Ruiz, M., Espirito Santo, C., Gessner, J., Wuertz, S. 2018b. How to improve foraging efficiency for restocking measures of juvenile Baltic sturgeon. *Aquaculture.* 49.
Carbon, S., Dietze, H., Lewis, S.E., Mungall, C.J., Munoz-Torres, M.C., Basu, S. et al. 2017. Expansion of the gene ontology knowledgebase and resources: The gene ontology consortium. *Nucleic Acids Res.* 45, 331–338. doi:10.1093/nar/gkw1108
Choi, C.Y., Min, B.H., Jo, P.G., Chang, Y.J. 2007. Molecular cloning of PEPCK and stress response of black porgy (*Acanthopagrus schlegeli*) to increased temperature in freshwater and seawater. *Gen. Comp. Endocrinol.* 152, 47–53. doi:10.1016/j.ygcen.2007.02.019
Ciudad, J., Cid, E., Velasco, A., Lara, J.M., Aijón, J., Orfao, A. 2002. Flow cytometry measurement of the DNA contents of G0/G1 diploid cells from three different teleost fish species. *Cytometry* 48, 20–25. doi:10.1002/cyto.10100
Crosland, M.W.J., Crozier, R.H. 1985. *Myrmecia pilulosa*, and ant with only one pair of chromosomes. *Science*. 231, 1278. doi:10.1126/science.231.4743.1278
Crow, K.D., Smith, C.D., Cheng, J.F., Wagner, G.P., Amemiya, C.T. 2012. An independent genome duplication inferred from Hox paralogs in the American paddlefish- A

representative basal ray-finned fish and important comparative reference. *Genome Biol. Evol*. 4, 937–953. doi:10.1093/gbe/evs067
David, L., Blum, S., Feldman, M.W., Lavi, U., Hillel, J. 2003. Recent duplication of the common carp (*Cyprinus carpio L.*) Genome as revealed by analyses of microsatellite loci. *Mol. Biol. Evol.* doi:10.1093/molbev/msg173
De Coster, W., D'Hert, S., Schultz, D.T., Cruts, M., Van Broeckhoven, C. 2018. NanoPack: Visualizing and processing long-read sequencing data. Bioinformatics. doi:10.1093/bioinformatics/bty149
de Crécy-Lagard, V., Hanson, A. 2013. Comparative Genomics, in: Brenner's Encyclopedia of Genetics: Second Edition. doi:10.1016/B978-0-12-374984-0.00299-0
De La Serrana, D.G., Mareco, E.A., Johnston, I.A. 2014. Systematic variation in the pattern of gene paralog retention between the teleost superorders ostariophysi and acanthopterygii. *Genome Biol. Evol*. 6, 981–987. doi:10.1093/gbe/evu074
Division, B., Biology, C., Cruz, S., Kasianowicz, J. 1996. Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl. Acad. Sci. U. S. A.* 93, 13770–13773.
Du, H., Wei, Q.W., Xie, X., Shi, L.L., Wu, J.M., Qiao, X.M. et al. 2014. Improving swimming capacity of juvenile Dabry's sturgeon, (*Acipenser dabryanus* Duméril, 1869) in current-enriched culture tanks. *J. Appl. Ichthyol*. 30, 1445–1450. doi:10.1111/jai.12591
Eder, B.L., Steffensen, K.D., Haas, J.D., Adams, J.D. 2015. Short-term survival and dispersal of hatchery-reared juvenile pallid sturgeon stocked in the channelized Missouri River. *J. Appl. Ichthyol*. 31, 991–996. doi:10.1111/jai.12881
Elbers, J.P., Rogers, M.F., Perelman, P.L., Proskuryakova, A.A., Serdyukova, N.A., Johnson, W.E. et al. 2019. Improving Illumina assemblies with Hi-C and long reads: an example with the North African dromedary. *Mol. Ecol. Resour.* doi:10.1111/1755-0998.13020
Elvira, B., Leal, S., Doadrio, I., Almodóvar, A. 2015. Current Occurrence of the Atlantic Sturgeon Acipenser oxyrinchus in Northern Spain: A New Prospect for Sturgeon Conservation in Western Europe. PLoS One 10, e0145728. doi:10.1371/journal.pone.0145728
Ewing, B., Hillier, L.D., Wendl, M.C. 1998. Base-Calling of Automated Sequencer Traces Using Phred. *Genome Res*. doi:10.1101/gr.8.3.186
Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R. et al. 2016. The reactome pathway knowledgebase. *Nucleic Acids Res*. doi:10.1093/nar/gkv1351
Fields, P.A., Graham, J.B., Rosenblatt, R.H., Somero, G.N. 1993. Effects of expected global climate change on marine faunas. *Trends Ecol. Evol.* 8, 361–367. doi:10.1016/0169-5347(93)90220-J
Finstad, B., Jonsson, N. 2001. Factors influencing the yield of smolt releases in Norway. *Nord. J. Freshw. Res*.
Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R. et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*. 269, 496–512. doi:10.1126/science.7542800
Fontana, F., Lanfredi, M., Kirschbaum, F., Garrido-Ramos, M.A., Robles, F., Forlani, A. et al. 2008. Comparison of karyotypes of *Acipenser oxyrinchus* and *A. sturio* by chromosome banding and fluorescent in situ hybridization. *Genetica*. 132, 281–286. doi:10.1007/s10709-007-9171-4
Fopp-Bayat, D., Kucinski, M., Liszewski, T., Teodorowicz, T., Laczynska, B., Lebeda, I. 2015. Genetic protocol of Atlantic sturgeon *Acipenser oxyrinchus* (L .) fry for

restocking the Vistula river, Poland. *J. Surv. Fish. Sicences*. 2, 1–10. doi:10.18331/SFS2015.2.1.1
Fredrich, F., Kapusta, A., Duda, A., Ebert, M., Gessner, J. 2008. Migratory behavior of young sturgeon, acipenser oxyrinchus mitchill, in the Oder River drainage. Preliminary results of a radio telemetric study in the Drawa River, Poland. *Arch. Polish Fish*. 16, 105–117. doi:10.2478/s10086-008-0009-1
Freyhof, J., Brooks, E. 2011. European Red List of Freshwater Fishes. doi:10.2779/85903
GAIN.Report, 2014. GAIN Report. Fish and Seafood Production and Trade Update. Moscow:Russia Federation. Retrieved from (GAIN Rep. Number RS1419. Date 3/18/2014). IUCN.
Gallagher, M.D., Matejusova, I., Nguyen, L., Ruane, N.M., Falk, K., Macqueen, D.J. 2018. Nanopore sequencing for rapid diagnostics of salmonid RNA viruses. *Sci. Rep*. doi:10.1038/s41598-018-34464-x
Gantner, M.L., Hazen, B.C., Conkright, J., Kralli, A. 2014. GADD45 regulates the thermogenic capacity of brown adipose tissue. *Proc. Natl. Acad. Sci.* 111, 11870–11875. doi:10.1073/pnas.1406638111
Gessner, J., 2000. Reasons for the decline of *Acipenser sturio* L. 1758 in central Europe, and attempts at its restoration. *Bol. Inst. Esp. Oceanogr.* 16, 117–126.
Gessner, J., Arndt, G.M., Kapusta, A., Gushin, S.A., Pilinkovskij, A., Povliunas, J. et al. 2019. HELCOM -Action Plan for the protection and recovery of the Baltic sturgeon *Acipenser oxyrinchus* in the Balstic Sea area. Helicom Conv. March 7th, 99p.
Gessner, J., Bartel, R., 2000. Sturgeon spawning grounds in the Odra River tributaries : A first assessment. *Bol. Inst. Esp. Oceanogr.* 16, 127–137.
Goffeau, A., Barrel, B., Bussey, R., Davis, R., Dujon, B., Feldmann, H. et al. 1996. Life with 6000 Genes. *Science*. 274, 546–567. doi:jyu
Gong, L., Wong, C.H., Idol, J., Ngan, C.Y., Wei, C.L. 2019. Ultra-long Read Sequencing for Whole Genomic DNA Analysis. *J. Vis. Exp*. 1–10. doi:10.3791/58954
Goodwin, S., Gurtowski, J., Ethe-Sayers, S., Deshpande, P., Schatz, M.C., McCombie, W.R. 2015. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Res*. doi:10.1101/gr.191395.115
Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I. et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* doi:10.1038/nbt.1883
Gradil, A.M., Wright, G.M., Speare, D.J., Wadowska, D.W., Purcell, S., Fast, M.D. 2014. The effects of temperature and body size on immunological development and responsiveness in juvenile shortnose sturgeon (*Acipenser brevirostrum*). *Fish Shellfish Immunol*. 40, 545–555. doi:10.1016/j.fsi.2014.07.036
Griffith, M., Griffith, O.L., Mwenifumbo, J., Goya, R., Morrissy, A.S., Morin, R.D. et al. 2010. Alternative expression analysis by RNA sequencing. *Nat. Methods*. doi:10.1038/nmeth.1503
Grunow, B., Noglick, S., Kruse, C., Gebert, M. 2011. Isolation of cells from Atlantic sturgeon *Acipenser oxyrinchus oxyrinchus* and optimization of culture conditions. *Aquat. Biol*. 14, 67–75. doi:10.3354/ab00383
Guo, L.T., Wang, S.L., Wu, Q.J., Zhou, X.G., Xie, W., Zhang, Y.J. 2015. Flow cytometry and K-mer analysis estimates of the genome sizes of *Bemisia tabaci* B and Q (Hemiptera: Aleyrodidae). *Front. Physiol*. doi:10.3389/fphys.2015.00144
Gut, I.G. 2013. New sequencing technologies. *Clin. Transl. Oncol*. 15, 879–881. doi:10.1007/s12094-013-1073-6
Hardie, D.C., Hebert, P.D. 2004. Genome-size evolution in fishes. *Can. J. Fish. Aquat. Sci*. 61, 1636–1646. doi:10.1139/f04-106

Hardie, D.C., Hebert, P.D.N. 2003. The nucleotypic effects of cellular DNA content in cartilaginous and ray-finned fishes. *Genome / Natl. Res. Counc. Canada = Génome / Cons. Natl. Rech*. Canada 46, 683–706. doi:10.1139/g03-040

Hare, E.E., Johnston, J.S. 2011. Genome Size Determination Using Flow Cytometry of Propidium Iodide-Stained Nuclei. doi:10.1007/978-1-61779-228-1_1

Havelka, M., Bytyutskyy, D., Symonová, R., Ráb, P., Flajšhans, M. 2016. The second highest chromosome count among vertebrates is observed in cultured sturgeon and is associated with genome plasticity. *Genet. Sel. Evol.* 48, 12. doi:10.1186/s12711-016-0194-0

Havelka, M., Hulák, M., Bailie, D.A., Prodöhl, P.A., Flajšhans, M. 2013. Extensive genome duplications in sturgeons: New evidence from microsatellite data. *J. Appl. Ichthyol.* 29, 704–708. doi:10.1111/jai.12224

Havelka, M., Kašpar, V., Hulák, M., Flajšhans, M. 2011. Sturgeon genetics and cytogenetics: A review related to ploidy levelsand interspecific hybridization. *Folia Zool*. 60, 93–103.

Havelka, M., Zhou, H., Hagihara, S., Ichimura, M., Fujimoto, T., Yamaha, E. et al. 2017. Spontaneous polyploidization in critically endangered *Acipenser mikadoi. Fish. Sci.* 83, 587–595. doi:10.1007/s12562-017-1083-3

Henkel, C.V., Burgerhout, E., de Wijze, D.L., Dirks, R.P., Minegishi, Y., Jansen, H.J. et al. 2012. Primitive duplicate hox clusters in the european eel's genome. *PLoS One* 7. doi:10.1371/journal.pone.0032231

Henson, J., Tischler, G., Ning, Z. 2012. Next-generation sequencing and large genome assemblies. *Pharmacogenomics*. doi:10.2217/pgs.12.72

Hirakawa, H., Sumitomo, K., Hisamatsu, T., Nagano, S., Shirasawa, K., Higuchi, Y. et al. 2019. *De novo* whole-genome assembly in *Chrysanthemum seticuspe* , a model species of Chrysanthemums, and its application to genetic and gene discovery analysis. *DNA Res*. doi:10.1093/dnares/dsy048

Hofmann, G.E., Todgham, A.E. 2010. Living in the Now: Physiological Mechanisms to Tolerate a Rapidly Changing Environment. *Annu. Rev. Physiol.* 72, 127–145. doi:10.1146/annurev-physiol-021909-135900

Howe, K., Clark, M.D., Torroja, C.F., Torrance, J., Berthelot, C., Muffato, M. et al. 2013. The zebrafish reference genome sequence and its relationship to the human genome. *Nature*. 496, 498–505. doi:10.1038/nature12111

Huang, D.W., Sherman, B.T., Lempicki, R.A. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* doi:10.1038/nprot.2008.211

Human Genome Sequencing Consortium, I. 2004. Finishing the euchromatic sequence of the human genome. *Nature*. 431, 931–945.

Inoue, J.G., Miya, M., Tsukamoto, K., Nishida, M. 2003. Basal Actinopterygian relationships: a mitogenomic perspective on the phylogeny of the "ancient fish." *Mol. Phylogenet. Evol*. 26, 110–120.

IPCC. 2013. Climate Change 2013. The Physical Science Bases. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge Univ. Press. Cambridge, United Kindom New York, NY, USA 881 pp. doi:10.1002/mds.26688

IUCN. 2018. The IUCN Red List of Threatened Species (Version 2018-1).

Jain, M. 2017. Nanopore sequencing and assembly of a human genome with ultra-long reads. *bioRxiv*. doi:http://dx.doi.org/10.1101/128835

Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A. et al. 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol*.

doi:10.1038/nbt.4060
Jansen, H., Dirks, R.P., Liem, M., Henkel, C.V., van Heusden, G.P.H., Lemmers, R.J.L.F. et al. 2017. De novo whole-genome assembly of a wild type yeast isolate using nanopore sequencing. *F1000Research.* 6, 618. doi:10.12688/f1000research.11146.1
Jansen, H.J., Liem, M., Jong-Raadsen, S.A., Dufour, S., Weltzien, F.A., Swinkels, W. et al. 2017b. Rapid de novo assembly of the European eel genome from nanopore sequencing reads. *Sci. Rep*. doi:10.1038/s41598-017-07650-6
Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M. et al. 2014. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* 24, 1384–1395. doi:10.1101/gr.170720.113
Kappenman, K.M., Fraser, W.C., Toner, M., Dean, J., Webb, M.A.H. 2009. Effect of Temperature on Growth, Condition, and Survival of Juvenile Shovelnose Sturgeon. *Trans. Am. Fish. Soc*. 138, 927–937. doi:10.1577/t07-265.1
Katopodis, C., Cai, L., Johnson, D. 2019. Sturgeon survival: The role of swimming performance and fish passage research. *Fish. Res.* 212, 162–171. doi:10.1016/j.fishres.2018.12.027
Keyvanshokooh, S., Gharaei, A. 2010. A review of sex determination and searches for sex-specific markers in sturgeon. *Aquac. Res*. 41, 1–7. doi:10.1111/j.1365-2109.2009.02463.x
Khandelwal, S. 1990. Chromosome evolution in the genus *Ophioglossum* L. *Bot. J. Linn. Soc*. 102, 205–217.
Khodorevskaya, R.P., Dovgopol, G.F., Zhuravleva, O.L. 1995. Formation of commercial sturgeon (*Acipenseriae*) stocks. *Proc. Int. Symp. sturgeons A.D. Gersh*, 137–150. doi:10.1007/0-306-46854-9_11
Kihara, H., Ono, T. 1926. Chromosomenzahlen und systematische Gruppierung der Rumex-Arten. Zeitschrift für Zellforsch. und Mikroskopische Anat. doi:10.1007/BF00391215
Kikugawa, K., Katoh, K., Kuraku, S., Sakurai, H., Ishida, O., Iwabe, N. et al. 2004. Basal jawed vertebrate phylogeny inferred from multiple nuclear DNA-coded genes. *BMC Biol*. 11, 1–11.
King, L., Lubinski, A., Spidle, P., 2001. Microsatellite DNA variation in Atlantic sturgeon (Acipenser). *Conserv. Genet*. 2, 103–119.
Kirschbaum, F., Ludwig, A., Hensel, E., Wurt, S., Kloas, W., Williot, P. et al. 2015. Status of the Project on Protection and Restoration of Atlantic sturgeon in Germany: Background, Current situation, and perspectives.
Klinkhardt, M., Tesche, M., Greven, H. 1995. Database of fish chromosomes. Westarp Wissenschaften, Magedeburg, Ger. doi:10.1093/database/baw012
Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., Phillippy, A.M. 2017. Canu : scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*. 27, 722–736. doi:10.1101/gr.215087.116.Freely
Krieger, J., Fuerst, P.A. 2002. Evidence for a slowed rate of molecular evolution in the order *Acipenseriformes*. *Mol. Biol. Evol*. 19, 891–7.
Krieger, J., Hett, A.K., Fuerst, P.A., Artyukhin, E., Ludwig, A. 2008. The molecular phylogeny of the order *Acipenseriformes* revisited. *J. Appl. Ichthyo*l. 24, 36–45. doi:10.1111/j.1439-0426.2008.01088.x
Kurosawa, G., Takamatsu, N., Takahashi, M., Sumitomo, M., Sanaka, E., Yamada, K. et al. 2006. Organization and structure of hox gene loci in medaka genome and comparison with those of pufferfish and zebrafish genome*s. Gene*. doi:10.1016/j.gene.2005.11.015
Langmead, B., Salzberg, S.L. 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*. 9, 357–359. doi:10.1038/nmeth.1923
Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-efficient

alignment of short DNA sequences to the human genome. Genome Biol. doi:10.1186/gb-2009-10-3-r25
Lebeda, I., Steinbach, C., Flajšhans, M., 2018. Flow cytometry for assessing the efficacy of interspecific gynogenesis induction in sturgeon 1819–1831. doi:10.1111/jfb.13623
Leger, A., Leonardi, T., 2019. pycoQC, interactive quality control for Oxford Nanopore Sequencing. J. Open Source Softw. doi:10.21105/joss.01236
Levin, A., 1995. Russian sturgeon, Acipenser gueldenstaedti Brandt, stocking in the Volga-Caspian Basin. Proc. Int. Symp. sturgeons 178–188. doi:10.1111/j.1439-0426.2011.01725.x
Li, H., 2018. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics. doi:10.1093/bioinformatics/bty191
Li, H., Durbin, R., 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. doi:10.1093/bioinformatics/btp698
Li, L.M., Wei, L., Jiang, H.Y., Zhang, Y., Zhang, X.J., Yuan, L.H., 2015. Rapid development of polymorphic microsatellite markers for the Amur sturgeon (Acipenser schrenckii) using next-generation sequencing technology. Genet. Mol. Res. 14, 7910–7913.
Lien, S., Koop, B.F., Sandve, S.R., Miller, J.R., Matthew, P., Leong, J.S. et al. 2016. The Atlantic salmon genome provides insights into rediploidization. Nature 533, 200–205. doi:10.1038/nature17164
Linares-Casenave, J., Werner, I., Van Eenennaam, J.P., Doroshov, S.I. 2013. Temperature stress induces notochord abnormalities and heat shock proteins expression in larval green sturgeon (Acipenser medirostris Ayres 1854). J. Appl. Ichthyol. 29, 958–967. doi:10.1111/jai.12220
Lu, H., Giordano, F., Ning, Z. 2016. Oxford Nanopore MinION Sequencing and Genome Assembly. Genomics, Proteomics Bioinforma. 14, 265–279. doi:10.1016/j.gpb.2016.05.004
Ludwig, A.,2006. A sturgeon view on conservation genetics. Eur. J. Wildl. Res. 52, 3–8. doi:10.1007/s10344-005-0006-2
Ludwig, A., Arndt, U., Lippold, S., Benecke, N., Debus, L., King, T.L. et al. 2008. Tracing the first steps of American sturgeon pioneers in Europe. BMC Evol. Biol. 8, 221. doi:10.1186/1471-2148-8-221
Ludwig, A., Belfiore, N.M., Pitra, C., Svirsky, V., Jenneckens, I. 2001. Genome duplication events and functional reduction of ploidy levels in sturgeon (Acipenser, Huso and Scaphirhynchus). Genetics 158, 1203–1215.
Ludwig, A., Debus, L., Lieckfeldt, D., Wirgin, I., Benecke, N., Jenneckens, I. et al. 2002. When the American sea sturgeon swam east. Nature 419, 447–448. doi:10.1038/419447a
Luo, D., Li, Y., Zhao, Q., Zhao, L., Ludwig, A., Peng, Z. 2019. Highly Resolved Phylogenetic Relationships within Order Acipenseriformes According to Novel Nuclear Markers. Genes (Basel). doi:10.3390/genes10010038
Marçais, G., Kingsford, C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27, 764–770. doi:10.1093/bioinformatics/btr011
Martin, K.J., Holland, P.W.H. 2014. Enigmatic orthology relationships between hox clusters of the African butterfly fish and other teleosts following ancient whole-genome duplication. Mol. Biol. Evol. 31, 2592–2611. doi:10.1093/molbev/msu202
McClintock, J.M., Carlson, R., Mann, D.M., Prince, V.E. 2001. Consequences of Hox gene duplication in the vertebrates: an investigation of the zebrafish Hox paralogue group 1 genes. Development.
Meller, A., Nivon, L., Branton, D., 2001. Voltage-driven DNA translocations through a

nanopore. Phys. Rev. Lett. 86, 3435–3438. doi:10.1103/PhysRevLett.86.3435
Metscher, B.D., Ahlberg, P.E., 1999. Zebrafish in context: Uses of a laboratory model in comparative studies. Dev. Biol. doi:10.1006/dbio.1999.9230
Meyer, A., Schartl, M., 1999. Gene and genome duplications in vertebrates: The one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. Curr. Opin. Cell Biol. 11, 699–704. doi:10.1016/S0955-0674(99)00039-3
Michael, T., Florian, J., Bemm, F., Motley, T., Sandoval, J.P., Loudet, O. et al. 2018. High contiguity Arabisponsis thaliana genome assembly with a single nanopore flow cell. Nat. Commun. 9, 541.
Mikheyev, A.S., Tin, M.M.Y. 2014. A first look at the Oxford Nanopore MinION sequencer. Mol. Ecol. Resour. doi:10.1111/1755-0998.12324
Moberg, O., Braithwaite, V.A., Jensen, K.H., Salvanes, A.G.V. 2011. Effects of habitat enrichment and food availability on the foraging behaviour of juvenile Atlantic Cod (Gadus morhua L). Environ. Biol. Fishes. doi:10.1007/s10641-011-9803-5
Mohler, J.W., King, M.K., Farrell, P.R. 2004. Growth and Survival of First-Feeding and Fingerling Atlantic Sturgeon under Culture Conditions. N. Am. J. Aquac. doi:10.1577/1548-8454(2000)062<0174:gasoff>2.3.co;2
Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., Wold, B., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods 5, 621–628. doi:10.1038/nmeth.1226
Moskalev, A., Plyusnina, E., Shaposhnikov, M., Shilova, L., Kazachenok, A., Zhavoronkov, A., 2012. The role of D-GADD45 in oxidative, thermal and genotoxic stress resistance. Cell Cycle 11, 4222–4241. doi:10.4161/cc.22545
Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y. et al. 2011. Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res. doi:10.1093/nar/gkr344
Near, T.J., Eytan, R.I., Dornburg, A., Kuhn, K.L., Moore, J.A., Davis, M.P. et al. 2012. Resolution of ray-finned fish phylogeny and timing of diversification. Proc. Natl. Acad. Sci. doi:10.1073/pnas.1206625109
Ojima, Y., Yamamoto, K., 1990. Cellular DNA contents of fishes determined by flow cytometry. La Kromosomo II 57, 1871–1888.
Oksala, N.K.J., Ekmekçi, F.G., Özsoy, E., Kirankaya, Ş., Kokkola, T., Emecen, G. et al. 2014. Natural thermal adaptation increases heat shock protein levels and decreases oxidative stress. Redox Biol. 3, 25–28. doi:10.1016/j.redox.2014.10.003
Palstra, A.P., Mes, D., Kusters, K., Roques, J.A.C., Flik, G., Kloet, K. et al. 2015. Forced sustained swimming exercise at optimal speed enhances growth of juvenile yellowtail kingfish (Seriola lalandi). Front. Physiol. doi:10.3389/fphys.2014.00506
Pedersen, R.A., 1971. DNA content, ribosomal gene multiplicity, and cell size in fish. J. Exp. Zool. 177, 65–78. doi:10.1002/jez.1401770108
Peng, G., Zhao, W., Shi, Z., Chen, H., Liu, Y., Wei, J. et al. 2016. Cloning HSP70 and HSP90 genes of kaluga (Huso dauricus) and the effects of temperature and salinity stress on their gene expression. Cell Stress Chaperones 21, 349–359. doi:10.1007/s12192-015-0665-1
Peng, Z., Ludwig, A., Wang, D., Diogo, R., Wei, Q., He, S. 2007. Age and biogeography of major clades in sturgeons and paddlefishes (Pisces: Acipenseriformes). Mol. Phylogenet. Evol. 42, 854–862. doi:10.1016/j.ympev.2006.09.008
Pevzner, P.A., Tang, H., Waterman, M.S. 2001. An Eulerian path approach to DNA fragment assembly. PNAS 98, 9748–9753.
Pierre, R., Parauka, F. 2006. (U.S. Fish & Wildlife Service). Acipenser oxyrinchus. The IUCN Red List of Threatened Species 2006 e.T245A107,

http://dx.doi.org/10.2305/IUCN.UK.2006.RLTS.T245A1.
Prystay, T.S., Eliason, E.J., Lawrence, M.J., Dick, M., Brownscombe, J.W., Patterson, D.A. et al. 2017. The influence of water temperature on sockeye salmon heart rate recovery following simulated fisheries interactions. Conserv. Physiol. doi:10.1093/conphys/cox050
Qu, H., Ajuwon, K.M. 2018. Cytosolic phosphoenolpyruvate carboxykinase is a response gene involved in porcine adaptation to heat stress. J. Anim. Sci. 96, 1724–1735. doi:10.1093/infdis/jiy452/5056504
Quattro, J.M., Greig, T.W., Coykendall, D.K., Bowen, B.W., Baldwin, J.D. 2002. Genetic issues in aquatic species management: The shortnose sturgeon (Acipenser brevirostrum) in the southeastern United States. Conserv. Genet. doi:10.1023/A:1015275602748
Que, Y., Xu, D., Shao, K., Xu, N., Li, W., Zhu, B. 2015. Identification and characterization of seventeen novel microsatellite markers for Dabry's sturgeon (Acipenser dabryanus). J. Genet. 93, 62–65. doi:10.1007/s12041-014-0363-2
Quinn, N.L., Mcgowan, C.R., Cooper, G.A., Koop, B.F., Davidson, S. 2011. Identification of genes associated with heat tolerance in Arctic charr exposed to acute thermal stress. Physiol Genomics 43, 685–696. doi:10.1152/physiolgenomics.00008.2011
Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2009. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. doi:10.1093/bioinformatics/btp616
Rochard, E., Castelnaud, G., Lepage, M. 1990. Sturgeons (Pisces: Acipenseridae); threats and prospects. J. Fish Biol. doi:10.1111/j.1095-8649.1990.tb05028.x
Romanenko, S.A., Biltueva, L.S., Serdyukova, N.A., Kulemzina, A.I., Beklemisheva, V.R., Gladkikh, O.L. et al. 2015. Segmental paleotetraploidy revealed in sterlet (Acipenser ruthenus) genome by chromosome painting. Mol. Cytogenet. 8, 90. doi:10.1186/s13039-015-0194-8
Salvanes, A.G.V., Moberg, O., Ebbesson, L.O.E., Nilsen, T.O., Jensen, K.H., Braithwaite, V.A., 2013. Environmental enrichment promotes neural plasticity and cognitive ability in fish. Proc. R. Soc. B Biol. Sci. doi:10.1098/rspb.2013.1331
Sanger, F., Nicklen, S., Coulson, A., 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U. S. A. 74, 5463–5467. doi:10.1093/annonc/mdp039
Schmidt, M.H.W., Vogel, A., Denton, A., Istace, B., Wormit, A., van de Geest, H. et al. 2017. Reconstructing The Gigabase Plant Genome Of Solanum pennellii Using Nanopore Sequencing. bioRxiv 1–23. doi:10.1101/129148
Scholz, S., Klüver, N., 2009. Effects of endocrine disrupters on sexual, gonadal development in fish. Sex. Dev. 3, 136–151. doi:10.1159/000223078
Shi, Y., He, M., 2014. Differential gene expression identified by RNA-Seq and qPCR in two sizes of pearl oyster (Pinctada fucata). Gene. doi:10.1016/j.gene.2014.01.031
Shin, S.C., Kim, H., Lee, J.H., Kim, H.W., Park, J., Choi, B.S. et al. 2019. Nanopore sequencing reads improve assembly and gene annotation of the Parochlus steinenii genome. Sci. Rep. doi:10.1038/s41598-019-41549-8
Sicuro, B., 2019. The future of caviar production on the light of social changes: a new dawn for caviar? Rev. Aquac. doi:10.1111/raq.12235
Smederevac-Lalic, M., Pešic, R., Cvejic, S., Simonovic, P. 2012. Socio-economic features of commercial fishery in the bordering upper Danube River area of Serbia. Environ. Monit. Assess. doi:10.1007/s10661-011-2140-5
Smith, T.I.J., Collins, M.C., Post, W.C., McCord, J.W. 2002. Stock enhancement of shortnose sturgeon: A case study, in: Biology, Management, and Protection of North American Sturgeons.

Solås, M.R., Skoglund, H., Salvanes, A.G.V., 2019. Can structural enrichment reduce predation mortality and increase recaptures of hatchery-reared Atlantic salmon Salmo salar L. fry released into the wild? J. Fish Biol. doi:10.1111/jfb.14004
Soshnikova, N., Dewaele, R., Janvier, P., Krumlauf, R., Duboule, D., 2013. Duplications of hox gene clusters and the emergence of vertebrates. Dev. Biol. doi:10.1016/j.ydbio.2013.03.004
Stingo, V., Rocco, L., 2001. Selachian cytogenetics: A review. Genetica 111, 329–347. doi:10.1023/A:1013747215866
Sulak, K.J., Randall, M.T., Clugston, J.P., 2014. Survival of hatchery Gulf sturgeon (Acipenser oxyrinchus desotoi Mitchill, 1815) in the Suwannee River, Florida: A 19-year evaluation. J. Appl. Ichthyol. 30, 1428–1440. doi:10.1111/jai.12607
Swartz, B., 2012. A marine stem-tetrapod from the devonian of western north america. PLoS One 7, e33683. doi:10.1371/journal.pone.0033683
Symonová, R., Havelka, M., Amemiya, C.T., Howell, W.M., Ko, T., Flaj, M. et al. 2017. Molecular cytogenetic differentiation of paralogs of Hox paralogs in duplicated and re-diploidized genome of the North American paddlefish (Polyodon spathula). BMC Genomics 18, 19. doi:10.1186/s12863-017-0484-8
Treangen, T.J., Salzberg, S.L., 2012. Repetitive DNA and next-generation sequencing: Computational challenges and solutions. Nat. Rev. Genet. doi:10.1038/nrg3117
Vaser, R., Sovic, I., Nagarajan, N., Sikic, M., 2017. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27, 737–746. doi:10.1101/068122
Venkatesan, B.M., Bashir, R., 2011. Nanopore sensors for nucleic acid analysis. Nat. Nanotechnol. doi:10.1038/nnano.2011.129
Venkatesh, B., Lee, A.P., Ravi, V., Maurya, A.K., Lian, M.M., Swann, J.B. et al. 2014. Elephant shark genome provides unique insights into gnathostome evolution. Nature 505, 174–179. doi:10.1038/nature12826
Verma, M., Kulshrestha, S., Puri, A., 2017. Genome sequencing. In: Keith J. (eds) Bioinformatics. Methods in Molecular Biology, Tissue Phenomics: Profiling Cancer Patients for Treatment Decisions. doi:10.1201/9781351134279
Vinogradov, A.E., 1998. Genome size and GC-percent in vertebrates as determined by flow cytometry: The triangular relationship. Cytometry 31, 100–109. doi:10.1002/(SICI)1097-0320(19980201)31:2<100::AID-CYTO5>3.0.CO;2-Q
Vurture, G.W., Sedlazeck, F.J., Nattestad, M., Underwood, C.J., Fang, H., Gurtowski, J. et al. 2017. GenomeScope: Fast reference-free genome profiling from short reads. Bioinformatics 33, 2202–2204. doi:10.1093/bioinformatics/xxxxx
Wagner, G.P., Amemiya, C., Ruddle, F., 2003. Hox cluster duplications and the opportunity for evolutionary novelties. Proc. Natl. Acad. Sci. doi:10.1073/pnas.2536656100
Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S. et al. 2014. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9, e112963. doi:10.1371/journal.pone.0112963
Wang, R., Liu, J., Lin, C., Huang, B., Pan, X., Wang, Y. 2011. Distribution and bioaccumulation of steroidal and phenolic endocrine disrupting chemicals in wild fish species from Dianchi Lake, China. Environ. Pollut. 159, 2815–2822. doi:10.1016/j.envpol.2011.05.013
Warren, R.L., Coombe, L., Mohamadi, H., Zhang, J., Jaquish, N.I., Jones, S.J.M. et al. 2019. ntEdit: scalable genome sequence polishing. Bioinformatics, btz400, https://doi.org/10.1093/bioinformatics/btz400
Waterhouse, R.M., Seppey, M., Simao, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G. et al. 2018. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol. Biol. Evol. 35, 543–548. doi:10.1093/molbev/msx319

Werner, I., Linares-Casenave, J., Van Eenennaam, J.P., Doroshov, S.I., 2007. The effect of temperature stress on development and heat-shock protein expression in larval green sturgeon (Acipenser mirostris). Environ. Biol. Fishes 79, 191–200. doi:10.1007/s10641-006-9070-z

Williot, P., Rochard, E., Castelnaud, G., Rouault, T., Brun, R., Lepage, M. et al. 2005. Biological characteristics of European Atlantic sturgeon, Acipenser sturio, as the basis for a restoration program in France. Sturgeon Biodivers. Conserv. 359–370. doi:10.1007/0-306-46854-9_24

Wu, A.R., Neff, N.F., Kalisky, T., Dalerba, P., Treutlein, B., Rothenberg, M.E. et al. 2014. Quantitative assessment of single-cell RNA-sequencing methods. Nat. Methods. doi:10.1038/nmeth.2694

Yebra-Pimentel, E.S., Gebert, M., Jansen, H.J., Jong-Raadse S.A., Dirks, R.P.H. 2019. Deep transcriptome analysis of the heat shock response in an Atlantic sturgeon (*Acipenser oxyrinchus*) cell line. *Fish Shellfish Immunol.* 88, 508-517.

Zhang, Y., Loughery, J.R., Martyniuk, C.J., Kieffer, J.D., 2016. Physiological and molecular responses of juvenile shortnose sturgeon (Acipenser brevirostrum) to thermal stress. Comp. Biochem. Physiol. A. Mol. Integr. Physiol. 203, 314–321. doi:10.1016/j.cbpa.2016.10.009

Zhou, X.Y., Li, M.Y., Abbas, K., Gao, Z.X., Wang, W.M., 2008. Comparison of ploidy level screening methods in Chinese dojo loach (Misgurnus anguillicaudatus). J. Appl. Ichthyol. doi:10.1111/j.1439-0426.2008.01116.x

Ziegeweid, J.R., Jennings, C.A., Peterson, D.L., 2008. Thermal maxima for juvenile shortnose sturgeon acclimated to different temperatures. Environ. Biol. Fishes 82, 299–307. doi:10.1007/s10641-007-9292-8

Ziemann, D.A., 2001. The potential for the restoration of marine ornamental fish populations through hatchery releases. Aquarium Sci. Conserv. 3, 107–117. doi:10.1023/A:1011367615465

Appendix: Papers I-III

I

Deep transcriptome analysis of the heat shock response in an embryo-derived Atlantic sturgeon (*Acipenser oxyrinchus*) cell line

Elena Santidrián Yebra-Pimentel[a,b*], Marina Gebert[c], Hans J. Jansen[a], Susanne A. Jong-Raadsen[a] and Ron P.H. Dirks[a]

a - ZF-screens B.V., 2333CH Leiden, The Netherlands. b - Department of Basic Sciences and Aquatic Medicine, Norwegian University of Life Sciences, 0454 Oslo, Norway. c - Working Group Aquatic Cell Technology and Aquaculture, Fraunhofer Research Institution for Marine Biotechnology and Cell Technology, 23562 Lübeck, Germany.

* - Corresponding author at: ZF-screens B.V., 2333CH Leiden, The Netherlands. E-mail address: elena@zfscreens.com

Abstract

Despite efforts to restore Atlantic sturgeon in European rivers, aquaculture techniques result in animals with high post-release mortality due to, among other reasons, their low tolerance to increasing water temperature. Marker genes to monitor heat stress are needed in order to identify heat-resistant fish. Therefore, an Atlantic sturgeon cell line was exposed to different heat shock protocols (30ºC and 35ºC) and differences in gene expression were investigated. In total 3,020 contigs (~1.5%) were differentially expressed. As the core of the upregulated contigs corresponded to heat shock proteins (HSP), the heat shock factor (HSF) and the HSP gene families were annotated in Atlantic sturgeon and mapped via Illumina RNA sequencing to identify heat-inducible family members. Up to 6 *hsf* and 76 *hsp* genes were identified in the Atlantic sturgeon transcriptome resources, 16 of which were significantly responsive to the applied heat shock. The previously studied *hspa1* (*hsp70*) gene was only significantly upregulated at the highest heat shock (35ºC), while a set of 5 genes (*hspc1, hsph3a*, *hspb1b*, *hspb11a,* and *hspb11b)* was upregulated at all conditions. Although the *hspc1 (hsp90a)* gene was previously used as heat shock-marker in sturgeons, we found that *hspb11a* is the most heat-inducible gene, with up to 3,296-fold higher expression in the treated cells, constituting the candidate gene markers for *in vivo* trials.

Keywords: *Acipenser oxyrinchus*, RNAseq, Temperature, Heat shock proteins (HSP), Transcriptome sequencing.

Introduction

Sturgeons are distributed in the northern hemisphere and are mostly anadromous, performing upriver migrations to spawn. They have been aquacultured for the last three decades due to the commercial value of their caviar [1], however wild populations are in severe decline [2,3] and currently the Atlantic sturgeon (*Acipenser oxyrinchus*, Mitchill, 1815) is extirpated from Europe [4].

In 1997, Germany and Poland started a project aiming to restore a self-sustaining Atlantic sturgeon population, derived from a Canadian broodstock, in the Baltic Sea. Rearing techniques have been improved in order to build an *ex-situ* broodstock locally. In addition, the Oder and Vistula river drainages, where more than 120,000 individuals have been released since 2005 [5], have been evaluated in search of possible spawning grounds [6]. The success of a restoration program depends both on politics (e.g., habitat preservation policies and regional and international cooperation) [7, 8] and on biology, (e.g., development of cultivation techniques adapted for restocking purposes). Aquacultured sturgeons are raised in stimulus-deprived tanks at constant temperature and photoperiod [9]; however, their distribution, abundance, and activity are determined by temperature, photoperiod and salinity fluctuations along the migration route [10,11]. An increase in water temperature, both locally due to the discharge of cooling waters or globally due to climate change, can affect swimming performance, predator avoidance, foraging behavior and shift the geographic distribution of species, especially of ectothermic animals such as fish, and result in local extinctions [12–14]. Thus, thermal adaptation is crucial for post-release survival.

As a result of physiological or chemical stress, proteins misfold and accumulate in the cytoplasm, activating a highly conserved and transient heat shock response (HSR), essential for proteostasis maintenance. HSR results from the rapid activation of the heat shock factor (HSF) by trimerization, hyperphosphorylation and translocation to the nucleus, were its DNA-binding domain (DBD) binds to the heat shock elements (HSE) located at the promoters of heat shock protein (*hsp*) genes, inducing their transcription [15]. Previous studies in *Saccharomyces cerevisiae* showed that heat-responsive gene expression attenuates within 20-40 minutes after induction [16]. Most vertebrates have multiple *hsf* genes: mammals have 4 [17], while the Spotted gar (*Lepisosteus oculatus*), sturgeon's closest living relative whose genome is sequenced, has 6 (*hsf1*, *hsf2*, *hsf3*, *hsf4*, *hsf5,* and *hsfy*).

HSPs, one of the largest and most conserved protein families, were first discovered in *Drosophila* [18] and have since then been described in all living organisms, from yeast [19] to mammals [20,21] and fish [22-24]. Its nomenclature refers to their molecular weights, however, since the human genome annotation the names have become confusing, with up to 10 different names for the same gene product. In 2009 a new nomenclature was proposed [25] based on the gene symbols that have been assigned to the HUGO Gene nomenclature committee (HGNC). This nomenclature was also used in this study, and classifies HSPs into 9 subfamilies: HspA (former Hsp70), HspH (former Hsp110), HspB (small Hsp), HspC (Hsp90), HspD (Hsp60), HspE (Hsp10) and the J domain-containing subfamilies (Hsp40) DNAJA, DNAJB and DNAJC.

The HSP family consists on both constitutively expressed and inducible members, some of which are responsive to heat and act by binding to the exposed hydrophobic amino acid residues of the misfolded proteins preventing undesired molecular interactions. The capacity of HSPs to bind misfolded proteins is regulated through allosteric mechanisms via ATP binding and hydrolysis, with the exception of the HspB subfamily which is ATP-independent.

RNA sequencing (RNAseq) is a high-throughput tool to quantify transcriptomic changes and, unlike qPCR, is unbiased, allowing gene discovery and quantification with good correlation with qPCR when both procedures are performed well [26,27]. RNAseq has previously been used in sturgeons to study sex-related genes, developmental genes or response to infection [28-33], but never to study the heat shock response. Many studies used RNAseq to address this in teleosts [34-36], while for sturgeons only qPCR or protein analysis approaches were used [37-42]. Still, the focus has only been on *hspa1 (hsp70)* and hspc1 (*hps90)*, as the HSP family is not fully annotated in sturgeons, which are also lacking a reference genome.

Recently, the AOXlar7y cell line (***A****cipenser* ***o****xyrinchus* **lar**vae n°**7**, tr**y**psin-digestion) was established by whole larvae trypsin digestion [43], providing the possibility of reducing the use of *in vivo* sturgeon experiments. Still, the presence of stem cells in the culture was not confirmed so, characterization of the cell type needs to be assessed.

The objective of this study was to make an inventory of all the HSPs present in the Atlantic sturgeon transcriptome resources and identify the most heat shock-inducible family members. To this end, we annotated the complete HSP family using transcriptomic resources from Atlantic sturgeon cell line, multiple Atlantic sturgeon organs and a publicly available dataset [31] and subsequently evaluated their response to heat. Although *in vivo* validation during temperature-challenge trials is needed, the *in vitro* experiment provides a general perspective of the HSR and candidate gene markers for the selection of thermotolerant individuals with better restoration fitness.

Materials and methods

Cell line samples

The AOXlar7y sturgeon cell line [43] was used for the heat shock experiments. Cells from passage 22 (P22) were thawed and seeded in a 25 cm^2 cell culture flask (Corning Life Sciences, Tewksbury, MA, USA) at 25°C, the optimal growing temperature (doubling time 110h), in Leibovitz-15 medium supplemented with 15% FCS (fetal calf serum), 100U/mL penicillin and 0.1mg/mL streptomycin.

Prior to the final experiment, the cells were propagated by trypsinization, homogenization and subculturing at a 1:3 ratio every 4 days, up to P29. To determine the heat tolerance of the AOXlar7y sturgeon cell line, the cells were exposed to 28°C, 30°C, 33°C or 35°C for 1, 2, 3 or 4 hours and survival was inspected under the microscope after 4, 8 and 24 hours of recovery. Based on this pilot experiment, the final experiment was performed as depicted in Figure 1: 24 individual flasks were seeded at 25°C, and after 24 hours the medium was renewed at either 25°C (n=6), 30°C (n=9) or 35°C (n=9). The flasks were then transferred to different incubators set at 25°, 30° and 35°C respectively and kept there for 1 hour. After

the heat shock all the flasks were transferred back to 25°C to recover. After 4, 8 and 24 hours of recovery, cells were lysed in QIAzol lysis reagent (Qiagen GmbH, Hilden, Germany) and stored at - 80° until further processing.

Atlantic sturgeon samples

An aquaculture-reared immature female (7,700g, 98cm) was provided by Fischzucht Rhönforelle GmbH in Gersfeld (Germany) on July 3rd, 2015. The specimen was euthanized, and tissue samples were taken in agreement with standardized fish processing methods at a licensed and registered processing unit (Fischzucht Rhönforelle GmbH). A total of 21 samples were taken through dissection, from caudal to rostral: caudal fin, dorsal fin, pelvic fin, skin, muscle, intestine (posterior), kidney, spleen, swim bladder, gonad (2 samples), intestine (anterior), stomach, liver, gallbladder, heart, gill, brain (anterior), brain (posterior), barbel and eye. Samples were preserved in RNAlater (Qiagen GmbH, Hilden, Germany) and subsequently stored at -80°C until further processing.

RNA extraction, Illumina library preparation, and RNA sequencing

RNA was extracted from the 21 juvenile Atlantic sturgeon tissue samples and from the 24 AOXlar7y cell line samples using the Qiagen miRNeasy Mini kit according to the manufacturer's instructions (Qiagen GmbH, Hilden, Germany). RNA concentration and integrity were analyzed with a Bioanalyzer 2100 total RNA Nano series II chip (Agilent, Santa Clara, USA). RNA libraries were prepared from 500 ng total RNA, using the Illumina TruSeq Stranded mRNA Sample Preparation Kit according to the manufacturer's instructions (Illumina Inc., San Diego, USA) and the resulting libraries were evaluated with a Bioanalyzer 2100 DNA 1000 series II chip (Agilent, Santa Clara, USA).
All the libraries were sequenced using an Illumina HiSeq2500 instrument. Tissue libraries were sequenced as paired-end 2x151 nucleotides (nt) reads up to a minimum of ~ 20 million reads for each tissue sample, while AOXlar7y libraries were sequenced as 1x51nt single-reads up to a minimum of ~10 million reads for each experimental condition.

De novo transcriptome assemblies and annotation

Three reference transcriptomes were produced in the present study. Firstly, all combined juvenile sturgeon organ reads were *de novo* assembled into cDNA contigs using the De Bruijn graph-based *de novo* assembler implemented in the CLC Genomics Workbench version 4.4.1 (CLC bio, Aarhus, Denmark), resulting in an Organ assembly. Secondly, the same software was used to produce a Cell assembly with the AOXlar7y sequence reads only. Finally, all sequence reads were combined into a Cell-Organ-Embryo (COE) transcriptome, including the reads from the AOXlar7y, the juvenile sturgeon organs and a set of embryonic reads previously published by others [31].
The quality of the *de novo* assemblies was assessed by the assembly contiguity (contig N50) and the percentage of AOXlar7y mapped reads. In order to link the best assembly to zebrafish Ensembl protein identifiers, BLASTX 2.2.31+ [44] similarity searches were conducted locally against the UniProt zebrafish using an E-value cut off of 1E-5.

Gene expression and gene ontology (GO) analysis

For the gene expression analysis, Bowtie2 (version 2.2.5) [45] was used to align Illumina reads from the 24 experimental samples against the 3 *de novo* assembled reference transcriptomes (Cell, Organ and COE contigs) and against the embryonic transcriptome available at NCBI [31]. The assembly with the highest mapping percentage was used as a reference for downstream analysis. The resulting files were filtered using SAMtools (version 1.2.) [46] to exclude secondary aligned reads. Then, the aligned reads were counted from the SAM alignment files using Python package HTSeq (version 0.5.3p9) [47], and the resulting TSV files were used for identification of differentially expressed contigs (DECs) using the Bioconductor package DESeq (version 1.30.0) [48] in R software (version 3.3.0). The top 100 expressed contigs in the unstressed cells were investigated to characterize the AOXlar7y cell line. Each experimental condition (30°C and 35°C) and timepoint (4, 8 and 24 hours) was compared to the control group, resulting in 6 DESeq pairwise comparisons. P-values were adjusted for multiple testing with the Benjamini-Hochberg procedure, which controls the false discovery rate (FDR). Contigs with an adjusted p-value (padj) < 0.05 were considered differentially expressed at each pairwise comparison between different treatments and timepoints. The set of zebrafish protein identifiers linked to these contigs was investigated in order to elucidate the broad transcriptomic changes. GO enrichment analysis was performed using the Fisher's Exact test with Bonferroni correction implemented by the PANTHER Overrepresentation test (released on 2017-12-05) using the GO Ontology database (released on 2018-07-03) [49]. As recommended by the Gene Ontology Consortium [50, 51], a custom reference list containing all the top zebrafish Ensembl gene identifiers linked to the expressed COE contigs was used for the analysis. GO terms with a p-value < 0.05 were considered overrepresented in each dataset.

Annotation and expression of Hsp and Hsf genes

Spotted gar, zebrafish and human HSP and HSF protein sequences [25] were retrieved from NCBI and used as queries to find the corresponding sturgeon orthologues. Proteins were blasted against the translated COE transcriptome assembly using the CLC Main Workbench (version 7.7.3.). When full open reading frames (ORF) were not found in the COE assembly, the Organ, Cell or embryo assembly [31] were used. The human HSP protein nomenclature according to Kampinga et al. [25] was used. After the annotation of all Atlantic sturgeon *hsp* and *hsf* genes, the AOXlar7y Illumina reads were mapped to the annotated genes and those with more than 5 mismatches were filtered out in order to improve stringency. DESeq analysis was performed, following the aforementioned pipeline, in order to find candidate *hsp* markers for *in vivo* trials. Significantly upregulated transcripts (padj>0.05) with at least 3-fold change (FC≥3) between the control and at least 5 experimental conditions were considered good markers in this experiment.

The AOXlar7y and organ sequencing reads were deposited at the NCBI Short Read Archive (SRA) database under the accession numbers SRP161542 and SRP161601 respectively. The Cell, Organ, and Cell-Organ-Embryo (COE) transcriptome shotgun assembly projects have

been deposited at DDBJ/EMBL/GenBank under the accession numbers GGWJ00000000, GGZX00000000 and GGZT00000000 respectively. The versions described in this paper are the first versions: GGWJ01000000, GGZX01000000, and GGZT01000000, respectively. The Atlantic sturgeon *hsp* and *hsf* ORFs were submitted to BankIT under the accession numbers MH777912-MH777987 and MH917287-MH917292 respectively.

Results

Heat shock treatment and RNA sequencing

The pilot experiment has shown that cells survived after being exposed to 28°C, 30C° or 33°C for up to 4 hours or to 35°C for 1 hour; however, acute mortality was found in cells exposed to 35°C for 2 hours or longer (data not shown). Based on this pilot, the final experiment was performed applying a mild or a severe heat shock at either 30ºC or 35ºC for 1 hour, and cells were harvested at 4, 8 and 24 hours after heat shock (Figure 1). Subsequently, high-quality RNA (average RIN value of 9.1) was isolated from all cell samples and from multiple organs of an aquacultured juvenile sturgeon specimen, and Illumina libraries were prepared and sequenced, resulting in altogether more than 1 billion reads (Table S1 in Supplementary data).

Heat shock-responsive genes in Atlantic sturgeon cell line AOXlar7y

To select the best reference for alignment of the sequencing reads, 3 Atlantic sturgeon *de novo* transcriptome assemblies were performed and evaluated. The set of reads derived from the AOXlar7y cell line was assembled to a 34.80 Mb Cell assembly containing 53,624 contigs with an N50 of 1.08 Kb. In parallel, reads from the Atlantic sturgeon organs were assembled to a 342.28 Mb Organ assembly containing 641,485 contigs with an N50 of 0.60 Kb. Finally, all sequencing reads were combined with a previously published set of ~ 380 million reads derived from Atlantic sturgeon embryos (SRA Accession number SRP069853) [31] and used in a *de novo* assembly that resulted in a 254 Mb Cell-Organ-Embryo (COE) assembly containing 203,131 contigs with an N50 of 1.87 Kb (Table 1).

Reads from the 24 AOXlar7y samples were aligned against the three assembled transcriptomes and the COE reference, which gave the highest overall mapping percentage (89.6%), was used for the downstream analysis (Table 1). In total 86,021 contigs (~ 42.3%) could be linked to zebrafish proteins (E-value <1E-5), corresponding to 23,436 unique zebrafish proteins and 17,319 genes (Table S2 in Supplementary data). The read counts and contig lengths were used to calculate the corresponding RPKM values (Table S3 in Supplementary data). A total of 168,739 contigs (83.07%) showed expression (RPKM value > 0) in at least one of the cell samples, and 27,425 (13.50%) showed expression in all of them.

Since it was hitherto unknown what tissue type was represented by the embryonic AOXlar7y cell line, the zebrafish proteins linked to the 100 most highly expressed contigs (highest RPKM mean) were examined. In total, 34 contigs corresponded to ribosomal proteins, 5 to keratins, 5 to actin genes, 5 to tubulin and other genes were also found. Overall, keratin 4 was the most expressed contig (Table S4 in Supplementary data).

DESeq analysis showed that 3,020 unique contigs (~ 1.5%) were differentially expressed in at least 1 of the treatments compared with the untreated cells, of which 2,302 were upregulated, 714 downregulated and 4 were up or downregulated depending on the experimental condition (Table S5 in Supplementary data).

At 4 and 8 hours after the 30°C heat shock, 79 and 91 contigs were upregulated and 23 and 64 were downregulated respectively, while after 24 h only 12 were downregulated and none upregulated. After the 35°C heat shock, respectively 1,338, 1,081 and 1,143 contigs were upregulated and 411, 123 and 198 were downregulated (Figure 2).

Venn diagrams were used to determine overlaps between the sets of heat shock-responsive contigs. After the 30°C heat shock, 140 unique contigs (0.07%) were upregulated, 30 of which at both 4 and 8 hours after heat shock (Figure 3a), while only 87 unique contigs (0.04%) were downregulated (Figure 3b). After the 35°C heat shock 2,257 unique contigs (1.11%) were upregulated, 398 of which were upregulated at all 3 timepoints (Figure 3c), and 664 unique contigs (0.33%) were downregulated, of which only were downregulated at all timepoints (Figure 3d).

Considering that one of our goals was to provide a list of consistently upregulated genes that need to be validated as markers *in vivo*, we examined the core set of 27 upregulated contigs (Figure 3e). Up to 18 corresponded to 5 heat shock proteins (*hspb1, hspb11a, hsph3a, hsph2*, and *hspc1*), while 5 corresponded to 4 different proteins: clusterin (*clu*), growth factor receptor bound protein 2b (*grb2b*), atrial natriuretic peptide receptor 2-like (*npr2*) and coiled-coil domain-containing protein 17 (*ccdc17*) (Tables 2 and S7 in Supplementary data). The remaining contigs did not output any blast hit or were uncharacterized proteins.

To link the differentially expressed contigs (DECs) to GO terms, the 3,020 unique contigs were blasted (BLASTx) against zebrafish proteins (E-value <1E-2), which resulted in 1,383 hits (45.8%) corresponding to 1,068 unique proteins and 1,017 zebrafish genes (Table S6 in Supplementary data). GO enrichment analysis for biological processes (BP) and molecular functions (MF) was performed (Table S8 in Supplementary data). With respect to upregulated genes after exposure to 30°C, cellular processes involved in the response to stimulus (e.g. 'cellular response to heat', 'cellular response to unfolded protein', 'response to temperature stimulus', 'response to unfolded protein'), protein folding (e.g. ' chaperone-mediated protein folding', 'protein refolding') and protein binding (e.g. 'heat shock protein binding', 'protein folding', 'misfolded protein binding') were overrepresented (Figure 4), however, no overrepresented terms were found after 8 or 24 hours. After the 35ºC heat shock overrepresented terms were found at all the timepoints: while GO terms related with cellular response to stimulus and stress (e.g. 'cell chemotaxis', 'leukocyte chemotaxis', 'inflammatory response') and cell migration (e.g. 'granulocyte migration', leukocyte migration', 'neutrophil migration') are overrepresented after 4 and 8 hours (Figure 5a and 5b), after 24 hours (Figure 5c) many more GO terms are overrepresented, being the 'response to stimulus' the predominant term (122 genes).

With respect to downregulated genes, overrepresentation of 'cell cycle' and 'mitotic cell cycle' GO terms was found only at 4 hours after 35ºC heat shock.

Annotation and expression of Hsp and Hsf family members in Atlantic sturgeon

Protein sequences of all known human, zebrafish and spotted gar HSP and HSF family members were retrieved from NCBI and used as queries to search for the corresponding Atlantic sturgeon orthologues in the COE, Organ, Cell or Embryo transcriptomes (Table S9 in Supplementary data).

Spotted gar HSF proteins were used to identify 6 *hsf* genes in sturgeon, which lacked the *hsf3* gene and had 2 *hsfy* genes (*hsfy1* and *hsfy2*). None of the *hsf* genes was differentially expressed at any time point after the mild and severe heat shock.

The former Hsp70 family is divided into 2 sub-families: HspA and HspA-related HspH (Hsp110). The HSPA family contains 13, 8 and 7 members in humans, zebrafish and spotted gar, respectively. Nine members were found in the Atlantic sturgeon transcriptome: *hspa1, hspa5, hspa8, hspa9, hspa12a, hspa12b, hspa13, hspa14a, hspa14b*. The HSPH family contains 4 members in both human and spotted gar, plus 1 duplicated gene in zebrafish. The *hsph1* member could not be retrieved in any of the available sturgeon assemblies and the *hsph3* member was found duplicated.

Humans, spotted gar and zebrafish contain the *hspb1*, *hspb2, hspb3, hspb4* (αA-crystallin), *hspb5* (αB-crystallin)*, hspb7, hspb8* and *hspb9* members of the small HSP family, whereas *hspb6* is missing in spotted gar, *hspb10* is only present in humans, and *hspb11*, *hspb12* and *hpsb15* are only present in zebrafish. With the exception of *hspb3*, orthologues of all spotted gar *hspb* genes could be found in the sturgeon transcriptome, including 2 *hspb1* genes (*hspb1a* and *hpsb1b*). The *hspb11* member, present only in zebrafish, could also be retrieved duplicated in the sturgeon transcriptome (*hspb11a* and *hspb11b*).

The HSPC family (former *hsp90*) has 5 members in humans and 4 in zebrafish and spotted gar, which lacks the *hspc2* member. The same fish members were found in the sturgeon transcriptome, with an additional *hspc3* duplication (*hspc3a* and *hspc3b*). The HSPD and HSPE families each contain only 1 single gene in humans, zebrafish and spotted gar, and orthologues of both of them could be found in sturgeon.

The new nomenclature for the former HSP40 family divides it into DNAJA, DNAJB and DNAJC families. We have found 4, 10 and 32 genes for each family, respectively.

Upon mapping the AOXlar7y Illumina reads against the annotated sturgeon HSP genes (Table S10 in Supplementary data) and differential expression analysis using DESeq, we found that 16 out of 76 HSP genes (21.%) were differentially expressed between the control and some of the experimental conditions. No HSPs transcripts were found differentially expressed between the control and 24 hours after the 30ºC heat shock. Within the HspA family, only *hspa1* was differentially expressed after 35ºC heat shock (Figure 6a), being up to ~1000-fold upregulated 4 hours after the 35ºC treatment.

Upregulation of 3 out of 4 HspH family members was observed (up to ~9-fold): *hsph2* and *hsph3a* were significantly upregulated at all the conditions, however, the *hsph3b* paralog was only slightly upregulated at after the 30° treatment (Figure 6a).

Five out of ten HspB family members and the single HSPC family member *hspc1* were upregulated after heat shock: while *hspb8* was only significantly upregulated at 4 hours after the 35°C heat shock and *hspb5* only at 35ºC, *hspb1b, hspb11a, hspb11b* and *hspc1* were consistently upregulated in all the treatments (excluding 24 hours after the 30°C treatment), with the *hspb11a* having the higher expression (as high as ~3296-fold compared to the

untreated cells) (Figure 6b). The HspE1 and HspD1 were not differentially expressed at any condition.

Within the DNAJA family (Figure 6c), only the *dnaja4* member was upregulated. Four DNAJB members were upregulated: *dnajb1* and *dnajb5* were only upregulated after the 35°C heat shock, *dnajb4* was also upregulated at 4 hours after the 30°C heat shock, and *dnajb2* was only upregulated after the 30°C heat shock.

Although the DNAJC family is the most extensive DNAJ subfamily, it contained only 1 heat shock inducible gene under our conditions, *dnajc3*, which was upregulated only at 24 hours after the 35ºC heat shock.

Discussion

The AOXlar7y cell line is easy to maintain and propagate, providing an excellent tool for examining the effects of different stressors. Healthy AOXlar7y cells have cubic to fibroblast-like morphology, and the RNAseq analysis in this study shows that keratin 4 is the most abundantly expressed mRNA in this cell line, suggesting that they are epithelial cells [52]. Although both the cell line and the donor embryo have been reported to be more tolerant to cold than heat [43, 53] the optimal temperature for cells is higher than for the entire animal (25°C and 18°C-23°C, respectively). After establishing the temperature tolerance limits, cells were exposed to a mild (30°C) and severe (35°C) non-lethal heat shock for 1 hour and, after 4, 8 and 24 hours of recovery, RNA was isolated and sequenced. By RNAseq we could get a general perspective on the heat-inducible genes, which may not truly represent the metabolic state of the cells *in vivo* and therefore needs validation, but still provides the sequence of all *hsp* genes and a general perspective of which genes may be more responsive to heat.

The COE reference transcriptome had a much higher alignment rate of cellular reads (89.6%) than the transcriptomes that were assembled from the individual data sets, and was therefore selected for identification of DECs. This reference provides a considerable number of sturgeon protein sequences that contribute to sturgeon research. The set of 3,020 DECs between the control and at least one of the treatments was blasted against zebrafish proteins, the most related species for which GO analysis is available. A total of 1,383 DECs could be assigned to zebrafish proteins, while the remaining contigs were noncoding, sturgeon-specific, missing in zebrafish or lacked sufficient homology with their zebrafish orthologues.

The DESeq results showed a very intense response at 35°C, initially involving energy-consuming gene upregulation and later also gene downregulation, which is an energy-saving mechanism to direct energy towards the repair of damaged molecules [36]. After a mild 30°C heat shock there were very few transcriptional changes at early timepoints, and even zero after 24 hours of recovery. The narrow temperature tolerance range between 30°C and 35°C where sturgeon cells switch from a mild to a severe heat shock response is in agreement with existing studies in green sturgeon (Acipe*nser medirostris*) [53] and Kaluga (*Huso dauricus*) [39].

The GO analysis showed a heat shock response at 35°C which increased with the recovery time. At 30°C the response was more specific (overrepresentation of less GO terms);

however, 24 hours after heat shock the cells showed the same transcriptional activity as the untreated cells, suggesting that the cells were already fully recovered after this mild heat shock (30ºC). The bulk of the 27 contigs that were upregulated at all conditions (excluding 30°C after 24 hours) corresponded with HSP genes, and included 4 other genes: *clu*, *grb2b, npr2 and ccdc17*. Clusterin is a molecular chaperone [54, 55] and, similarly to *npr2*, is involved in cell survival after apoptosis induction [56-58]. Coiled-coil domains are involved in the regulation of gene expression, but the specific function of *ccdc17* has not been studied in detail [59]. As both the DESeq and GO results point to the HSP genes as central players in the heat shock response, we have annotated the entire family and analyzed their response to heat.

In total 76 Hsp and 6 Hsf genes could be retrieved in the Atlantic sturgeon transcriptomes, including 4 partial sequences and 78 full ORFs: 33 (40.24%) in the COE assembly and 45 (54.87%) in either the embryo, cell or organ transcriptomes.

In our experiment, only 16 out of 76 Hsp genes were heat-inducible, for which the coefficient of variation (CV) was lower than 30% in most of the cases (89.58%). HspB was found to be the most heat responsive family with less than 30% CV except for *hspb5*, which had high CV at all the timepoints after the 30ºC heat shock. The HspB family is ATP-independent and characterized by the presence of a conserved α-crystallin domain [60]. Their expression has been shown to enhance the post-stimulus survival of mammalian cells [61] and, besides the molecular chaperone activity, some members have additional cellular functions: *hspb1* (*hsp27*) and *hspb5*, induced by heat shock in zebrafish [62], are known to inhibit apoptosis [63-65], while *hspb8* has kinase activity [66]. Unfortunately, the human *hspb11* gene proposed by Kampinga [25], previously known as intraflagellar transport protein 24 (*ift25*), lacks the α-crystallin domain and its nomenclature hasn't been approved [67]. In fact *hspb11*, also known as *hsp30*, exists in all vertebrates except mammals [68], and doesn't share any amino acid sequence similarity with the human *ift25*. Since *hspb11* absent in both human and spotted gar genomes, the zebrafish protein sequence was used to retrieve the Atlantic sturgeon orthologue.

The few published studies on HSPs in *Acipenserids* were performed in species other than Atlantic sturgeon, focused only on *hspa1* (*hsp70)* and *hspc1* (*hsp90)* and didn't use RNAseq. Using qPCR, *hspa1* and *hspc1* were found expressed in both unstressed and heat-shocked Kaluga juvenile tissues; however, *hspa1* was found to be more inducible by cold than heat [39]. In contrast, *hspa1* showed higher expression after heat than cold stress in both white (*A. transmontanus*) and green (*A. medirostris*) sturgeon larvae [69]. Existing studies found a heat-dependent increase of deformities accompanied by an increase of HspA1 and HspA5 (Hsp78) and a decrease of HspD1 (Hsp60) protein levels [70]. If heat stress is not lethal, the accumulation of HSPs may lead to the tolerance of more severe and otherwise fatal stresses [71,72]. Some studies suggest that heat shock experienced by the parental fish or gametes could improve larvae thermotolerance, resulting in higher survival rates and lower incidence of deformities linked to high HspD1 and HspC1 protein levels [53]. After heat shock, Hsp levels were higher in a hot-adapted subspecies of doctor fish (*Garra rufa*) than in the non-adapted, indicating that hsp levels provide thermotolerance [73].

Overall, 5 HSPs are consistently upregulated (FC≥3) after all the treatments (excluding 24 h after the 30ºC heat shock) and are candidate markers for *in vivo* validation: *hsph3a*, *hspb1b*, *hspb11a*, *hspb11b* and *hspc1*. Interestingly, this list excludes *hspa1*, which is only upregulated after severe heat shock in our study. Although *hspc1*, previously found upregulated *in vivo,* is also included in the list included in the list, *hspb11a* has as much as ~3296-fold upregulation compared to the control cells, and is the best candidate marker for *in vivo* validation and trials.

Hspb11 is an intron-less gene, enabling fast expression without major splicing events [74]. Heat shock was shown to induce accumulation of hspb11 mRNA in Atlantic salmon [75], Chinook salmon (1250-fold) [76], red band trout (200-fold) [77], zebrafish [67], clawed frog [70], heat-tolerant Artcic charr [78] and killfish [79], with the latter showing more upregulation in heat-tolerant southern populations than the northern counterparts.

In conclusion, the AOXlar7y cell line provides the opportunity to reduce *in vivo* experiments on Atlantic sturgeon, an extirpated species in Europe. We have performed RNAseq on heat-shocked cells to get a general perspective of heat inducible genes, and identified and annotated 6 HSF and 76 HSP genes. Only 16 *hsp* transcripts were significantly upregulated after the applied treatment of which 5 were common to all treatments and timepoints, excluding 24h after the 30ºC heat shock treatment. These genes had at least a 3-fold increase in expression and one of them, hspb11a, has as much as a 3,296-fold increase. These genes are candidate markers for the selection of thermotolerant individuals and should be validated *in vivo*.

Funding

This work was supported by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Actions: Innovative Training Network "IMPRESS" [Grant Agreement No. 642893].

Acknowledgements

The authors wish to sincerely thank Peter Gross and Vojta Kaspar for the help provided during sturgeon sampling and Elena Sarropoulou for her permission and support providing the Atlantic sturgeon embryonic reads and reading through the manuscript.

Supplementary Material

Table S1. Sample list and corresponding number of sequencing reads per sample and SRA Accession number.

Table S2. List of top zebrafish BLASTx hits linked to of the COE transcriptomic contigs.

Table S3. Number of sequencing reads mapped to the COE reference assembly per AOXlar7y sample, expressed in raw and normalized values (reads per kilobase and million, RPKM).

Table S4. Top 100 expressed contigs in the untreated cell line (25ºC 4 and 24 h), indicating the RPKM mean and standard deviation (SD) and best BLASTx hits.

Table S5. DESeq results per comparison.

Table S6. List of differentially expressed contigs (DECs) in at least one condition and best zebrafish BLASTx hit.

Table S7. List of the always upregulated (n=27) COE contigs and best BLASTx hits.

Table S8. List of differentially represented GO terms per comparison.

Table S9. List of annotated *Hsp* and *Hsf* genes in Atlantic sturgeon and BankIT accession.

Table S10. Number of AOXlar7y sequencing reads mapped to the list of annotated genes and DESeq analysis results.

References

[1] Bronzi P, Rosenthal H, Gessner J. 2011. Global sturgeon aquaculture production: An overview. J Appl Ichthyol, 27. 169-175. https://doi.org.10.1111/j.1439-0426.2011.01757.x

[2] Dudgeon D, Arthington AH, Gessner MO, Kawabata Z, Knowler DJ, et al. 2006. Freshwater Biodiversity: Importance, Threats, Status and Conservation Challenges. Biol Rev Camb Philos Soc, 81(2). 163–82. https://doi.org/10.1017/S1464793105006950.

[3] Freyhof J and Brooks E. 2011. European Red List of Freshwater Fishes. Luxembourg: Publications Office of the European Union. https://doi.org/10.2779/85903.

[4] Elvira B, Leal S, Doadrio I, Almodóvar A. 2015. Current Occurrence of the Atlantic Sturgeon *Acipenser Oxyrinchus* in Northern Spain: A New Prospect for Sturgeon Conservation in Western Europe. Plos One, 10(12): e0145728. https://doi.org/10.1371/journal.pone.0145728.

[5] Kolman R, Kapusta A, Duda A, Wiszniewski G. 2011. Review of the Current Status of the Atlantic Sturgeon *Acipenser Oxyrinchus Oxyrinchus* Mitchill 1815, in Poland: Principles, Previous Experience, and Results. J Appl Ichthyol, 27(2). 186–91. https://doi.org/10.1111/j.1439-0426.2011.01680.x.

[6] Gessner J and Bartel R. 2000. Sturgeon Spawning Grounds in the Odra River Tributaries : A First Assessment. Boletin Instituto Espanol de Oceanografia, 16 (1–4). 127–37.

[7] Gessner J. 2000. Reasons for the Decline of *Acipenser Sturio* L., 1758 in Central Europe, and Attempts at Its Restoration. Boletin Instituto Espanol de Oceanografia, 16(1–4). 117–26.

[8] Kirschbaum F, Ludwig A, Hensel E, Wurtz S, Kloas W, et al. 2004. Status of the Project on Protection and Restoration of Atlantic Sturgeon in Germany: Background, Current Situation, and Perspectives. 36-53 in Species differentiation and population identification in the sturgeons Acipenser sturio L. and Acipenser oxyrinchus. J. Gessner and J. Ritterhoff editors, BfN Skripten 101: 122pp.

[9] Mohler JW. 2000. Early Culture of the American Atlantic Sturgeon *Acipenser oxyrinchus oxyrinchus* Mitchill, 1815 and Preliminary Stocking Trials. Boletin Instituto Español de Oceanografia, 16(1–4). 203–8.

[10] Beitinger TL, Bennett WA, McCauley RW. 2000. Temperature Tolerances of North American Freshwater Fishes Exposed to Dynamic Changes in Temperature. Environ Biol Fishes, 58. 237–75. https://doi.org/10.1023/A:1007676325825.

[11] Gradil AM, Wright GM, Speare DJ, Wadowska DW, Purcell S, et al. 2014. The Effects of Temperature and Body Size on Immunological Development and Responsiveness in Juvenile Shortnose Sturgeon (*Acipenser brevirostrum*). Fish Shellfish Immunol, 40(2). 545–55. https://doi.org/10.1016/j.fsi.2014.07.036.

[12] Fields PA, Graham JB, Rosenblatt RH, Somero GN. 1993. Effects of Expected Global Climate Change on Marine Faunas. Trends Ecol Evol, 8(10). 361–67. https://doi.org/10.1016/0169-5347(93)90220-J.

[13] Paragamian, VL, Beamesderfer RCP, Ireland SC. 2005. Status, population dynamics, and future prospects of the endangered Kootenai River white sturgeon population with and without hatchery intervention. Trans Am Fish Soc, 134. 518-532. https://doi.org/10.1577/T03-011.1

[14] Hofmann GE and Todgham AE. 2010. Living in the Now: Physiological Mechanisms to Tolerate a Rapidly Changing Environment. Annu Rev Physiol, 72(1). 127–45. https://doi.org/10.1146/annurev-physiol-021909-135900.

[15] Andrews GK, Harding MA, Calvet JP, Adamson ED. 1987. The Heat-Shock Response in Hela-Cells Is Accompanied By Elevated Expression of the C-Fos Protooncogene. Mol Cell Biol, 7(10). 3452–58. https://doi.org/10.1128/MCB.7.10.3452

[16] Morano KA, Grant CM, Moye-Rowley WS. 2012. The Response to heat shock and Oxidative stress in Saccharomyces cerevisiae. Genetics, 190(4). 1157-1195.

[17] Åkerfelt M, Morimoto RII, Sistonen L. 2010. Heat Shock Factors: Integrators of Cell Stress, Development and Lifespan. Nat Rev Mol Cell Biol, 11(8). 545–55. https://doi.org/10.1038/nrm2938.

[18] Ritossa F. 1962. A New Puffing Pattern Induced by Temperature Shock and DNP in drosophila. Experientia, 18(12). 571–73. https://doi.org/10.1007/BF02172188.

[19] Miller MJ, Xuong NH, Geiduschek EP. 1979. A Response of Protein Synthesis to Temperature Shift in the Yeast *Saccharomyces Cerevisiae*. Proc Natl Acad Sci U S A, 76(10). 5222–25. https://doi.org/10.1073/pnas.76.10.5222.

[20] Slater A, Cato ACB, Sillar GM, Kioussis J, Burdon RH. 1981. The Pattern of Protein Synthesis Induced by Heat Shock of HeLa Cells. Eur J Biochem, 117(2). https://doi.org/10.1111/j.1432-1033.1981.tb06343.x.

[21] Wu B, Hunt C, Morimoto R. 1985. Structure and Expression of the Human Gene Encoding Major Heat Shock Protein HSP70. Mol Cell Biol, 5(2). 330–41. https://doi.org/10.1128/MCB.5.2.330.21

[22] Krone PH, Lele Z, Sass JB. 1997. Heat Shock Genes and the Heat Shock Response in Zebrafish Embryos. Biochem Cell Biol, 75. 487–97. https://doi.org/10.1139/o97-083.

[23] Råbergh CMI, Airaksinen S, Soitamo A, Björklund HV, Johansson T, et al. 2000. Tissue-Specific Expression of Zebrafish (*Danio rerio*) Heat Shock Factor 1 MRNAs in Response to Heat Stress. J Exp Biol, 1824. 1817–24.

[24] Hori TS, Gamperl AK, Afonso LOB, Johnson SC, Hubert S, et al. 2010. Heat-Shock Responsive Genes Identified and Validated in Atlantic Cod (*Gadus Morhua*) Liver, Head Kidney and Skeletal Muscle Using Genomic Techniques. BMC Genomics, 17. 72-93. https://doi.org/10.1186/1471-2164-72.

[25] Kampinga HH, Hageman J, Vos MJ, Kubota H, Tanguay RM, et al. 2009. Guidelines for the Nomenclature of the Human Heat Shock Proteins. Cell Stress Chaperones, 14(1). 105–11. https://doi.org/10.1007/s12192-008-0068-7.

[26] Wu AR, Nef NF, Kalisky T, Dalerba P, Treutlein B, et al. 2014. Quantitative assessment of single-cell RNA-sequencing methods. Nat Methods, 11(1). 41-46.

[27] Shi Y and He M.2014. Differential gene expression identified by RNA-Seq and qPCR in two sizes of pearl oyster (*Pinctada fucata*). Gene, 538(2). 313-322. https://doi.org/10.1016/j.gene.2014.01.031

[28] Cheng Y, Xia Y, Shao C, Han L, Chen X, et al. 2016. Discovery and identification of candidate sex-related genes based on transcriptome sequencing of Russian sturgeon (*Acipenser gueldenstaedtii*) gonads. Physiol Genomics, 48(7). 464-476. doi: 10.1152/physiolgenomics.00113.2015.

[29] Yue H, Li C, Du H, Zhang S, Wei Q. 2015. Sequencing and *de novo* assembly of the gonadal transcriptome of the endangered Chinese sturgeon (*Acipenser sinensis*). PLoS One, 10(6):e0127332. https://doi.org/10.1371/journal.pone.0127332

[30] Song W, Jiang K, Zhang F, Lin Y, Ma L. 2016. RNA-sequencing of the sturgeon *Acipenser baerii* provides insights into expression dynamics of morphogenic differentiation and developmental regulatory genes in early versus late developmental stages. BMC Genomics,17. 564 doi: 10.1186/s12864-016-2839-3.

[31] Kaitetzidou E, Ludwig A, Gessner J, Sarropoulou E,. 2017. Expression Patterns of Atlantic Sturgeon (*Acipenser oxyrinchus*) During Embryonic Development. G3 (Bethesda), 7(2). 533–542. https://doi.org/10.1534/g3.116.036699.

[32] Li S, Zhang Y, Cao Y, Wang D, Liu H, et al. 2017. Transcriptome profiles of Amur sturgeon spleen in response to infection. Fish Shellfish Immunol, 70. 451-460.

[33] Li S, Wang D, Cao Y, Zhang Y, Liu H, et al. 2018. Transcriptome profile of Amur sturgeon (*Acipenser schrenckii*) liver provides insights into immune modulation in response to *Yersinia ruckeri* infection. Aquaculture, 492. 137-146.

[34] Kassahn KS, Crozier RH, Pörtner HO, Caley MJ. 2009. Animal Performance and Stress: Responses and Tolerance Limits at Different Levels of Biological Organisation. Biol Rev Camb Philos Soc, 84(2). 277–92. https://doi.org/10.1111/j.1469-185X.2008.00073.x.

[35] Komoroske LM, Connon RE, Jeffries KM, Fangue NA. 2015. Linking Transcriptional Responses to Organismal Tolerance Reveals Mechanisms of Thermal Sensitivity in a Mesothermal Endangered Fish. Mol Ecol, 24(19). 4960–81. https://doi.org/10.1111/mec.13373.

[36] Jesus TF, Grosso AR, Almeida-Val VMF, Coelho MM. 2016. Transcriptome Profiling of Two Iberian Freshwater Fish Exposed to Thermal Stress. J Therm Biol, 55. 54–61. https://doi.org/10.1016/j.jtherbio.2015.11.009.

[37] Ni M, Wen H, Li J, Chi M, Ren Y, et al. 2014. Two HSPs Gene from Juvenile Amur Sturgeon (*Acipenser schrenckii*): Cloning, Characterization and Expression Pattern to Crowding and Hypoxia Stress. Fish Physiol Biochem, 40(6). 1801–16. https://doi.org/10.1007/s10695-014-9969-9.

[38] Lee S, Hung SSO, Fangue NA, Haller L, Verhille CE, et al. 2016. Effects of Feed Restriction on the Upper Temperature Tolerance and Heat Shock Response in Juvenile Green and White Sturgeon. Comp Biochem Physiol A Mol Integr Physiol, 198. 87–95. https://doi.org/10.1016/j.cbpa.2016.04.016.

[39] Peng G, Zhao W, Shi Z, Chen H, Liu Y, et al. 2016. Cloning HSP70 and HSP90 Genes of Kaluga (*Huso dauricus*) and the Effects of Temperature and Salinity Stress on Their Gene Expression. Cell Stress Chaperones, 21(2). 349–59. https://doi.org/10.1007/s12192-015-0665-1.

[40] Zhang Y, Loughery JR, Martyniuk CJ, Kieffer JD. 2016. Physiological and Molecular Responses of Juvenile Shortnose Sturgeon (*Acipenser Brevirostrum*) to Thermal Stress. Comp Biochem Physiol A Mol Integr Physiol, 203. 314–21. https://doi.org/10.1016/j.cbpa.2016.10.009.

[41] Yuan X, Zhou YH, Huang YP, Guo WT, Johnson D, et al. 2017. Effects of Temperature and Fatigue on the Metabolism and Swimming Capacity of Juvenile Chinese Sturgeon (*Acipenser sinensis*). Fish Physiol Biochem, 43(5). 1279-1287. https://doi.org/10.1007/s10695-017-0371-2

[42] Zhang Y and Kieffer JD. 2017. The Effect of Temperature on the Resting and Post-Exercise Metabolic Rates and Aerobic Metabolic Scope in Shortnose Sturgeon *Acipenser brevirostrum*. Fish Physiol Biochem 43(5), 1245-1252. https://doi.org/10.1007/s10695-017-0368-x.

[43] Grunow B, Noglick S, Kruse C, Gebert M. 2011. Isolation of Cells from Atlantic Sturgeon *Acipenser Oxyrinchus Oxyrinchus* and Optimization of Culture Conditions. Aquat Biol, 14. 67–75. https://doi.org/10.3354/ab00383.

[44] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, et al. 2009. BLAST+: Architecture and Applications. BMC Bioinformatics,10. 1–9. https://doi.org/10.1186/1471-2105-10-421.

[45] Langmead B and Salzberg SL. 2012. Fast Gapped-Read Alignment with Bowtie 2. Nat Methods, 9(4). 357–59.

[46] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. 2009. The Sequence Alignment/Map Format and SAMtools. Bioinformatics, 25(16). 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

[47] Anders S, Pyl PT, Huber W. 2015. HTSeq-A Python Framework to Work with High-Throughput Sequencing Data. Bioinformatics, 31(2). 166–69. https://doi.org/10.1093/bioinformatics/btu638.

[48] Anders S and Huber W. 2010. Differential Expression Analysis for Sequence Count Data. Genome Biol, 11. R106. https://doi.org/10.1186/gb-2010-11-10-r106.

[49] Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, et al. 2003. PANTHER: A Library of Protein Families and Subfamilies Indexed by Function. Genome Res, 13(9). 2129–2141. https://doi.org/10.1101/gr.772403.

[50] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. 2000. Gene Ontology: Tool for the Unification of Biology. Nat Genetics, 25(1). 25–29. https://doi.org/10.1038/75556.

[51] Carbon S, Dietze H, Lewis SE, Mungall CJ, Munoz-Torres MC, et al. 2017. Expansion of the Gene Ontology Knowledgebase and Resources: The Gene Ontology Consortium. Nucleic Acids Res, 45(D1). D331–38. https://doi.org/10.1093/nar/gkw1108.

[52] Hudson DL. 2002. Keratins as Markers of Epithelial Cells. In: Wise C. (eds) Epithelial Cell Culture Protocols. Methods in Molecular Medicine, 188. Humana Press.

[53] Linares-Casenave J, Werner I, van Eenennaam JP, Doroshov SI. 2013. Temperature Stress Induces Notochord Abnormalities and Heat Shock Proteins Expression in Larval Green Sturgeon (*Acipenser medirostris*, Ayres 1854). J Appl Ichthyol, 29(5). 958–67. https://doi.org/10.1111/jai.12220.

[54] Humphreys DT, Carver JA, Easterbrook-Smith SB, Wilson MR. 1999. Clusterin has chaperone-like activity similar to that of small heat shock proteins. J Biol Chem, 274(11). 6875-6881. https://doi.org/10.1074/jbc.274.11.6875.

[55] Carver JA, Rekas A, Thorn DC, Wilson MR. 2003. Small heat-shock proteins and clusterin: intra- and extracellular molecular chaperones with a common mechanism of action and function. IUBMB Life, 55(12): 661-668. https://doi.org/10.1080/15216540310001640498

[56] Viard I, Wehrli P, Jornot L, Bullani R, Vechietti JL, Schifferli, Tschopp J, French LE. 1999. Clusterin gene expression mediates resistance to apoptotic cell death induced by heat shock and oxidative stress. J Invest Dermatol, 112(3). 290-296. https://doi.org/10.1046/j.1523-1747.1999.00531.x

[57] Jarolim S, Ayer A, Pillay B, Gee AC, Phrakaysone A, et al. 2013. *Saccharomyces Cerevisiae* Genes Involved in Survival of Heat Shock. G3 (Bethesda), 3(12). 2321–33. https://doi.org/10.1534/g3.113.007971.

[58] Phithakrotchanakoon C, Puseenam A, Phaonakrop N, Roytrakul S, Tanapongpipat S, et al. 2018. Hac1 Function Revealed by the Protein Expression Profile of a OtHAC1 Mutant of Thermotolerant Methylotrophic Yeast Ogataea Thermomethanolica. Mol Biol Rep, 45(5). 1311–1319. https://doi.org/10.1007/s11033-018-4287-4.

[59] Nobeyama Y and Nakagawa H. 2016. Aberrant DNA Methylation in Keratoacanthoma. PLoS ONE, 11(10). 1–14. https://doi.org/10.1371/journal.pone.0165370.

[60] Narberhaus F. 2002. Alpha-Crystallin-Type Heat Shock Proteins: Socializing Minichaperones in the Context of a Multichaperone Network. Microbiol Mol Biol Rev, 66(1). 64–93. https://doi.org/10.1128/MMBR.66.1.64-93.2002.

[61] Jäättelä, M. 1999. Escaping cell death: survival proteins in cancer. Exp Cell Res, 248(1). 30-43.

[62] Long Y, Li L, Li Q, He X, Cui Z. 2012. Transcriptomic Characterization of Temperature Stress Responses in Larval Zebrafish. PLoS ONE, 7(5). https://doi.org/10.1371/journal.pone.0037209.

[63] Arrigo AP. 2007. The Cellular 'Networking' of Mammalian Hsp27 and Its Functions in the Control of Protein Folding, Redox State and Apoptosis. Adv Exp Med Biol, 594. 14–26. https://doi.org/10.1007/978-0-387-39975-1_2.

[64] Wang K and Spector A. 1996. Alpha-Crystallin Stabilizes Actin Filaments and Prevents Cytochalasin-Induced Depolymerization in a Phosphorylation-Dependent Manner. Eur J Biochem, 242(1). 56–66. https://doi.org/10.1111/j.1432-1033.1996.0056r.x.

[65] Kamradt MC, Chen F, Sam S, Cryns VL. 2002. The Small Heat Shock Protein AB-Crystallin Negatively Regulates Apoptosis during Myogenic Differentiation by Inhibiting Caspase-3 Activation. J Biol Chem, 277(41). 38731–38736. https://doi.org/10.1074/jbc.M201770200.

[66] Smith CC, Yu YX, Kulka M, Aurelian L. 2000. A Novel Human Gene Similar to the Protein Kinase (PK) Coding Domain of the Large Subunit of Herpes Simplex Virus Type 2 Ribonucleotide Reductase (ICP10) Codes for a Serine-Threonine PK and Is Expressed in Melanoma Cells. J Biol Chem, 275(33). 25690–25699. https://doi.org/10.1074/jbc.M002140200.

[67] Kappé G, Boelens WC, de Jong WW. 2010. Why Proteins without an α-Crystallin Domain Should Not Be Included in the Human Small Heat Shock Protein Family HSPB. Cell Stress Chaperones, 15(4). 457–61. https://doi.org/10.1007/s12192-009-0155-4.

[68] Marvin M, O'Rourke D, Kurihara T, Juliano CE, Harrison KL, et al. 2008. Developmental Expression Patterns of the Zebrafish Small Heat Shock Proteins. Dev Dyn, 237(2). 454–63. https://doi.org/10.1002/dvdy.21414.

[69] Wang W, Deng DF, de Riu N, Moniello G, Hung SSO. 2013. Heat Shock Protein 70 (HSP70) Responses in Tissues of White Sturgeon and Green Sturgeon Exposed to Different Stressors. North Am J Aquac, 75(2). 164–69. https://doi.org/10.1080/15222055.2012.747457.

[70] Werner I, Linares-Casenave J, van Eenennaam JP, Doroshov SI. 2007 The Effect of Temperature Stress on Development and Heat-Shock Protein Expression in Larval Green Sturgeon (*Acipenser medirostris*). Environ Biol Fish, 79(3-4). 191–200. https://doi.org/10.1007/s10641-006-9070-z.

[71] Iwama GK, Vijayan MM, Forsyth RB, Ackerman PA. 1999. Heat Shock Proteins and Physiological stress in fish. Am Zool, 39(6). 901-909.

[72] Demeke A and Tassew A. 2015. Heat Shock Protein and their Significance in Fish Health. Research & Reviews: J Veterinary Sci, 2(1). 66-75.

[73] Oksala NKJ, Ekmekçi FG, Özsoy E, Kirankaya S, Kokkola T, et al. 2014. Natural thermal adaptation increases heat shock protein levels and decreases oxidative stress. Redox Biol, 3. 25-28.

[74] Heikkila JJ. 2017. The expression and function of hsp30-like small heat shock protein genes in amphibians, birds, fish and reptiles. Comp Biochem Physiol A Mol Integr Physiol, 203. 179-192.

[75] Zarate J and Bradley TM. 2003. Heat shock proteins are not sensitive indicators of hatchery stress in salmon. Aquaculture, 223. 175-187.

[76] Tomalty KMH, Meek MH, Stephens MR, Rincon G, Fangue NA, et al. Transcriptional response to acute thermal exposure in juvenile Chinook salmon determined by RNAseq. G3 (Bethesda) 5, 1335-1349.

[77] Chen Z, Farrel AP, Matala A, Narum SR. 2018. Mechanisms of thermal adaptation and evolutionary potential of conspecific population to changing environments. Mol Ecol 27(3), 659-674.

[78] Quinn NL, McGowan CR, Cooper GA, Koop BF, Davison WS. 2011. Identification of genes associated with heat tolerance in Arctic char exposed to acute thermal stress. Physiol Genomics, 43. 685-696.

[79] Healy TM, Tymchuk WE, Osborne EJ, Schulte PM. 2010. Heat shock response of killfish (*Fundulus heteroclitus*): candidate gene and heterologous microarray approaches. Physiol Genomics, 41. 171-184.

.

Assembly	**Cell**	**Organ**	**Embryo**	**COE**
Input reads	cell	organs	embryo	cell + organs + embryo
n. reads (million)	364,86	639,59	380,08	1424,41
Contigs	53,624	641,485	179,564	203,131
Assembly size (Mb)	34.80	342.28	166.71	254.00
N50 (Kb)	1.08	0.60	1.94	1.87
Max (Kb)	15.63	16.64	54.44	34.02
Accession number	GGWJ01000000	GGZX01000000	GEUL01000000	GGZT01000000
Mapped reads (%)	74.0	54.0	75.0	89.6

Table 1. Overview of assemblies' statistics and read mapping. List and number of input sequence reads for each assembly, number of resulting contigs, assembly length in Megabases (Mb), contig N50 in Kilobases (Kb), maximum contig length in Kilobases (kb) and percentage of cell-derived sequencing reads mapped to each assembly.

COE contig/s ID	Gene description	Sturgeon gene	Other names
75185, 75186, 75188	Heat shock protein beta-11	*hspb11a/b*	*hsp30*
81106, 81108	Heat shock protein beta-1	*hspb1*	*hsp27, hsp25*
68968	Heat shock 70kDa protein 4L	*hsph3a*	*hspa4l, apg1*
66436, 3989, 4740, 133784, 161880, 133767, 63068, 79530, 69039, 69847, 169012	Heat shock protein 90kDa alpha family class A member 1	*hspc1*	*hsp90aa1.2, hsp86, hsp90, hsp89*
113709	Heat shock 70kDa protein 4	*hsph2*	*hspa4, apg2, hsp110*
68283, 7082	Clusterin	*clu*	*clu*
22979	Growth factor receptor-bound protein 2	*grb2*	*grb2*
166695	atrial natriuretic peptide receptor 2	*npr2*	*npr2*
64489	Coiled-coil domain-containing protein 17	*ccdc17*	*ccdc17*
46154, 75199	n.a.	n.a.	n.a.
62264, 62266;	Uncharacterized protein	n.a.	n.a.

Table 2. List of the core 27 up-regulated contigs and corresponding gene description, sturgeon gene name and other names present in the literature.

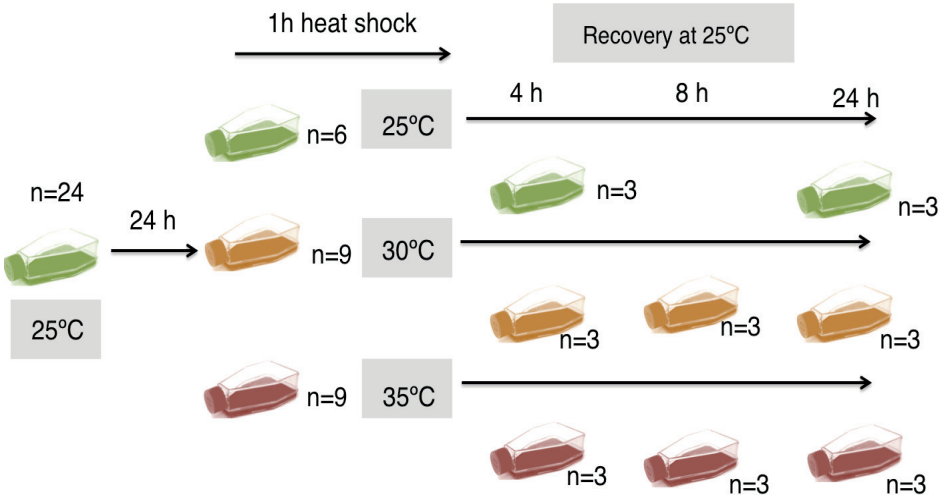


Figure 1. Experimental design. The AOXlar7y cell line was seeded in 25mL flasks and cultured at 25ºC for 24 hours (h). Then, the cells were treated at either 30º or 35º for 1 hour and then transferred back to 25ºC for recovery. RNA was harvested after recovery for 4, 8 and 24 hours and sequenced with Illumina HiSeq2500.

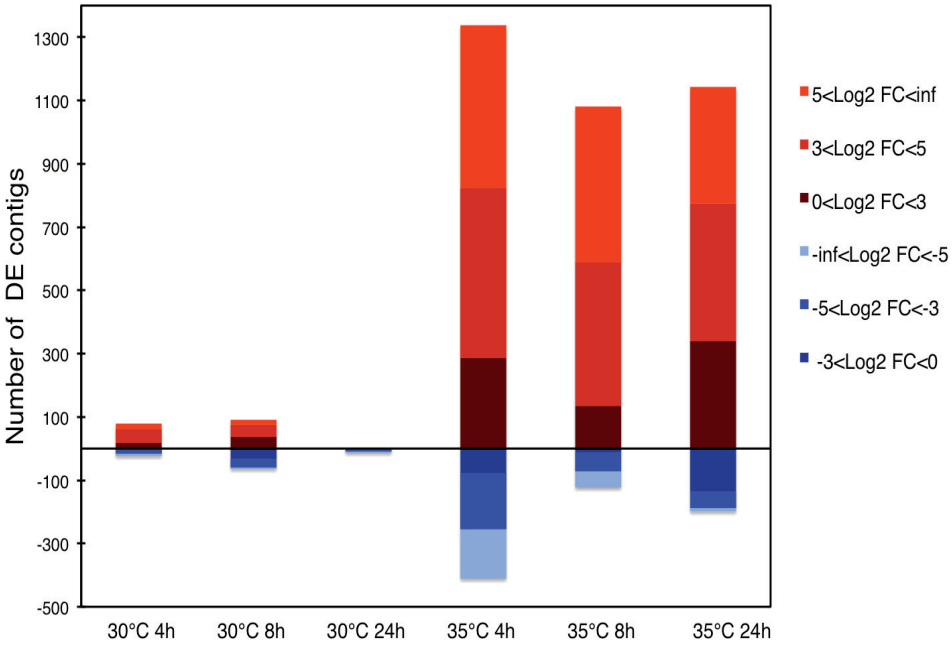


Figure 2. Number of differentially expressed contigs (DECs) per condition. Positive values represent upregulated contigs and negative values represent downregulated contigs. Different brightness represent different fold changes (FC) ranges.

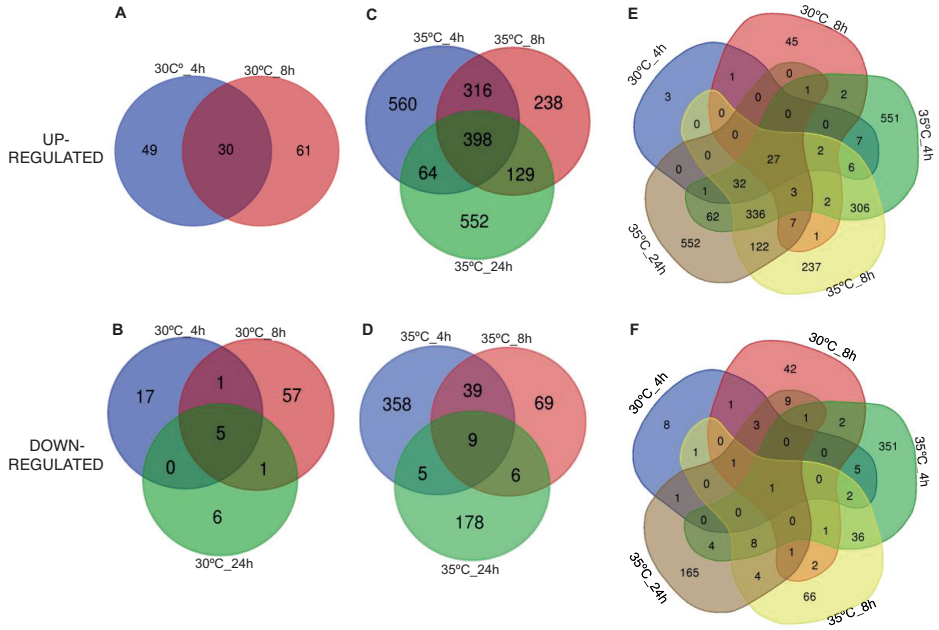


Figure 3. Venn diagrams showing the differentially expressed contigs (DECs) overlaps per condition. Up and down-regulated contigs at 30ºC and 35ºC at three timepoints (4, 8 and 24h) and overlap of upregulated and downregulated contigs at all the conditions, excluding 30ºC after 24 hours.

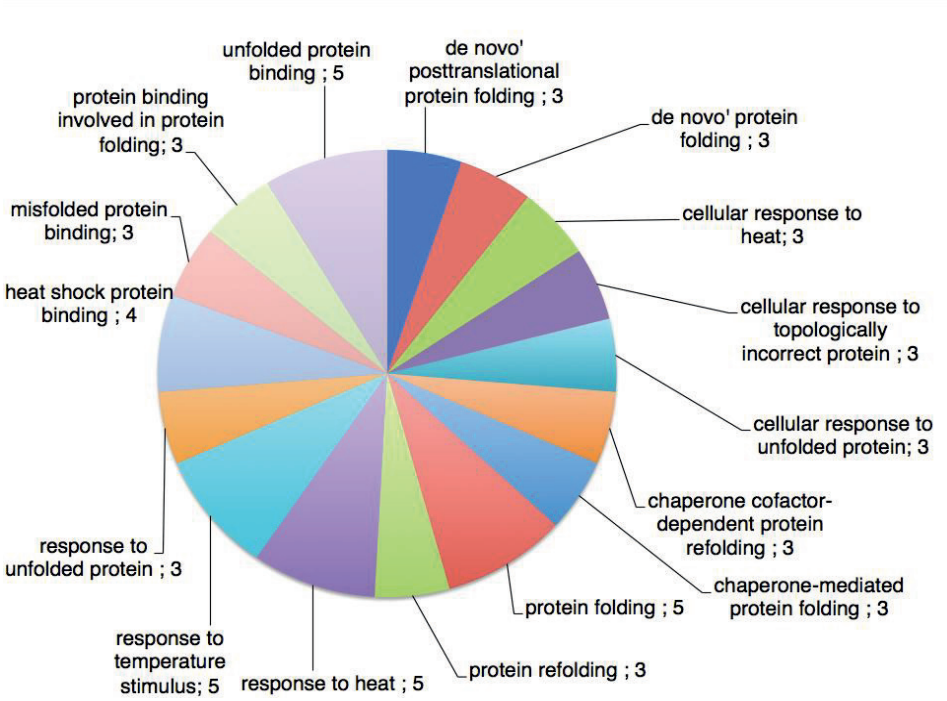


Figure 4. Pie chart showing overrepresented gene ontology (GO) terms at 30º. Biological process and molecular functions overrepresented at 30ºC after 4 hours, including the number of differentially expressed genes in each term.

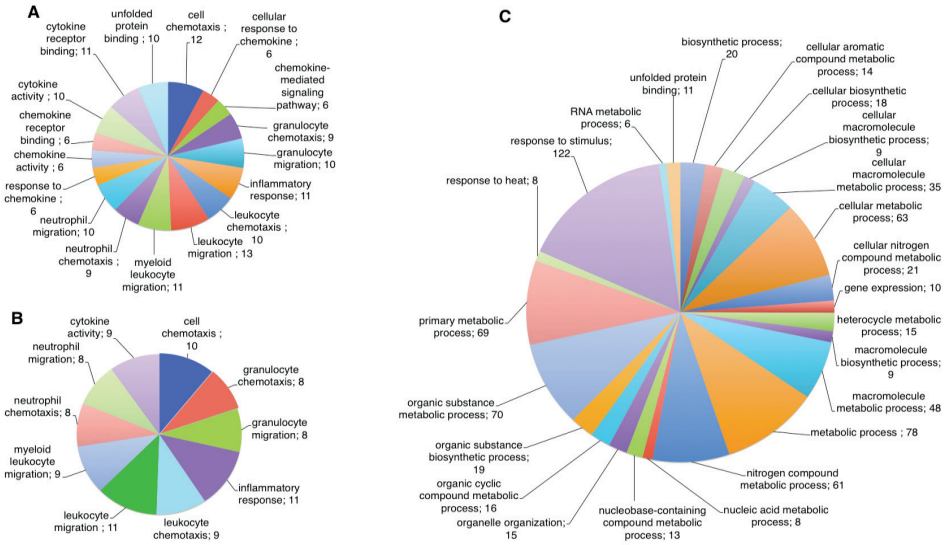


Figure 5. Pie chart showing overrepresented gene ontology (GO) terms at 35º. Biological process and molecular functions overrepresented at 35ºC after **A:** 4 hours; **B:** 8 hours; **C:** 24 hours, including the number of differentially expressed genes in each term

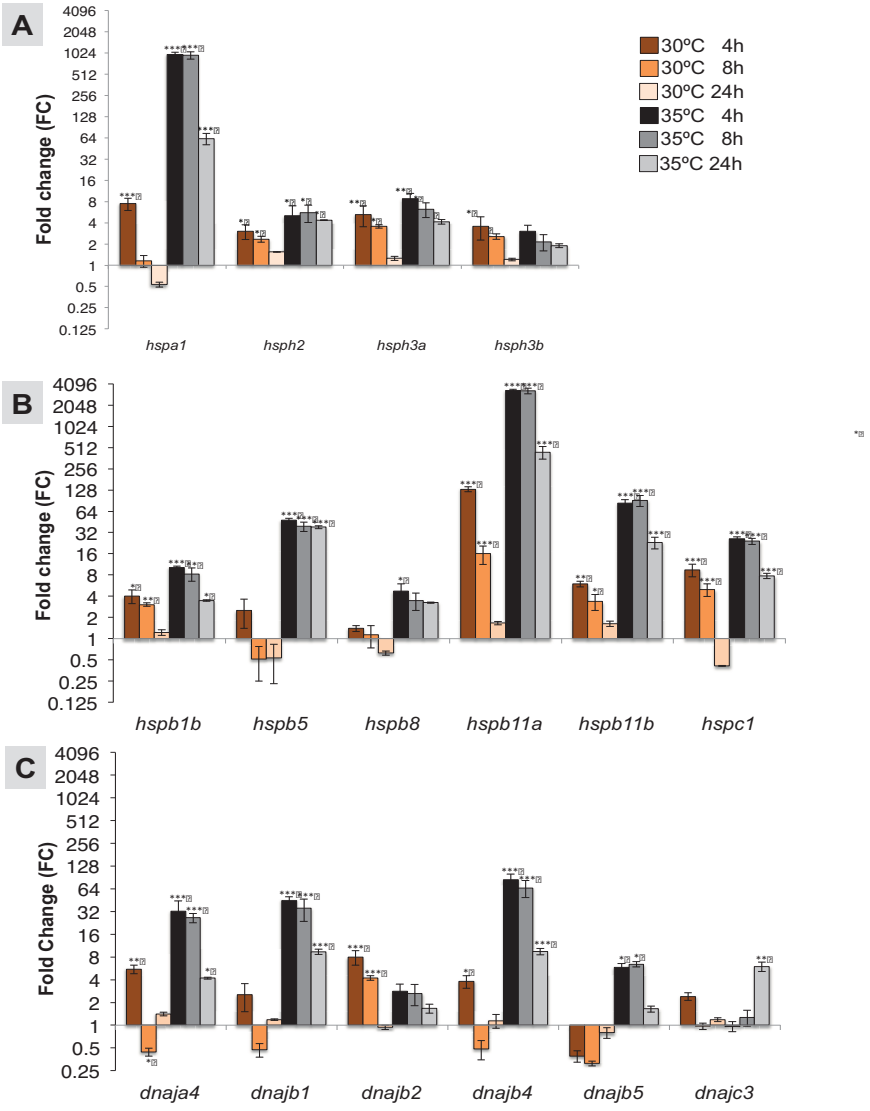


Figure 6. Fold change (FC) of the heat shock inducible genes per condition. a: HspA and HspH genes**; b:** HspB and HspC genes; **c:** DNAJ genes**.** Differences are considered significant when padj < 0.05. *padj < 0.05, **padj < 0.001, ***padj < 0.0001.

Deep transcriptome analysis of the heat shock response in an embryo-derived Atlantic sturgeon (*Acipenser oxyrinchus*) cell line

Deep transcriptome analysis of the heat shock response in an embryo-derived Atlantic sturgeon (Acipenser oxyrinchus) cell line

Elena Santidrián Yebra-Pimentel[a,b*], Marina Gebert[c], Hans J. Jansen[a], Susanne A. Jong-Raadsen[a] and Ron P.H. Dirks[a]

Supplementary Material

https:// doi.org/10.1016/j.fsi.2019.03.014.

Supplementary tables are available at:
https://www.dropbox.com/s/zitd49x6ros5a0z/Article1_supplementary_Tables.zip?dl=0

Supplementary Tables

- **Table S1.** Sample list and corresponding number of sequencing reads per sample and SRA Accession number.
- **Table S2.** List of top zebrafish BLASTx hits linked to of the COE transcriptomic contigs.
- **Table S3.** Number of sequencing reads mapped to the COE reference assembly per AOXlar7y sample, expressed in raw and normalized values (reads per kilobase and million, RPKM).
- **Table S4.** Top 100 expressed contigs in the untreated cell line (25ºC 4 and 24 h), indicating the RPKM mean and standard deviation (SD) and best BLASTx hits.
- **Table S5.** DESeq results per comparison. **Contig ID:** contig number in the reference transcriptome; **Base Mean:** Mean of mapped reads in all samples used for comparison; **Base Mean A:** Mean of mapped reads in the control group (25ºC 4 and 24 h); **Base Mean B:** Mean of mapped reads in the comparison group divided by the

mean of mapped reads in the control group. **Fold Change:** Mean of mapped reads in the comparison group divided by the mean of mapped reads in the control group; **log2FoldChange:** log2 of the Fold Change. Positive values indicate upregulation and negative values downregulation; **p-value:** Probability value; **padj:** Probability value adjusted for multiple testing with the Benjamini-Hochberg procedure;

- **Table S6.** List of differentially expressed contigs (DECs) in at least one condition and best zebrafish BLASTx hit.
- **Table S7.** List of the always upregulated (n=27) COE contigs and best BLASTx hits.
- **Table S8.** List of differentially represented GO terms per comparison.
- **Table S9.** List of annotated Hsp and Hsf genes in Atlantic sturgeon and BankIT accession.
- **Table S10.** Number of AOXlar7y sequencing reads mapped to the list of annotated genes and DESeq analysis results.

II

Temperature training improves transcriptional homeostasis after heat shock in juvenile Atlantic sturgeon (*Acipenser oxyrinchus*)

Elena Santidrián Yebra-Pimentel[a,b,*], Bruno Reis[c], Jörn Gessner[d], Sven Wuertz[d], and Ron P.H. Dirks[a].

[a] ZF-screens B.V., 2333CH Leiden, The Netherlands; [b] Department of Basic Sciences and Aquatic Medicine, Norwegian University of Life Sciences, 0454 Oslo, Norway; [c] Centro Interdisciplinar de Investigação Marinha e Ambiental (CIIMAR), Universidade do Porto, 4450-208 Matosinhos, Portugal; [d] Leibniz Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany.

* Corresponding author: Elena Santidrián Yebra-Pimentel, ZF-screens B.V., 2333CH Leiden, The Netherlands (elena@zfscreens.com).

Abstract

Exposure to high temperatures can lead to thermotolerance in fish, which is hypothesized to potentially improve post-release survival in species under restocking programs, like Atlantic sturgeon. The aim of this study was to determine whether Atlantic sturgeon juveniles exposed to a four-week temperature treatment respond differently to a subsequent heat shock than juveniles exposed to heat shock for the first time (naive fish). Response to heat shock was assessed by mapping the liver transcriptome. In total 838 unique contigs were differentially expressed between the trained and the control group (592 downregulated, 261 upregulated and 15 down- or upregulated, depending on the condition), corresponding to genes involved in the response to heat, tissue damage, proteolysis and metabolism. Temperature-trained fish showed 2-4-fold fewer dysregulated contigs than naive fish, indicating their ability to maintain and recover homeostasis faster. During heat shock, *hspc1* was upregulated in both experimental groups, while *hspa1* and *dnaja4* were exclusively upregulated in the control. Overall, compensatory mechanisms were observed in addition to the heat shock response. Only two genes, *fgg* and *apnl*, were upregulated at nearly all timepoints in both groups. Peptidases were more strongly downregulated in control fish, which also showed a reduction in lipid metabolism during recovery. Keratins, *pck1, gadd45ga* and *gadd45gb* were differentially expressed between trained and control fish, and due to their roles in tissue protection and ER stress reduction, they might be responsible for the maintenance of the transcriptional homeostasis observed in trained fish.

Keywords: *Acipenser oxyrinchus*; heat shock proteins (HSP); training; temperature; transcriptome, fitness.

1. Introduction

Wild sturgeon (*Acipenseridae*) populations are in severe decline and are currently considered among the most endangered fish species worldwide [1]. Due to their late onset of maturity, these diadromous fish are especially vulnerable to anthropogenic and climate impacts. Atlantic sturgeon (*Acipenser oxyrinchus*) populations used to inhabit the Atlantic seaboard of North America from Canada to Florida as well as European waters, being restricted lately to the Baltic Sea. However, the Baltic population is currently extinct [2].

Despite efforts to reintroduce the Atlantic sturgeon into the Baltic [2], major concerns are addressing the post-release mortality [3]. Common hatchery practices are characterized by high stocking densities in stimulus-deprived tanks with fixed abiotic conditions, such as constant temperature, leading to reduced fitness of farmed sturgeons when compared to their wild conspecifics [4]. In order to increase restoration success, innovative rearing techniques, such as enrichment of abiotic factors, are increasingly applied to increase the fitness of juveniles produced for release [3,5,6].

Moreover, climate change has globally affected discharge and water temperatures with recent models predicting increases of 2ºC to 6ºC in the next century [7,8]. Increased water temperature has particularly adverse impacts on ectothermic animals such as fish, affecting their swimming performance, predator avoidance and foraging behavior and resulting in ecosystem shifts [9,10]. Therefore, understanding the mechanisms of thermal adaptation is critical in order to adapt restoration programs.

Cells that have been exposed to thermal stress can acquire increased resistance against the detrimental effects of a subsequent thermal stress [11]. This phenomenon, called thermotolerance, implies the induction of the heat shock response (HSR) that triggers the transcription of heat shock protein (HSP) genes [12,13], resulting in an enhanced resistance against protein denaturation and aggregation.

Previous studies have unraveled the transcriptional changes after temperature treatments and have documented a large number of genes associated with heat shock (HS) and thermal adaptation [14-16], including the highly conserved HSPs (e.g. *hspa8*) [17-20], and genes controlling cell cycle, energy metabolism and protein folding, repair and biosynthesis [21,22].

Temperature adaptation results from phenotypic plasticity and has been shown to occur among conspecific populations occupying different environments [23-25]. Thermotolerance may play an essential role in enhancing performance and survival of fish under variable thermal conditions.

The current study aimed to assess whether exposure to fluctuating temperatures comparable to those experienced in the wild can provide increased protection during a subsequent temperature increase event in Atlantic sturgeon. In the training phase, the fish were subjected to differing four-week temperature treatments followed by exposure to a HS. RNA sequencing (RNAseq) was used to map the transcriptomic changes in the liver, as the most important organ in cellular metabolism, detoxification, stress response and adaptation

[26]. Results were quantified to confirm the response of HSP genes but also to identify additional genes and biological pathways, providing helpful information towards understanding of molecular mechanisms underlying the heat stress response in Atlantic sturgeon.

2. Materials and methods

2.1. Experimental animals and sampling

The experiments were performed at the facilities of the Leibniz-Institute of Freshwater Ecology and Inland Fisheries (IGB, Berlin, Germany) using fish from the regular reproduction and rearing of the *ex situ* stock. The experimental work was carried out under the regulations of the German Animal Welfare Law in compliance with the EU Directive 2010/63/EU and approved by the regulatory administration (LAGeSo G 0305/15). Three-month-old Atlantic sturgeon juveniles (14.9 ± 4.1 g body weight; 16.8 ± 1.8 cm total length) were randomly distributed to two raceway units (2.40 m * 0.225 m * 0.1 m) at a natural photoperiod and acclimatized for 7 days. Two experimental groups were established: during the 30 days prior to the HS experiment, the control group (C; n=12) was constantly kept at 20ºC, while the trained group (T; n=12) was subjected to 7 rounds of a temperature challenge effectuated by an increase of the water temperature from 20ºC to 24ºC over 12 hours and a subsequent decline to 20ºC over the following 12 hours. The seven rounds of temperature challenge were randomly distributed over the 30-day period to avoid habituation. The temperature fluctuactions represent values typically observed in the recipient water body (Oder River) for the stocked fish during summer.
The HS challenge (Fig. 1) comprised the simultaneous exposure of both groups to the temperature challenge. At the beginning of the temperature increase (t=0h), at the peak of the heat treatment (t=12h), at the return to basal temperature (t=24h) and after 24 hours of recovery (t=48h), three animals per group (n=3) were euthanized with MS222 (300ppm), decapitated and their livers were sampled. Temperature was recorded hourly throughout the training period and every 15 min during the temperature challenge on day 31 using PRTemp101 data loggers (MadgeTech, USA). Oxygen was determined daily with a Hach HQ40d multimeter (> 8 mg/L).

2.2. RNA isolation, Illumina libraries and sequencing

RNA was extracted with TRIzol [27]. Integrity and concentration were analyzed with a Bioanalyzer 2100 total RNA Nano series II chip (Agilent, Santa Clara, USA). RNA libraries were prepared from 500 ng total RNA, using the Illumina TruSeq Stranded mRNA Sample Preparation Kit according to the manufacturer's instructions (Illumina Inc., San Diego, USA) and the resulting libraries were evaluated with a Bioanalyzer 2100 DNA 1000 series II chip (Agilent, Santa Clara, USA). One biological replicate of the trained group (12h after HS) was lost due to RNA degradation. All the libraries were sequenced using an Illumina HiSeq2500 instrument as single 51nt reads up to a minimum of 10 million reads per sample. Image analysis and base calling were done using the Illumina pipeline.

2.3. Transcript quantification, differential expression and Gene Ontology (GO) analysis

Bowtie2 (version 2.2.5) [28] was used to align the Illumina reads to the recently published cell-organ-embryo (COE) *de novo* transcriptome assembly and to the complete annotated set of Atlantic sturgeon HSP and heat shock factor (HSF) genes [29], available at GenBank (Accession numbers MH777912-MH777987 and MH917287-MH917292, respectively). SAMtools (version 1.2.) [30] was used to exclude secondary alignments, and the remaining alignments were counted using the HTSeq Python library (version 0.5.3p9) [31]. Read counts for 203,131 contigs were analyzed in R (version 3.3.0) using the Bioconductor package DESeq (version 1.30.0) [32]. P-values were adjusted for multiple testing with the Benjamini-Hochberg procedure, which controls the false discovery rate (FDR). Contigs with an adjusted p-value (padj) < 0.05 were considered differentially expressed at each pairwise comparison between different treatments and timepoints and were blasted against the zebrafish proteome (BLAST 2.8.1+) with an E-value cutoff of 1E-5. Reads were normalized between samples and corrected for contig length using the RPKM (reads per kilobase million) correction.

GO enrichment analysis was performed using the Fisher's Exact test with Bonferroni correction implemented by the PANTHER Overrepresentation test (released on 2017-12-05) using the GO Ontology database [33] released on 2018-07-03. GO terms with a p-value < 0.05 were considered overrepresented in each dataset.

3. Results

To determine whether prior exposure to a four-week temperature treatment affects the response of Atlantic sturgeon juveniles to a new HS, we exposed both control and trained fish to an experimental temperature challenge (Fig. 1) and measured mRNA levels in the liver of 3 fish per treatment and timepoint using Illumina RNAseq. We obtained between 10.3 and 20.3 million sequencing reads per sample, 80.4-95.7% of which could be aligned to a cell-organ-embryo (COE) sturgeon transcriptome (Accession n. GGZT01000000) including 82 previously annotated HSP cDNA contigs (Table S1) [29]. The Illumina reads were deposited at the NCBI Short Read Archive (SRA) database under the accession number PRJNA534021.

3.1. General transcriptional response

DESeq analysis showed 838 unique differentially expressed contigs (DECs) in at least 1 experimental group and timepoint, 592 of which were downregulated, 261 were upregulated, and 15 were up or downregulated, depending on the treatment and timepoint (Table S2). DECs were blasted against the zebrafish proteome (BLAST 2.8.1+) with an E-value cutoff of 1E-5, resulting in 420 hits, 337 protein versions and 321 genes (Table S3). We therefore analyzed differential expression of these 838 contigs between trained fish and their respective controls.

In order to determine general trends in gene expression both during HS and recovery, the numbers of up and downregulated contigs per experimental group were plotted in Venn diagrams (Fig. 2). At all tested time points, the number of differentially expressed genes in

the control exceeded those in the trained group by a factor of 2-4. During the HS peak (t= 12h) higher numbers of downregulated than upregulated contigs were observed in both groups, while during the HS relaxation period, when the temperature was restored to initial levels (t= 24h), the number of downregulated contigs decreased in both groups (Fig. 2B). During recovery (t= 48h) downregulation increased to higher values than during the HS peak, but the control group showed about two times more upregulated (N=81) and downregulated contigs (N=265) than the trained group (N=45 and N=142, respectively) (Table S2).

3.2. Heat shock response

During the HS peak (t =12h) both experimental groups expressed high levels of HSPs and molecular chaperones*: hspc1, serpinh1a* and *serpinh1b* were significantly upregulated in both fish groups, while *hspa1* and *dnaja4* were only upregulated in the control fish, for which response to heat (GO: 0009408) was overrepresented (FDR= 3.81E-4) (Fig. 3 and Tables S2 and S4). During the HS relaxation period HSPs were no longer upregulated, however, *dnajc14* was downregulated in the control group. Finally, during recovery (t=48), control fish showed downregulation of *hspa5*. Before the temperature challenge no HSPs were differentially expressed between the control and the trained group, yet, genes involved in cell cycle regulation, like growth arrest and DNA-damage-inducible gamma a and gamma b (*gadd45ga* and *gadd45gb*) and 4-aminobutyrate aminotransferase (*abat*) were significantly upregulated in trained fish. Of these, *gadd45ga* and *gadd45gb* returned and stayed at basal levels during the temperature challenge in both groups*, while abat* remained upregulated in trained fish during the HS peak. Coiled-coil domain containing 17 *(ccdc17)* was also upregulated in response to temperature in both fish groups, but returned to basal levels during recovery.

3.3. Tissue damage

Except for trained fish during the HS peak, contigs corresponding to actinoporin-like (*apnl)* and fibrinogen gamma chain (*fgg)* were the only ones that were upregulated in all conditions (Fig. 3). Both genes were also upregulated in trained fish before the temperature challenge. On the other hand, keratins (*krt5, krt94, krt97*) were downregulated before the temperature challenge in trained fish and during the HS peak in control fish. Catabolic processes (GO:0009056) were underrepresented during HS in the control group, while no GO terms were differentially represented in trained fish (Tables S2 and S4).

3.4. Proteolysis

More than 10 peptidases were differentially expressed through the temperature challenge in both experimental groups. Among them, carboxypeptidases A4 and B1 (*cpa4, cpb1*), elastase-like 3 (*elal3*), chymotrypsin-like elastase (*cela*), trypsin (*try*) and protein disulfide isomerase family A member 2 (*pdia2*) were downregulated in the liver of trained fish before the temperature challenge and during recovery, and during the HS relaxation period in the control fish. Chymotrypsin-like protease (*ctrl*) and elastase-like 2 (*elal2*) were only

downregulated in the control fish during HS relaxation (Fig. 4). Contigs corresponding to cathepsin D (*ctsd*) and Ba (*ctsba*) were downregulated in trained fish before the temperature challenge, of which *ctsba* was also downregulated in trained fish during HS peak and *ctsba* was downregulated in control fish at all the timepoints and during HS relaxation in trained fish. Only one peptidase, matrix metallopeptidase 2 (*mmp2*), and one phosphorylase, phosphorylase kinase gamma 1 a (*phkg1a*), were upregulated during the experiment: *mmp2* during HS peak and relaxation and *phkg1a* during HS relaxation and recovery. Overall, proteolysis (GO:0006508) was underrepresented in trained fish during the HS relaxation (FDR= 2.75E-5) and during recovery (FDR= 7.18E-4) (Fig. 4, Tables S2 and S4).

3.5. Metabolic response

Phosphoenolpyruvate carboxykinase (*pck1*) expression was induced in trained fish before the temperature challenge and during HS relaxation, and during recovery in both experimental groups. No more genes were upregulated in the trained fish during the temperature challenge; however, apolipoproteins (*apoa4b.1* and *apoa4b.2*) were responsive to HS both during the peak and relaxation period in control fish. Despite this observation, no metabolism-related GO terms were found overrepresented. Fatty acid desaturase 2 (*fads2*) was upregulated in trained fish before the temperature challenge and in control fish during HS relaxation, while strongly downregulated during recovery in both groups.
The majority of dysregulated genes involved in metabolic processes were downregulated. One third of these genes, including lectin galactosidase-binding soluble 9 like 1 and 3 (*lgals9l1*, *lgals9l3*), solute carrier family 15 member 1 b (*slc15a1b*), S100 calcium binding protein A 10 b (*s100a10b*) were downregulated in both experimental groups at all conditions, and anterior gradient related 2 (*agr2*) was also downregulated at all conditions with the exception of trained fish during recovery. More than half of the downregulated genes were uniquely downregulated during recovery in either both groups (e.g. *cyp51, dhcr24, gck*) or only in the control fish (e.g. *shbg, dhcr7, fa2h*). Overall, lipid metabolic process (GO:0006629) was underrepresented in the control group during recovery (FDR= 2.09E-4) (Tables S2 and S4).

4. Discussion

4.1. General transcriptional response

After a four-week temperature treatment, fish were exposed to HS and the liver transcriptome was mapped in order to study the response in comparison to naive fish. In this experiment a fluctuation of the temperature regime comparable to the fluctuations observed in the restocking ecosystem (Oder River) during the summer affected the transcriptional response to a subsequent HS. In both groups, the vast majority of transcriptional changes corresponded to downregulation, showing similar temporal patterns in both experimental groups: dysregulated genes decreased during HS relaxation and increased again during recovery. Even though the pattern was similar in both groups, the control group showed 2 to 4 times more DECs than the trained one. This response was characterized by particularly increased downregulation, indicating that naïve fish initiate a more complex response to

maintain (during HS) and reestablish (during recovery) transcriptional homeostasis. In addition to the upregulation of heat shock specific genes (HSPs), other physiological networks are involved in this response, including cell cycle regulation, proteolysis, tissue damage and lipid and carbohydrate metabolism.

4.2. Heat shock response

A directed response to heat was observed in both experimental groups during the HS peak and was absent at other timepoints, even during HS relaxation, which proves that the HSR is a transient defense mechanism. While *hspc1* and *serpinh1 a* and *b* were upregulated in both experimental groups, *hspa1* and *dnaja4* were only upregulated in control fish, in which heat shock response was an overrepresented GO term. This is supported by previous studies using warm-adapted fish that showed that *serpinh1a* was upregulated in salmonids, which also showed high levels of *hspc1* in response to HS [34], while the cyprinid doctor fish (*Garra rufa*) had higher *hspa1* (*hsp70*) protein levels in the muscle [25]. To date, *in vivo* experiments with sturgeons have not provided a temperature treatment previous to the HS, addressed only a limited number of HSP genes and used approaches other than RNAseq (e.g. western blot, qPCR). In addition, results depended on the species and age of the exposed fish, finding that *hspa1* was HS inducible in both white (*A. transmontanus*) and green (*A. medirostris*) sturgeon larvae [35,36] while it was more responsive to cold in Beluga sturgeon juveniles (*Huso huso*) [37]. In our recently published RNAseq experiment [29] we exposed a cell line derived from Atlantic sturgeon larvae to two HS treatments (5ºC and 10ºC increase), identified and annotated 76 HSP genes in the COE reference transcriptome, and analyzed DECs. The cell line showed 3.6 times more DECs than the juveniles used in this study, including *serphin1b* and 17 HSP genes. Although *hspc1, hspa1* and *dnaj4* were among these genes (up to 26, 966 and 32 fold upregulation, respectively), *hspb11a* was the most upregulated gene (> 3,000-fold). However, the HS applied *in vivo* in this experiment resulted in a much milder upregulation, and while *hspc1* was upregulated in both groups (up to 11.5-fold), *hspa1* was only upregulated provided the absence of training (6.5-fold) and *hspb11a* expression was not affected. HSP genes are known to be responsive to other stress besides HS and, overall, only 3 out of 76 were upregulated in the liver under these experimental conditions and selected timepoints.

Apoptosis and cell cycle arrest during HS is known to be an energy-saving mechanism to direct energy towards the repair of damaged proteins [38]. This is the case for the gadd45 family [39], which confers HS resistance and increased lifespan in *Drosophila* [40-42] and is upregulated (*gadd45ga* and *b)* before the temperature challenge in trained fish.

Finally, *ccdc17* was also upregulated during the HS peak in both naïve and trained fish, and although this gene family is known to be involved in gene expression regulation [43], the specific function of this family member has not been studied in detail and its relevance remains unclear.

The few studies that examined the effects of thermal cycles on thermotolerance in fish show that the mechanisms involved go beyond the HSR and cell cycle regulation, including biological processes related to protein folding, energy metabolism and biosynthesis [44,45].

4.3. Tissue damage

With the exception of trained fish during the HS peak, *apnl* and *fgg* were upregulated at all timepoints in both experimental conditions. Although fibrinogen genes are constitutively expressed in the liver, the expression can be increased during inflammatory stress [46] and is induced by freezing in the wood frog (*Rana sylvatica*) [47]. Actinoporins are pore-forming proteins which have previously been identified to play an important role in water stress in moss [48,49]. Although, to our knowledge, none of these genes have been attributed to HS protection, *apnl* was also upregulated in the cell line [29] and further research should be performed to investigate its relevance. Keratins are important tissue-protecting genes and, when associated with HSPs, are targets of the caspase-mediated proteolysis [50]. Three family members (*krt5, krt94, krt97*) were downregulated in trained fish before HS but returned to basal levels during the HS peak. Opposite to this, they were very strongly downregulated (>190-fold) during the HS peak in control fish, indicating lack of protection.

4.4. Proteolysis

Many proteases (e.g. *cpa4, cpb1*, *elal3*, *cela*, t*ry* and *pdia2*) were downregulated in the liver of trained fish before the temperature challenge and during recovery, and earlier in control fish (during HS relaxation). In addition, most of them (e.g. *cela*) were more downregulated in control (1,424-fold) than in trained fish (11-fold). Many of these proteases (e.g. *pdia2 and agr2)* are involved in protein maturation and secretion [51].
Only one protease (*mmp2)* and one phosphorylase (*phkg1a)*, which are known to be responsive to heat [52], were upregulated during the experiment. Overall, proteolysis was underrepresented in both trained and naïve fish during recovery.

4.5. Metabolic response

Previous studies already indicated that genes involved in metabolism regulation (e.g. glycolysis, gluconeogenesis, tricarboxylic acid cycle, lipid metabolism) could be either induced or repressed during HS [22,53,54]. Among the genes involved in carbohydrate metabolism, *pck1*, upregulated in both experimental conditions, is involved in gluconeogenesis, whereas *gck*, involved in glycolysis, was downregulated during recovery in both groups. Pck1 was first upregulated in trained fish (before temperature challenge and during HS relaxation) and afterwards, during recovery, upregulated in both experimental groups. Previous studies already described *pck1* upregulation during HS in pigs and fish [55,56], possibly triggered by cortisol and resulting in a decrease of endoplasmic reticulum (ER) stress.
The remaining dysregulated genes had important functions in lipid metabolism, the majority of them being downregulated, with the exception of *fads2* and *apoa4b 1* and *2*. *Hspa1*, which was upregulated in control fish, is known to degrade apolipoproteins [57], which are likely upregulated in control fish during HS peak as a counteracting response. Even though *fads2* was initially upregulated in trained fish, it was downregulated in both experimental groups during recovery, like the majority of the genes in this category (e.g. *cyp51, dhcr24).*

Still, some genes (e.g. *shbg, dhcr7, fa2h*) were only downregulated in the control group, for which the "lipid metabolism" GO term was underrepresented during recovery. Temperature is an abiotic factor that can affect the fluidity of the membranes by changing the content of membrane lipids and the degree of unsaturation [58,59]. Therefore, downregulation of *fads2* and genes related to cholesterol biosynthesis (e.g. *dhcr7, dhcr24, cyp51*) are indicative of membrane rearrangements after HS recovery, and were also previously reported in rainbow trout [60].

It is important to take in consideration that we have analyzed the HSR using only one tissue, while previous studies have reported tissue-specific patterns in response to HS stress [38]. Also, while the temperature increase in previous experiments using sturgeons was up to 10ºC above the optimal growing temperature [22,36,37], in our study the temperature challenge was designed to meet the natural fluctuations in the recipient water body (+ 5ºC).

In conclusion, we have found that training reduces gene dysregulation during the temperature challenge, possibly due to faster homeostasis recovery or reduced sensitivity. Most of the HSP genes were not involved in the observed HSR and *hspa1* upregulation was only observed in naïve fish. Compared to the cell line, the HSR was milder due to the less severe heat shock and potentially due to compensatory mechanisms at the organismal level involving multiple physiological networks. Still, the experimental design, sampled tissue and selected timepoints could also have influenced the results. Control fish showed increased proteolysis, downregulation of protective genes (e.g. keratins) during HS and reduced lipid metabolism during recovery. It can be predicted that trained fish will be better adapted for the environmental fluctuation after release than the untrained fish, improving their acclimatization and survival. The differences between the two groups reveal large potential for improvement of rearing conditions in conservation aquaculture and still, new techniques, like the exposure of the broodstock to HS to produce more resistant offspring [61] could be investigated.

Funding

This work was supported by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Actions: Innovative Training Network "IMPRESS" [Grant Agreement No. 642893].

Acknowledgements

The authors would like to thank Eva Kreuz for her support for and contribution to the sampling and laboratory work.

Conflicts of Interest

The authors declare no conflict of interest related to this publication.

Supplementary Material

Table S1. Sample list, corresponding number of sequencing reads per sample, accession and number and normalized (RPKM, reads per kilobase and million) reads mapped to the reference transcriptome, including only contigs with more than 0 RPKM in at least one sample.

Table S2. DESeq results (padj <0.05) per comparison.

Table S3. List of DECs in at least one condition, including query (contig) size in nucleotides, log2FoldChange and padj value in each of the comparisons and best BLASTx zebrafish hit.

Table S4. List of under and overrepresented gene ontology (GO) terms per condition.

References

[1] IUCN. (2018). The IUCN Red List of Threatened Species. Gland, Switzerland.

[2] Gessner, J., Arndt, G. M., Kapusta, A., Gushin, S. A., Pilinkovskij, A., Povliunas, J., Medne, R., Purvina, S., Tambets, M. and Moeller, P. R. (2019). HELCOM -Action Plan for the protection and recovery of the Baltic sturgeon *Acipenser oxyrinchus* in the Baltic Sea area. March 7th, 99p. Helicom Conv.

[3] Camara-Ruiz, M., Santo, C. E., Gessner, J., & Wuertz, S. (2019). How to improve foraging efficiency for restocking measures of juvenile Baltic sturgeon (*Acipenser oxyrinchus*). *Aquaculture*, 502, 12-17.

[4] Sulak, K. J., Parauka, F., Slack, W. T., Ruth, R. T., Randall, M. T., Luke, K., Mettee, M. F. and Price, M. E. (2016). Status of scientific knowledge, recovery progress, and future research directions for the Gulf sturgeon, *Acipenser oxyrinchus desotoi* Vladykov, 1955. *J. Appl. Ichthyol.* 32, 87–161.

[5] Salvanes, A. G. V., Moberg, O., Ebbesson, L. O. E., Nilsen, T. O., Jensen, K. H., Braithwaite, V. A. (2013). Environmental enrichment promotes neural plasticity and cognitive ability in fish. *Proc R Soc Biol Sci.* doi:10.1098/rspb.2013.1331.

[6] Mes, D., von Krogh, K., Gorissen, M., Mayer, I., Vindas, M. A. (2018). Neurobiology of wild and Hatchery-Reared Atlantic Salmon: How Nurture Drives Neuroplasticity. *Front Behav. Neurosci.* doi:10.3389/fnbeh.2018.00210.

[7] Habary, A., Johansen, J. L., Nay, T. J., Steffensen, J. F. and Rummer, J. L. (2016). Adapt, move, or die- how will tropical coral reef fishes cope with ocean warming? *Glob. Chang. Biol.* 23, 566–577.

[8] IPCC. (2013). Climate change: The physical science bases. Contribution of working group I to the fifth assessment report of the Intergovernmental Panel on Climate Change. Cambridge, United Kindom. New York, NY, USA. Cambridge Univ. Press.

[9] Fields, P. A., Graham, J. B., Rosenblatt, R. H. and Somero, G. N. (1993). Effects of expected global climate change on marine faunas. *Trends Ecol. Evol.* 8, 361–367.

[10] Hofmann, G. E. and Todgham, A. E. (2010). Living in the Now: Physiological Mechanisms to Tolerate a Rapidly Changing Environment. *Annu. Rev. Physiol.* 72, 127–145.

[11] Kampinga, H. H. (1993). Thermotolerance in mammalian cells. *J. Cell Sci.* 104, 11–17.

[12] Dietmar, K. (2005). Molecular and Evolutionary basis of the cellular Stress Response. *Annu. Rev. Physiol*. 67, 225–257.

[13] Hirsch, C., Gauss, R. and Sommer, T. (2006). Coping with stress : cellular relaxation techniques. *Trends Cell Biol.* 16, 657–663.

[14] Morris, M. R. J., Richard, R., Leder, E. H. and Barrett, R. D. H.. (2014). Gene expression plasticity evolves in response to colonization of freshwater lakes in threespine stickleback. *Mol. Ecol*. 23, 3226–3240.

[15] Narum, S. R. and Campbell, N. R. (2015). Transcriptomic response to heat stress among ecologically divergent populations of redband trout. *BMC Genomics*. 16, 103.
[16] Quinn, N. L., Mcgowan, C. R., Cooper, G. A., Koop, B. F. and Davidson, S. (2011). Identification of genes associated with heat tolerance in Arctic charr exposed to acute thermal stress. *Physiol Genomics.* 43, 685–696.
[17] Dubeau, S. F., Pan, F., Tremblay, G. C. and Bradley, T. M. (1998). Thermal shock of salmon in vivo induces the heat shock protein hsp 70 and confers protection against osmotic shock. *Aquaculture.* 168, 311–323.
[18] Iwama, G. K., Vijayan, M. M., Forsyth, R. B. and Ackerman, P. A. (1999). Heat shock proteins and physiological stress in fish 1. Am. Zool. 39, 901–909.
[19] Smith, T. R., Tremblay, G. C. and Bradley, T. M. (1999). Characterization of the heat shock protein response of Atlantic salmon (*Salmo salar*). *Fish Physiol. Biochem.* 20, 279–292.
[20] Zarate, J. and Bradley, T.M. (2003). Heat shock proteins are not sensitive indicators of hatchery stress in salmon. *Aquaculture*. 223, 175–187.
[21] Jeffries, K. M., Hinch, S. G., Sierocinski, T., Pavlidis, P. and Miller, K. M. (2014). Transcriptomic responses to high water temperature in two species of Pacific salmon. *Integr Comp Biol*. 39, 901–909.
[22] Liu, S., Wang, X., Sun, F., Zhang, J., Feng, J., Liu, H., Rajendran, K. V, Sun, L., Zhang, Y., Jiang, Y., et al. (2013). RNA-Seq reveals expression signatures of genes involved in oxygen transport, protein synthesis, folding , and degradation in response to heat stress in catfish. *Physiol Genomics.* 45, 462–476.
[23] Kavanagh, K. D., Haugen, T. O., Gregersen, F., Jernvall, J. and Vøllestad, L.A. (2010). Contemporary temperature-driven divergence in a Nordic freshwater fish under conditions commonly thought to hinder adaptation. *BMC Evol. Biol.* 10, 350.
[24] Keller, I. and Seehausen, O. (2012). Thermal adaptation and ecological speciation. *Mol. Ecol.* 21, 782–799.
[25] Oksala, N. K. J., Ekmekçi, F. G., Özsoy, E., Kirankaya, Ş., Kokkola, T., Emecen, G., Lappalainen, J., Kaarniranta, K. and Atalay, M. (2014). Natural thermal adaptation increases heat shock protein levels and decreases oxidative stress. *Redox Biol.* 3, 25–28.
[26] Ryckaert, J., Pasmans, F., Tobback, E., Duchateau, L., Decostere, A., Haesebrouck, F., Sorgeloos, P. and Bossier, P. (2010). Heat shock proteins protect platyfish (*Xiphophorus maculatus*) from *Yersinia ruckeri* induced mortality. *Fish Shellfish Immunol.* 28, 228–231.
[27] Reiser, S., Wuertz, S., Schroeder, J., Kloas, W. and Hanel, R. (2011). Risks of seawater ozonation in recirculation aquaculture – Effects of oxidative stress on animal welfare of juvenile turbot (*Psetta maxima*, L.). *Aquat. Toxicol.* 105, 508–517.
[28] Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods.* 9, 357–359.
[29] Yebra-Pimentel, E. S., Gebert, M., Jansen, H. J., Jong-Raadsen, S. and Dirks, R. P. H. (2019). Deep transcriptome analysis of the heat shock response in an embryo-derived Atlantic sturgeon (*Acipenser oxyrinchus*) cell line. *Fish Shellfish Immunol.* 88, 508-517.
[30] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 25, 2078–2079.
[31] Anders, S., Pyl, P. T. and Huber, W. (2015). HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics*. 31, 166–169.
[32] Anders, S. and Huber, W. 2010. Differential expression analysis for sequence count data. *Genome Biol*. 11, 106.

[33] Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A. and Narechania, A. (2003). PANTHER: A library of protein families and subfamilies indexed by function. *Genome Res.* 13, 2129–2141.
[34] Akbarzadeh, A., Günther, O. P., Houde, A. L., Li, S., Ming, T. J., Jeffries, K. M., Hinch, S. G. and Miller, K. M. (2018). Developing specific molecular biomarkers for thermal stress in salmonids. *BMC Genomics.* 19, 749.
[35] Wang, W., Deng, D. F., De Riu, N., Moniello, G. and Hung, S. S. O. (2013). Heat shock protein 70 (HSP70) responses in tissues of white sturgeon and green sturgeon exposed to different stressors. *N. Am. J. Aquac.* 75, 164–169.
[36] Werner, I., Linares-Casenave, J., Van Eenennaam, J. P. and Doroshov, S.I. (2007). The effect of temperature stress on development and heat-shock protein expression in larval green sturgeon (*Acipenser mirostris*). *Environ. Biol. Fishes.* 79, 191–200.
[37] Peng, G., Zhao, W., Shi, Z., Chen, H., Liu, Y., Wei, J. and Gao, F. (2016). Cloning HSP70 and HSP90 genes of kaluga (*Huso dauricus*) and the effects of temperature and salinity stress on their gene expression. *Cell Stress Chaperones*. 21, 349–359.
[38] Jesus, T. F., Grosso, A. R., Almeida-Val, V. M. F. and Coelho, M. M. (2016). Transcriptome profiling of two Iberian freshwater fish exposed to thermal stress. *J. Therm. Biol*. 55, 54–61.
[39] Salvador, J. M., Brown-clay, J. D. and Jr, A. J. F. (2013). Gadd45 in stress signaling, cell cycle control and apoptosis. *Exp. Med. Biol.* 793, 1–19
[40] Buckley, B. A. and Somero, G. N. (2009). CDNA microarray analysis reveals the capacity of the cold-adapted Antarctic fish *Trematomus bernacchii* to alter gene expression in response to heat stress. *Polar Biol.* 32, 403–415.
[41] Gantner, M. L., Hazen, B. C., Conkright, J. and Kralli, A. (2014). GADD45 regulates the thermogenic capacity of brown adipose tissue. *Proc. Natl. Acad. Sci*. 111, 11870–11875.
[42] Moskalev, A., Plyusnina, E., Shaposhnikov, M., Shilova, L., Kazachenok, A. and Zhavoronkov, A. (2012). The role of D-GADD45 in oxidative, thermal and genotoxic stress resistance. *Cell Cycle.* 11, 4222–4241.
[43] Nobeyama, Y. and Nakagawa, H. (2016). Aberrant DNA methylation in keratoacanthoma. *PLoS One.* 11, 1–14.
[44] Eliason, E. J., Clark, T. D., Hague, M. J., Hanson, L. M., Gallagher, Z. S., Jeffries, K. M., Gale, M. K., Patterson, D. A., Hinch, S. G. and Farrell, A. P. (2011). Differences in thermal tolerance among sockeye salmon populations. *Science*. 332, 109–112.
[45] Miller, K. M., English, K. K., Cooke, S. J., Patterson, D. A., Farrell, A. P., Donnison, L., Kaukinen, K. H., Hruska, K. A., Hinch, S. G., Ginther, N., et al. (2011). Genomic signatures predict migration and spawning failure in wild canadian Salmon. *Science.* 331, 214–217.
[46] Fuller, G. M. and Zhang, Z. (2010). Transcriptional Control Mechanism of Fibrinogen Gene Expression. *Ann. N. Y. Acad. Sci.* 936, 469–479.
[47] Cai, Q. and Storey, K. B. (1997). Freezing-induced genes in wood frog (*Rana sylvatica*): fibrinogen upregulation by freezing and dehydration. *Am. J. Physiol*. 272, 1480–1492.
[48] Hoang, Q. T., Cho, S. H., McDaniel, S. F., Ok, S. H., Quatrano, R. S. and Shin, J. S. (2009). An actinoporin plays a key role in water stress in the moss *Physcomitrella patens*. *New Phytol.* 184, 502–510.
[49] Wang, X., Liu, Y. and Yang, P. (2012). Proteomic studies of the abiotic stresses response in model moss – *Physcomitrella patens. Front. Plant Sci.* 3, 1–8.
[50] Prasad, S., Soldatenkov, V. A., Srinivasarao, G. and Dritschilo, A. (1998). Identification of keratins 18, 19 and heat-shock protein 90β as candidate substrates of

proteolysis during ionizing radiation-induced apoptosis of estrogen-receptor negative breast tumor cells. *Int. J. Oncol.* 13, 757–764.
[51] Dumartin, L., Alrawashdeh, W., Trabulo, S. M., Radon, T. P., Steiger, K., Feakins, R. M., Di Magliano, M. P., Heeschen, C., Esposito, I., Lemoine, N. R., et al. (2017). ER stress protein AGR2 precedes and is involved in the regulation of pancreatic cancer initiation. *Oncogene*. 36, 3094–3103.
[52] Loyau, T., Hennequet-Antier, C., Coustham, V., Berri, C., Leduc, M., Crochet, S., Sannier, M., Duclos, M. J., Mignon-Grasteau, S., Tesseraud, S., et al. (2016). Thermal manipulation of the chicken embryo triggers differential gene expression in response to a later heat challenge. *BMC Genomics.* 17, 1–15
[53] Hori, T. S., Gamperl, A. K., Afonso, L. O. B., Johnson, S. C., Hubert, S., Kimball, J., Bowman, S. and Rise, M. L. (2010). Heat-shock responsive genes identified and validated in Atlantic cod (*Gadus morhua*) liver, head kidney and skeletal muscle using genomic techniques. *BMC Genomics*. 11, 72.
[54] Podrabsky, J. E. and Somero, G. N. (2004). Changes in gene expression associated with acclimation to constant temperatures and fluctuating daily temperatures in an annual killifish *Austrofundulus limnaeus*. *J. Exp. Biol*. 207, 2237–2254.
[55] Qu, H. and Ajuwon, K. M. (2018). Cytosolic phosphoenolpyruvate carboxykinase is a response gene involved in porcine adaptation to heat stress. *J. Anim. Sci.* 96, 1724–1735.
[56] Choi, C. Y., Min, B. H., Jo, P. G. and Chang, Y. J. (2007). Molecular cloning of PEPCK and stress response of black porgy (*Acanthopagrus schlegeli*) to increased temperature in freshwater and seawater. *Gen. Comp. Endocrinol.* 152, 47–53.
[57] Gusarova, V., Caplan, A. J., Brodsky, J. L. and Fisher, E. A. (2001). Apoprotein B Degradation is promoted by the molecular chaperones hsp90 and hsp70. *J. Biol. Chem.* 276, 24891–24900.
[58] Balogh, G., Péter, M., Glatz, A., Gombos, I., Török, Z., Horváth, I., Harwood, J. L. and Vígh, L. (2013). Key role of lipids in heat stress management. *FEBS Lett.* 587, 1970–1980.
[59] Pan, J., Tristram-Nagle, S., Kučerka, N. and Nagle, J.F. (2008). Temperature dependence of structure, bending rigidity, and bilayer interactions of dioleoylphosphatidylcholine bilayers. *Biophys. J.* 94, 117–124.
[60] Mellery, J., Geay, F., Tocher, D. R., Kestemont, P., Debier, C., Rollin, X. and Larondelle, Y. (2016). Temperature increase negatively affects the fatty acid bioconversion capacity of rainbow trout (*Oncorhynchus mykiss*) fed a linseed oil-based diet. *PLoS One*. 11, 1–24.
[61] Linares-Casenave, J., Werner, I., Van Eenennaam, J. P. and Doroshov, S. I. (2013). Temperature stress induces notochord abnormalities and heat shock proteins expression in larval green sturgeon (*Acipenser medirostris* Ayres 1854). *J. Appl. Ichthyol*. 29, 958–967.

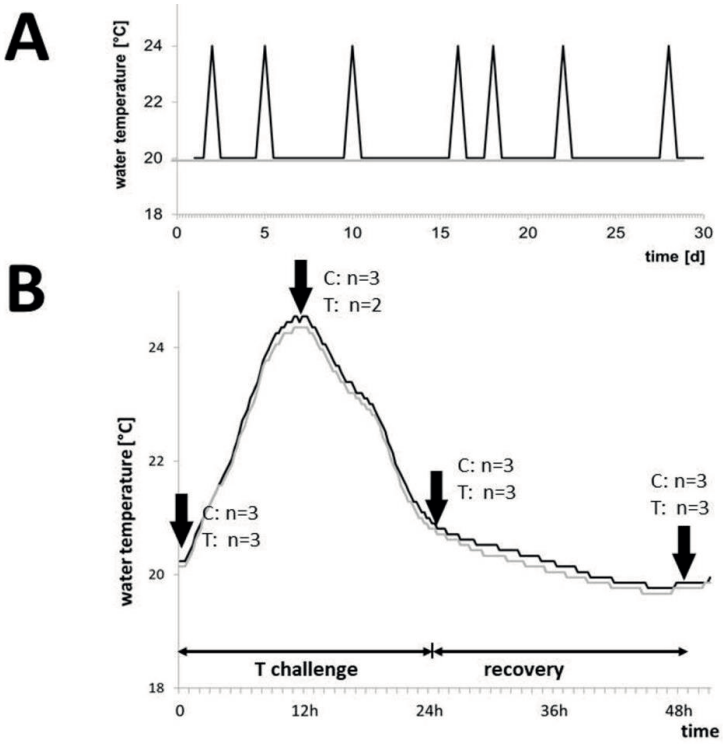


Fig. 1. Experimental setup involving A) a 30-day (d) training period followed B) by a heat shock. The training period **(A)** comprised seven temperature challenges in trained fish (T) and rearing at a constant temperature of 20°C in control fish (C). The heat shock **(B)** (data from PRTemp101 data loggers) involved an increase in the water temperature from 20ºC to 24ºC in 12 hours (h) and a relaxation with a temperature decrease to 20°C. The subsequent recovery phase was 24 h long. Samples were taken at 0 h, 12 h, 24 h and 48 h as indicated by the arrows.

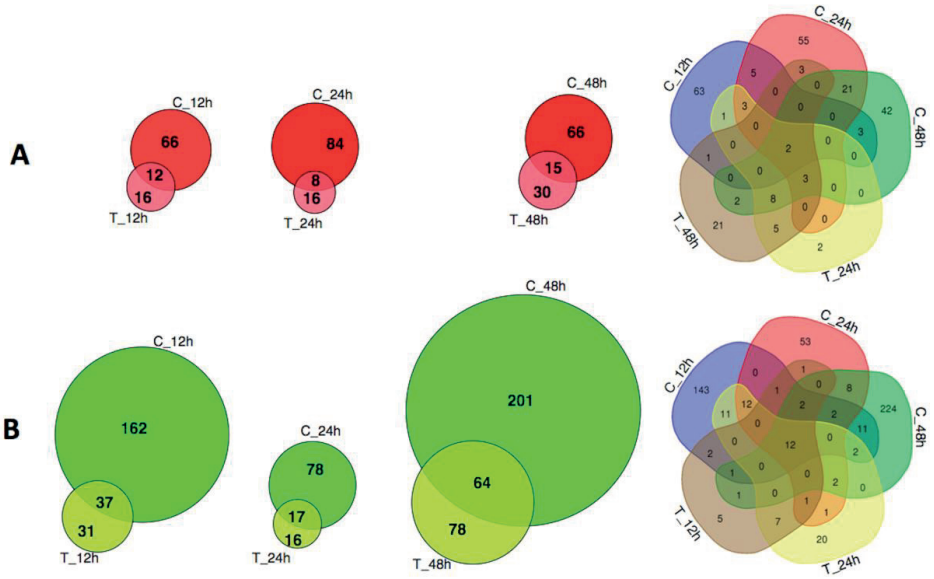
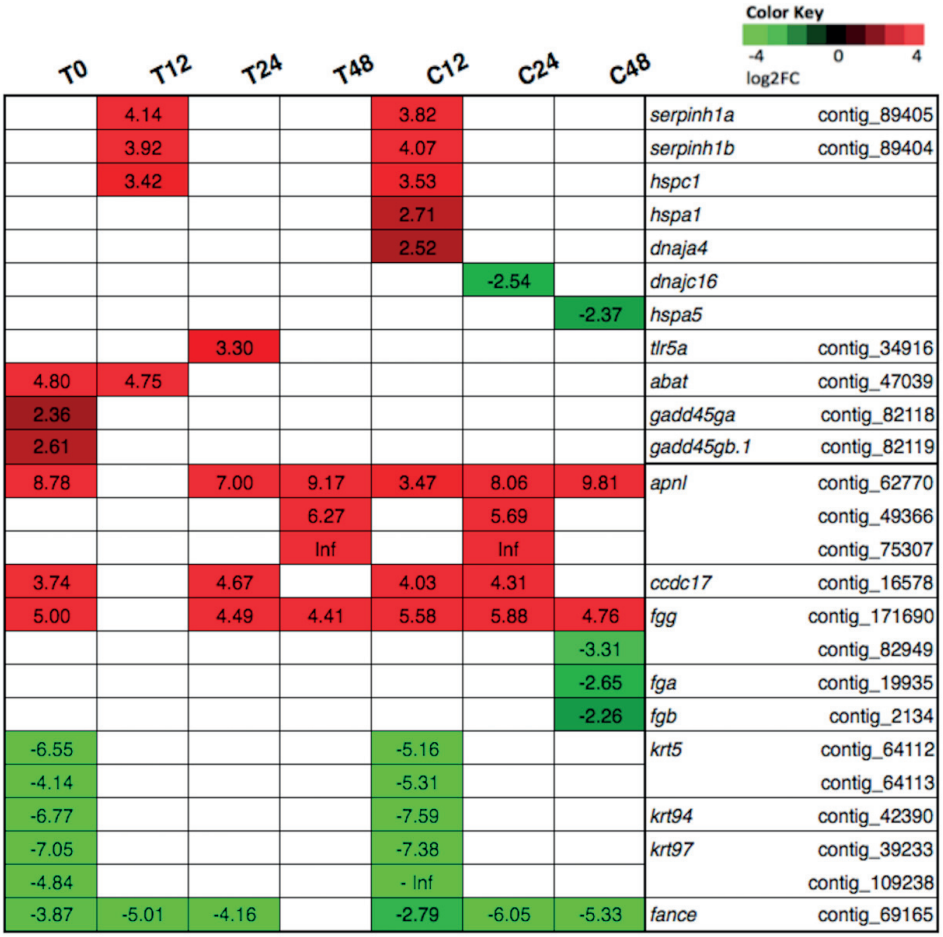
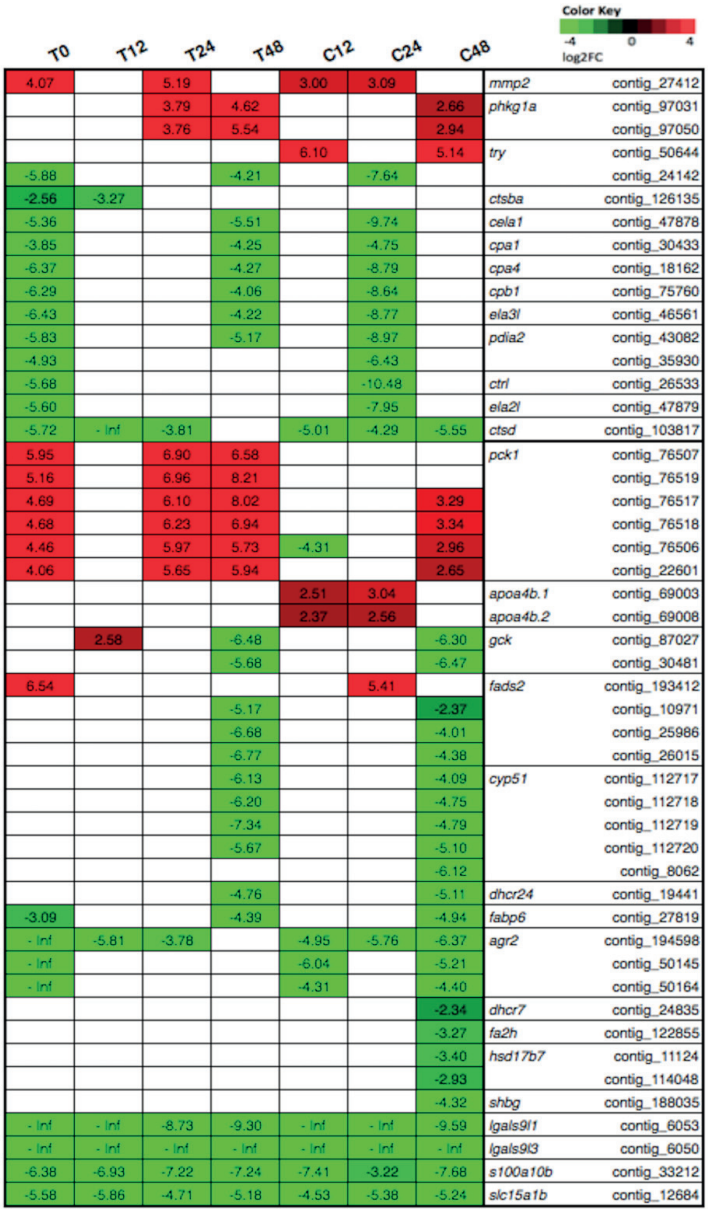


Fig. 2. Venn Diagrams showing DECs and overlaps between experimental groups at different timepoints. A: Overlaps of upregulated contigs in the control (C) and trained (T) fish during heat shock peak (12h), heat shock relaxation (24h) and recovery (48h) and throughout the experiment, for which T_12h showed no overlap and was therefore not represented. **B:** Overlaps of downregulated contigs in the control (C) and trained (T) fish during heat shock peak (12h), heat shock relaxation (24h) and recovery (48h) and throughout the experiment, for which T_48h showed no overlap and was therefore not represented.



T0	T12	T24	T48	C12	C24	C48		
	4.14			3.82			*serpinh1a*	contig_89405
	3.92			4.07			*serpinh1b*	contig_89404
	3.42			3.53			*hspc1*	
				2.71			*hspa1*	
				2.52			*dnaja4*	
					-2.54		*dnajc16*	
						-2.37	*hspa5*	
		3.30					*tlr5a*	contig_34916
4.80	4.75						*abat*	contig_47039
2.36							*gadd45ga*	contig_82118
2.61							*gadd45gb.1*	contig_82119
8.78		7.00	9.17	3.47	8.06	9.81	*apnl*	contig_62770
			6.27		5.69			contig_49366
			Inf		Inf			contig_75307
3.74		4.67		4.03	4.31		*ccdc17*	contig_16578
5.00		4.49	4.41	5.58	5.88	4.76	*fgg*	contig_171690
						-3.31		contig_82949
						-2.65	*fga*	contig_19935
						-2.26	*fgb*	contig_2134
-6.55				-5.16			*krt5*	contig_64112
-4.14				-5.31				contig_64113
-6.77				-7.59			*krt94*	contig_42390
-7.05				-7.38			*krt97*	contig_39233
-4.84				- Inf				contig_109238
-3.87	-5.01	-4.16		-2.79	-6.05	-5.33	*fance*	contig_69165

Fig. 3. Heat map of selected genes related to heat shock response, cell cycle and tissue damage. T and C denote trained and control groups, respectively. Color key indicates log2 fold change (FC) values. Contigs numbers and corresponding gene symbols of the best BLAST hit are provided for all genes.



T0	T12	T24	T48	C12	C24	C48		
4.07		5.19		3.00	3.09		*mmp2*	contig_27412
		3.79	4.62			2.66	*phkg1a*	contig_97031
		3.76	5.54			2.94		contig_97050
				6.10		5.14	*try*	contig_50644
-5.88			-4.21		-7.64			contig_24142
-2.56	-3.27						*ctsba*	contig_126135
-5.36			-5.51		-9.74		*cela1*	contig_47878
-3.85			-4.25		-4.75		*cpa1*	contig_30433
-6.37			-4.27		-8.79		*cpa4*	contig_18162
-6.29			-4.06		-8.64		*cpb1*	contig_75760
-6.43			-4.22		-8.77		*ela3l*	contig_46561
-5.83			-5.17		-8.97		*pdia2*	contig_43082
-4.93					-6.43			contig_35930
-5.68					-10.48		*ctrl*	contig_26533
-5.60					-7.95		*ela2l*	contig_47879
-5.72	- Inf	-3.81		-5.01	-4.29	-5.55	*ctsd*	contig_103817
5.95		6.90	6.58				*pck1*	contig_76507
5.16		6.96	8.21					contig_76519
4.69		6.10	8.02			3.29		contig_76517
4.68		6.23	6.94			3.34		contig_76518
4.46		5.97	5.73	-4.31		2.96		contig_76506
4.06		5.65	5.94			2.65		contig_22601
				2.51	3.04		*apoa4b.1*	contig_69003
				2.37	2.56		*apoa4b.2*	contig_69008
	2.58		-6.48			-6.30	*gck*	contig_87027
			-5.68			-6.47		contig_30481
6.54					5.41		*fads2*	contig_193412
			-5.17			-2.37		contig_10971
			-6.68			-4.01		contig_25986
			-6.77			-4.38		contig_26015
			-6.13			-4.09	*cyp51*	contig_112717
			-6.20			-4.75		contig_112718
			-7.34			-4.79		contig_112719
			-5.67			-5.10		contig_112720
						-6.12		contig_8062
			-4.76			-5.11	*dhcr24*	contig_19441
-3.09			-4.39			-4.94	*fabp6*	contig_27819
- Inf	-5.81	-3.78		-4.95	-5.76	-6.37	*agr2*	contig_194598
- Inf				-6.04		-5.21		contig_50145
- Inf				-4.31		-4.40		contig_50164
						-2.34	*dhcr7*	contig_24835
						-3.27	*fa2h*	contig_122855
						-3.40	*hsd17b7*	contig_11124
						-2.93		contig_114048
						-4.32	*shbg*	contig_188035
- Inf	- Inf	-8.73	-9.30	- Inf	- Inf	-9.59	*lgals9l1*	contig_6053
- Inf	- Inf	- Inf	- Inf	- Inf	- Inf	- Inf	*lgals9l3*	contig_6050
-6.38	-6.93	-7.22	-7.24	-7.41	-3.22	-7.68	*s100a10b*	contig_33212
-5.58	-5.86	-4.71	-5.18	-4.53	-5.38	-5.24	*slc15a1b*	contig_12684

Fig. 4. Heat map of selected genes related to proteolysis and metabolism. T and C denote trained and control groups, respectively. Color key indicates log2 fold change (FC) values. Contigs numbers and corresponding gene symbols of the best BLAST hit are provided for all genes.

Temperature training improves transcriptional homeostasis after heat shock in juvenile Atlantic sturgeon (*Acipenser oxyrinchus*)

Temperature training improves transcriptional homeostasis after heat shock in juvenile Atlantic sturgeon (*Acipenser oxyrinchus*)

Elena Santidrián Yebra-Pimentel[a,b,*], Bruno Reis[c], Jörn Gessner[d], Sven Wuertz[d], and Ron P.H. Dirks[a].

Supplementary Material

Supplementary tables are available at:
https://www.dropbox.com/s/th8y9tv748o0dvf/Article2_supplementary_Tables.zip?dl=0

Supplementary Tables:

- **Table S1.** Sample list, corresponding number of sequencing reads per sample, accession and number and normalized (RPKM, reads per kilobase and million) reads mapped to the reference transcriptome, including only contigs with more than 0 RPKM in at least one sample.
- **Table S2**. DESeq results (padj <0.05) per comparison. **T:** trained fish; **C:** control fish; 0, 12, 24 and 48 stand for the timepoints sampled during heat shock. **Contig ID**: contig number in the reference transcriptome; **BaseMean**: Mean of mapped reads (RPKM) in all samples used in the comparison; **BaseMean A:** Mean of mapped reads in the control group (C0, control before temperature challenge); **BaseMean B:** Mean of mapped reads in the compared group; **Fold Change:** Mean of mapped reads in the comparison group divided by the mean of mapped reads in the control group; **Log2 Fold Change:** log2 of the Fold Change. Positive values indicate upregulation and negative values downregulation; **pval:** probability value; **padj:** probability value adjusted for multiple testing with the Benjamini-Hochberg procedure; **Lowest BLASTx E-value:** BLAST expectation value; **BLASTx Danio rerio Accession protein/ Accession gene/ gene symbol/ gene description:** ENSEMBL identifier for the best protein hit in the zebrafish

proteome; **BLASTx Danio rerio Accession gene:** ENSEMBL protein identifier/ ENSEMBL gene identifier/ gene name/ gene description of the best BLASTx hit in the zebrafish proteome.

- **Table S3.** List of DECs in at least one condition, including query (contig) size in nucleotides, log2FoldChange and padj value in each of the comparisons and best BLASTx zebrafish hit.
- **Table S4.** List of under and overrepresented gene ontology (GO) terms per condition. **D. rerio REFLIST:** the number of genes contained in each GO term in the reference list; **Upload:** number of differentially expressed genes contained in the GO term; **Expected:** expected number of genes include in the GO category; **Over/Under:** Over or underrepresentation of the GO term. Negative (green) values indicate underrepresentation, while positive (red) values indicate overrepresentation; **Fold Enrichment:** Number of observed genes in each category divided by the number of expected genes; **P-value:** probability value**; FDR:** False discovery rate.

III

The Atlantic sturgeon (*Acipenser oxyrinchus*) genome provides insights into genome evolution in paleotetraploids

Elena Santidrián Yebra-Pimentel,[*,1,2] Aurora Campo,[3] Sahana Shivaramu,[4] Christiaan V. Henkel,[2] Hans J. Jansen,[1] Susanne A. Jong-Raadsen, [1] and Ron P. Dirks[1]

1 - ZF-screens B.V., J. H. Oortweg 19, 2333 CH Leiden, The Netherlands.

2 - Norwegian University of Life Sciences, Faculty of Veterinary Medicine, Department of Basic Science and Aquatic Medicine, Oslo, Norway.

3 - Laboratory of Biology of Aquatic Organisms and Ecosystems (BOREA), UMR CNRS 7208-IRD207-UPMC, Muséum National d'Histoire Naturelle, 7 rue Cuvier, CP32 75231 Paris Cedex 05, France.

4 - South Bohemian Research Center for Aquaculture and Biodiversity of Hydrocenoses, Faculty of Fisheries and Protection of Waters. University of South Bohemia in České Budējovice, Vodňany, Czech Republic

[*] **Author for Correspondence:** Elena Santidrián Yebra-Pimentel, ZF-Screens B.V. Leiden, The Netherlands. elena@zfscreens.com

Abstract

The availability of a reference genome improves research on many fields; however, genomic resources are lacking for Atlantic sturgeon, which is under a restoration program in Europe. Sturgeons have high chromosome counts resulting from several lineage-specific whole-genome duplication events (WGD), but their ploidy status is still controversial. We have *de novo* assembled the Atlantic sturgeon genome sequence using Illumina and nanopore sequencing technologies. Ploidy level, genome size and heterozygosity were estimated using microsatellite loci analysis, flow cytometry and k-mer counting. Additionally, Hox gene clusters were analyzed to infer the evolutionary position of Atlantic sturgeon among gnathostomata and their expression was assessed in 4 stages of embryonic development. Microsatellite analysis showed disomic patterns in 5 loci and tetrasomy in 3 loci. Flow cytometry results suggested a genome size of 1.1 Gb, while k-mer counting inferred 652 Mb and high heterozygosity (5.36%). However, our nanopore assembly size was 1.28 Gb (N_{50} 0.97 Mb). We found 76 Hox genes and 1 pseudogene distributed over 7 clusters, with one of the HoxC clusters being lost. Hox paralogs shared an average identity of 92% at the protein level, while synteny was not fully conserved between the HoxA paralogs. Phylogenetic reconstruction indicated that they were resulting from a sturgeon-specific WGD (SR), independent from paddlefish and common for the two analyzed sturgeon species. The developed genomic resources serve as a basis for new research in Atlantic sturgeon, which is likely a paleotetraploid species in the process of rediploidization,

based on the found paralogs, microsatellite results and discrepancy between estimated genome size and total assembly size.

Keywords: Hox, genome, sequencing, sturgeon, nanopore, genome duplication, tetraploidization, rediploidization.

Introduction

Sturgeons (order *Acipenseriformes*) date back to ~ 350 Mya (Hughes et al. 2018) and occupy an important position in evolution as a basal lineage of ray-finned fish (*Actinopterygii*) that diverged from teleosts more than 250 Mya (Betancur et al. 2013). The order contains two sister families that diverged ~204 Mya (Peng et al. 2007; Luo et al. 2019): paddlefish (*Polyodontidae)* and sturgeons *(Acipenseridae)*. The first one contains only 2 extant species (*P. spathula* and *P. gladius*), while the second contains 25 species divided into 4 genera and 25 species, distributed in 3 clades: the basal Sea clade, containing only the Atlantic and European sturgeons, and the Atlantic and Pacific clade.

Seven out of the eight sturgeon species that have traditionally inhabited the European rivers and estuaries are currently considered critically endangered (Freyhof and Brooks, 2011). The situation of the Atlantic sturgeon (*Acipenser oxyrinchus*) in Europe is particularly alarming as it is extinct in the wild since the depletion of the Baltic population (Ludwig et al. 2002; Ludwig 2008). The remaining Atlantic sturgeon populations, distributed from Florida (US) to Labrador (Canada) and categorized as "near threatened" (Pierre and Parauka, 2006), are currently being used to repopulate the Baltic (Gessner and Bartel 2000).

Today, 318 fish genomes have been deposited at the NCBI database, most of them corresponding to teleosts, but also some chondrichthyans like Little Skate (*Leucoraja erinacea*) and Elephant shark (*Callorhinchus miii*) and the holostean Spotted gar (*Lepisosteus oculatus*) (King et al. 2011; Venkatesh et al. 2014; Braasch et al. 2016). Good genomic resources have a strong impact on research in multiple fields including molecular biology, physiology, ecology and evolution. Regarding Atlantic sturgeon, transcriptome resources consist of an assembly including 4 stages of embryonic development, a multiorgan assembly originating from 20 juvenile tissues, an embryo-derived cell line assembly, and a cell-organ-embryo (CEO) assembly combining all the aforementioned resources (Grunow et al. 2011; Kaitetzidou et al. 2017; Yebra-Pimentel et al. 2019). The only assembled sturgeon genome, corresponding to sterlet (*Acipenser ruthenus*), was recently deposited at NCBI (GCA_000209225.1). However, it is not associated with any publication and the assembly, performed solely with Illumina reads, is quite fragmented (1.83 Gb genome divided over 215,913 scaffolds, with an N_{50} of 191.06 kb).

Acipenseriformes have an extraordinary tendency for genome duplication and polyploidization and have undergone several independent lineage-specific duplication events (Crow et al. 2012; Symonová et al. 2017; Havelka et al. 2016). While the American paddlefish (*Polyodon spathula*) has a diploid chromosome number or 120, sturgeons have variable chromosome numbers within the family and ploidy status is still controversial and remains unresolved. Recent investigations suggest two ploidy scales: the "evolutionary scale" and the "functional scale". While the first suggests tetraploid (4n), octaploid (8n) and

dodecaploid (12n) relationships among species of ~ 120 (e.g. Atlantic sturgeon), ~ 250 and ~ 500 chromosomes respectively, the latter scale presumes diploid (2n), tetraploid (4n) and hexaploid (6n) relationships (Vasil'ev et al. 2010). In addition, triploid individuals are sporadically found within diploid species (Blacklidge and Bidwell 1993). Previous studies using microsatellite markers, direct chromosome visualization, Feulgen Image Analysis (FIA) densitometry and flow cytometry categorized the Atlantic sturgeon as a diploid species (Ludwig et al. 2001; Havelka et al. 2013) with 2n=120 chromosomes (Tagliavini et al. 1999; Fontana et al. 2008) and a genome size of 2.14 Gb (Blacklidge and Bidwell 1993; Hardie and Hebert 2003).

Evidence for a teleost-specific whole-genome duplication (WGD), known as the TSGD or 3R, was initially revealed by the presence of duplicated Hox genes (Amores 1998; Prince et al. 1998; Bruce et al. 2001), which are conserved homeodomain-containing transcription factors related with developmental patterning organized in genomic clusters that exhibit colinearity between their position and their temporal and spatial expression (McGinnis and Krumlauf 1992; Duboule 2007). Invertebrates contain a single Hox cluster with a variable number of genes, from only 6 in the worm *C. elegans* to a total of 14 in the cephalochordate *B. floridae* (Rosa et al. 1999; Minguillón et al. 2005). As a result of the two rounds of WGD events (2R) affecting vertebrates that occurred 600 Mya (Meyer and Schartl 1999), most vertebrates, including cartilaginous fishes (sharks), sarcopterygians (coelacanth, tetrapods) and holosteans (Spotted gar), typically contain 4 Hox paralog clusters (named A, B, C and D), with few exceptions, like the Little Skate which has lost the complete HoxC cluster (Venter et al. 2001; Martinez and Amemiya 2002; Amemiya et al. 2013; Venkatesh et al. 2014; Braasch et al. 2016).

As a result of the TSGD, some teleosts like the European eel (*Anguilla anguilla*) contain up to 8 Hox clusters (termed a and b) with 73 genes (Henkel et al. 2012a), while others, like zebrafish and medaka, have lost one or more clusters (Lynchu and Conery 2000; Kondrashov et al. 2002; Kuraku and Meyer 2009). Additional lineage-specific duplication events have been described in several teleost families, among them the salmonid-specific WGD (Ss4R) around ~80 Mya (Macqueen and Johnston 2014; Lien et al. 2016) and the more recent duplication in common carp (*Cyprinus carpio*), dated from ~ 12 Mya (David et al. 2003; Yuan et al. 2010; Henkel et al. 2012b; Wang et al. 2012). The American paddlefish, sturgeons closest relative, has two HoxA and Hox D clusters (termed α and β) derived from a paddlefish-specific WGD (termed as PR), that occurred ~42 Mya (Crow et al. 2012); however, the full genome sequence of the American paddlefish is not available and the HoxB and HoxC gene complement is unknown. Although Crow et al. suggested that the PR duplication seemed to have occurred after the divergence from sturgeons, the lack of genomic resources for sturgeons made it impossible to confirm this hypothesis.

The aim of this work was to generate genomic resources, determine the ploidy status and infer the phylogenetic relationships of the endangered Atlantic sturgeon among jawed vertebrates. To this aim, we have sequenced the Atlantic sturgeon genome using Illumina and nanopore technologies, resulting in a highly contiguous and complete genome assembly. Hox-coding genes were annotated and used to perform phylogeny and synteny analyses in order to infer the molecular evolution of sturgeons among gnathostomata.

Additionally, transcriptomic reads derived from 4 stages of embryonic development were mapped to the annotated Hox genes in order to confirm their transcriptional activity and evaluate their expression levels (Kaitetzidou et al. 2017).

Materials and methods

Sturgeon samples

A juvenile Atlantic sturgeon female (7.7 kg, 98 cm) was provided by Fischzucht Rhönforelle GmbH in Gersfeld (Germany) on July 3rd, 2015. The specimen was euthanized and samples were taken in agreement with standardized methods for fish processing at a licensed and registered processing unit (Fischzucht Rhönforelle GmbH). Blood samples were taken from the gills using heparin-flushed needles and immediately snap-frozen in liquid nitrogen. In addition, 20 tissue samples were taken through dissection, as described previously (Yebra-Pimentel et al. 2019).

Illumina and nanopore sequencing

For the Illumina sequencing, DNA was extracted from snap-frozen blood using the DNeasy Blood & Tissue kit (Qiagen GmbH, Hilden, Germany). DNA concentration was measured using fluorometric quantitation (Qubit, Thermo Fisher Scientific) and subsequently sheared using a nebulizer (Life Technologies). Paired-end libraries were prepared from 5 μg of sheared DNA using the Paired-End Sequencing Sample Prep kit according to the manufacturer's description (Illumina). For the size selection step, the 400 base pair (bp) band was cut from the agarose gel, purified and amplified by 10 PCR cycles. The resulting library was analyzed using a Bioanalyzer 2100 DNA 1000 series II chip and sequenced using an Illumina HiSeq2500 instrument with a paired-end read length of 151 nucleotides (nt). Image analysis and base calling were done by the Illumina pipeline. The quality of the raw Illumina reads was inspected using FastQC software (Andrews 2010) and sequencing adapters were removed using BBmap (BBMap - Bushnell B.)

For the nanopore sequencing, high molecular weight (HMW) DNA (>60 kb) was isolated from Atlantic sturgeon liver, blood, spleen, kidney and heart using the Qiagen Genomic-Tip 100/G (Qiagen). DNA concentration was measured with fluorometric quantification while length was estimated through automatic electrophoresis (4200 TapeStation System, Agilent). The isolated DNA was used either intact or sheared to a size of ~20Kb with a g-Tube (Covaris) to prepare 27 nanopore libraries that were run in either MinION (n=24) or PromethION (n=3) Oxford nanopore technologies (ONT) sequencing devices. Nanopore libraries were prepared according to the ONT protocols using either 1D-ligation (versions LSK108 and LSK109) or rapid protocol (versions RAD002 and RAD004). The MinKNOW software was used to control the sequencing process and the read files were base called with either albacore (versions 2.0.2 and 2.1.3) or guppy. Nanopore sequencing adapters were removed from the raw data using Porechop (version 0.2.1. http://github.com/rrwick/Porechop). Reads were filtered to a minimum average read quality Phred score of 7 using Nanofilt (version 2.2.0) and, after filtering, the quality and length were inspected with NanoPlot (De Coster et al. 2018).

Atlantic sturgeon genome assembly

The genome assembly was performed according to the strategy depicted in **Figure 1**. Initially, sequencing reads from paired-end Illumina libraries were used to build an Illumina assembly using the De Bruijn graph-based *de novo* assembler implemented in the CLC Genomics Workbench version 4.4.1 (CLC bio, Aarhus, Denmark). A total of 8 *de novo* genome assembly rounds were performed using different k-mer (21,31,51 and 63) and bubble size (50, 240, 300 and 400bp) settings combinations, in order to tune the setting for the optimal contiguity (highest N50) and completeness (largest assembly size).
Filtered nanopore reads were assembled using the Tulipa-julia assembly software, an improved version of the TULIP assembler(Jansen et al. 2017), which applies the overlap layout assembly strategy with a heuristic approach that results in increased speed of the assembly process: instead of finding overlaps among error-prone reads by aligning all versus all reads, reads are aligned to unique sparse sequences of the genome, called seeds. The seeds were obtained by fragmenting the Illumina contigs into DNA sequences of 1.5 kb in length. Filtered nanopore reads were aligned to the seeds using Minimap2 and the alignment file was thereafter used by the Tulipa-julia assembler to produce an assembly graph, where the seeds were represented as nodes linked by the nanopore reads aligning to them. After removing ambiguities in the assembly graph, the assembler produced contiguous sequences, which still needed post-assembly correction. The assembled contigs were corrected with 4 iterations of Racon (Vaser et al. 2017), followed by 2 iterations of Pilon (Walker et al. 2014). Genome contiguity was evaluated based on N_{50} and number of contigs, while the genome completeness was assessed with Benchmarking Universal Single-Copy Orthologs using BUSCO (version 3) against the vertebrate odb9 dataset (Waterhouse et al. 2018).

Estimation of ploidy, genome size and heterozygosity

Previously described microsatellite markers were amplified and analyzed as described in Havelka et al (2013) in order to assess the ploidy level of the specimen used for sequencing (Havelka et al. 2013). Initially 13 microsatellites were tested for amplification in Atlantic sturgeon DNA and, from all together 11 successfully amplified markers, 8 were chosen for subsequent analysis based on their level of polymorphism. The allele numbers of AciG35 (Börk et al. 2008), AfuG135 (Welsh et al., 2003), Aox45 (King et al., 2001), Atr117, Atr109 (Rodzen and May 2002), Spl101, Spl163 and Spl173 (McQuown et al. 2000) were assessed in order to infer the ploidy level.
Additionally, Atlantic sturgeon red blood cells (RBC) were processed for flow cytometry analysis in order to estimate the genome size. European eel (*Anguilla anguilla*) RBC were processed simultaneously and used as reference (genome size 1.1Gb, haploid) (Henkel et al. 2012a). Snap-frozen blood was diluted 1:50 in cold PBS and centrifuged at 2,500g for 5 minutes at 4ºC. The cellular pellet was suspended again in PBS and centrifuged at the same conditions. Cells were fixed in 70% cold ethanol for 30 min, washed with PBS and filtered through a 40 µm filter. RNA was digested with RNase A (0.5 µg/mL) followed by

propidium iodide (PI) staining (50 μg/mL). Flow cytometry analysis was performed up to a minimum of 4,000 events (cells) per sample and the median fluorescence (MF) was calculated. The Atlantic sturgeon genome size (GS) was calculated by [($GS_{European\ eel}/MF_{European\ eel}$) x $MF_{Atlantic\ sturgeon}$].

We have also used Jellyfish (version 2.2.6) (Marçais and Kingsford 2011) to count k-mers in Illumina reads after removal of low-quality bases (< 30) and trimming to a uniform length of 100 nt, in order to estimate the haploid genome size, repetitive content and heterozygosity. K-mer frequency histograms were obtained counting 19, 21, 23 and 25-mers in the filtered Illumina reads and the resulting histograms were analyzed using the algorithm applied by GenomeScope (Vurture et al. 2017).

In order to investigate the contig coverage distribution in the nanopore genome assembly, a subset of Illumina genomic reads was aligned to the assembled contigs using BWA-MEM (version 0.7.15) (Li and Durbin, 2009). The number of alignments per contig was counted and normalized to the contig length in order to estimate the coverage distribution in the final genome.

Evolutionary studies on developmental genes

Hox amino acid sequences from key vertebrate species, namely spotted gar, European eel, zebrafish and paddlefish, were retrieved from the Ensembl release 91 and NCBI databases, blasted (tblastn) against the Atlantic sturgeon genome and manually annotated using the CLC Main Workbench 7 software (QIAGEN Bioinformatics). The annotated Hox coding sequences (**Table S4**) were blasted against the embryonic and multiorgan assemblies in order to confirm the presence of open reading frames (ORF) in the transcriptomic resources.

In order to confirm the transcriptional activity of the annotated Hox genes and analyze their expression, transcriptomic reads derived from 4 stages of embryonic development (mid-gastrula, end of gastrulation, hatching, and 2 days post-hatching) (Kaitetzidou et al. 2017) were mapped to the Hox coding sequences (CDS) using bowtie2 (version 2.2.5.), and secondary alignments were filtered with SAMtools (version 1.2). Mapped reads were counted using the python package HTSeq (version 0.5.3p9), normalized, and expressed as reads per kilobase per million reads (RPKM).

In order to reconstruct the phylogeny of Atlantic sturgeon Hox clusters, the protein sequences of several vertebrate representatives were retrieved from the NCBI database, including two basal chondrichtians (hornshark and ghostshark), two sarcopterygians (human and coelacanth), two teleosts (European eel and zebrafish), and four non-teleost ray-finned fish: spotted gar, American paddlefish (sturgeon's sister clade), Sterlet and Atlantic sturgeon. For each of the species, protein sequences were concatenated per cluster and subsequently concatenated superalignments were performed using CLUSTAL included in SeaView (version 4.6.1), followed by manual curation. Maximum likelihood phylogenetic tree for each Hox clusters were built in the online platform CIPRES (www.phylo.org) using RAxML black-box (Stamatakis 2014) with 1,000 bootstrap replicates and JTT substitution matrix. Ghostshark was used as outgroup for all the clusters, and Hornshark was additionally used for HoxB and HoxC clusters. The protein sequences used in the tree are

described in the **Table S4**. The resulting phylogenetic tree was displayed using Figtree (version 1.4.3).

Finally, synteny analysis was performed on the HoxA cluster paralogs and on the HoxC cluster, using the Spotted gar as reference, and including ghostshark, human, coelacanth, zebrafish and medaka. The genomic regions of the Atlantic sturgeon Hox clusters were blasted against the non-redundant protein sequence database and manually analyzed with the CLC Main Workbench 7 software on the assembled genome in order to identify neighboring genes.

Results

Illumina and nanopore sequencing

Using the Illumina HiSeq2500 sequencer we have produced 548,599,202 paired reads of 151 nt (82.2 Gb) with a mean quality score of 35.6. Assuming a genome size of 1.28 Gb, this corresponded to ~64-fold coverage of the genome.

Using ONT devices and filtering out reads with a mean quality score below 7, we have obtained 58.26 Gb (~45.5-fold coverage) of nanopore sequencing data distributed over 7,860,939 sequencing reads with a mean quality score of 10.5 and a read N_{50} of 13.13 kb (**Figure S1**).

Genome assembly, correction and annotation

The Illumina reads were used for *de novo* assembly of a first draft genome sequence of Atlantic sturgeon. Multiple k-mer and bubble size setting were tested. The best assembly resulted from k-mer size 63 and bubble size 250 bp. The Illumina genome assembly had a length of 1.37 Gbp divided over 978,147 contigs with an N50 of 3.7 kb (**Figure S2**).

In order to obtain seeds for the nanopore genome assembly, Illumina contigs were split into fragments of 1.5 kb in length, yielding 581,096 high quality seeds (0.68-fold coverage of the estimated genome size). Filtered nanopore reads were aligned to the Illumina seeds and thereafter Tulipa-julia was used to generate the assembly graph. The resulting assembly contained 574,518 seeds (750.78 Mb) and 525.9 Mb links derived from nanopore sequencing reads, resulting in an assembly length of 1.26 Gb distributed over 4,979 contigs with an N_{50} of 0.93 Mb. The assembly was finally corrected with 4 iterations of Racon (version 1.3.2) and 2 iterations of Pilon (version 1.23), resulting in an assembly length of 1.28 Gb distributed over 4,947 contigs with an N_{50} of 0.97 Mb (**Figure 2**). The BUSCO analysis could retrieve 2,401 complete BUSCOS out of 2,586 (92.8%), of which 76% were single-copy (**Figure S3**).

Estimation of ploidy, genome size and heterozygosity

The analysis of eight microsatellite markers showed disomic allelic patterns at Aox45, Atr109, Spl173, Spl163 and AfuG135 loci whereas tetrasomic allelic patterns were observed at Atr117, Spl101 and AciG35 loci.

Flow cytometry results showed a median BLUE 695/40-A fluorescence of 12,162 and 12,135 for Atlantic sturgeon and European eel respectively; meaning the genome size of the Atlantic sturgeon specimen was ~1.1 Gbp (haploid) (**Figure S4**). However, k-mer counting of filtered Illumina reads estimated a smaller haploid genome size. Out of the 4 k-mers tested, the 19-mer frequency histogram had the best fit to the GenomeScope model (97.4%). The GenomeScope profile (**Figure 3**) showed a high error peak (orange line) at coverage ~12.5, a very high heterozygous k-mer peak at coverage 25 and a lower homozygous k-mer peak at coverage 50. Therefore the model estimated a ~653 Mb genome size (haploid), extremely high heterozygosity (5.36%) and high percentage of repetitive content (~26%). A subset of Illumina reads was also aligned to the nanopore assembly in order to investigate the coverage distribution of the nanopore assembly. Out of the 407,913,678 reads that were used to align to the nanopore assembly, 407,908,728 aligned to it (99.99%), corresponding to 61.59 Gb of data in sequencing reads. The average contig coverage was 48.1-fold, which would correspond to a genome size of 1.28 Gb. However, the contig coverage distribution was skewed to the right and showed a peak at coverage ~38 (**Table S1, Figure S5**).

Evolutionary studies on developmental genes

The Atlantic sturgeon genome sequence contained 7 Hox clusters (**Figure 4**): HoxA, HoxB and HoxD clusters were duplicated, while HoxC was present as a single cluster. In total, the Hox complement was very complete and conserved in both duplicates, and consisted of in total 76 genes and one pseudogene (HoxD14β) (**Table S2**). All Hox genes contained 2 coding exons with the exception of HoxD14α, which consisted of 3 coding exons. The sturgeon Hox complement in each of the duplicated clusters was identical to Spotted gar with the exception of paralog group (PG) 14 of cluster HoxD. While HoxD14 is a pseudogene in the spotted gar, in Atlantic sturgeon we found a functional HoxD14 gene in the α cluster and a pseudogene in the β cluster, with a stop codon at exon 2 resulting in a truncated protein lacking the homeodomain. In sterlet, both HoxD14 paralogs seemed to be pseudogenes, and additionally exon 2 of the HoxD3α, HoxD4α and HoxD4β genes were not present in the genome assembly retrieved from NCBI. Unlike the examined teleosts, the cluster length in the Atlantic sturgeon duplicates is quite conserved: For clusters HoxA and HoxB one of the paralogs is ~10% longer as a result of larger intergenic content, while both HoxD clusters have the same length and HoxC, the only cluster present as a single copy, is the longest (187.5 kb) due to the presence of larger intergenic regions and introns. The amino acid sequence divergence between paralog genes ranged from 0.5% for HoxD3 to 29.45% for HoxB2, being 8.04% the average divergence. Interestingly, three polycistronic messengers were found in the multiorgan assembly, namely HoxA5α-HoxA3α, HoxA3β-HoxA2β and Hox6β-Hox A5β. All the annotated Hox genes were transcribed at one or more of scrutinized embryonic development stages (**Table S3**). HoxC genes showed the highest average expression levels at all the evaluated timepoints, while HoxA genes showed the lowest expression level, which increased as embryonic development progressed. HoxB and HoxD genes had the highest average expression level during the complete gastrula timepoint. During the mid-gastrula (MG) period, HoxB10β, HoxC1 and both HoxB1 and

HoxD1 paralogs showed the highest expression level (**Figure 5A**), while upon gastrula completion (GC) the expression levels of HoxA3β, HoxB7β, HoxB8β, HoxB9β, HoxC8 and HoxD9β increased (**Figure 5B**). After hatching only the HoxB7β and HoxC1 genes remained highly expressed (**Figure 5C**), both of which were also highly expressed 2 days post-hatching (**Figure 5D**).

Using the protein accessions provided in **Table S4**, phylogenetic trees were built for each of the Hox clusters using the sharks as outgroup. The resulting trees showed a Hox clusters split in Sarcopterygian and Actinopterygian clades (**Figure 6**). Among sarcopterygians, the coelacanth had a basal position to the human. Hox clusters of coelacanth showed very little divergence when compared with human, which in turn had particularly high divergence, especially for clusters HoxB and HoxD according to the branch length. Actinopterygians were divided into chondrosteans and neopterygians, which were further divided into holosteans (Spotted gar), branching at the base, and the sister clade of teleosts. Within teleosts, two different groups indicated the duplication of the clusters in agreement with the TSGD, followed by speciation events. In each group of duplicate Hox clusters, the eel had a basal position to zebrafish. In addition, zebrafish showed one of the duplicates particularly divergent according to the branch length and also to the loss of the HoxD cluster. Among chondrosteans, the differentiation event leading to paddlefish and sturgeon groups was followed by a single WGD event affecting paddlefish (PR) in the included HoxA and HoxD clusters, and another independent duplication event within the sturgeon group.

Synteny studies suggested remarkable conservation levels of syntenic blocks between both HoxA duplicates with a few exceptions downstream of the HoxA clusters. The neighboring genes *pde11al*, *stk3, fam221* and *c7orf31* suffered asymmetric losses in the Atlantic sturgeon genome, with the first three being absent in the HoxAα cluster synteny and the last one being lost in the HoxAβ paralog (**Figure 7**). Teleost Hox paralogs showed low synteny conservation and many asymmetric losses of both Hox-coding and neighboring genes. Considering the HoxC cluster, we have also found conserved synteny compared with Spotted gar and other teleosts and sarcopterygians; however, both Hox-coding and neighboring genes were single copy in the Atlantic sturgeon genome, suggesting that the entire cluster, including neighboring genes, were either lost after the SR event or not affected by it.

Discussion

In order to generate genomic resources, and to elucidate the ploidy level and phylogenetic relationships of the endangered Atlantic sturgeon, we have sequenced its genome using Illumina and nanopore technologies. Additionally, genome size and heterozygosity were also investigated and evolutionary studies were performed using Hox cluster sequences in order to elucidate the position of sturgeons among jawed vertebrates.

In this work, we present for the first time a highly contiguous and complete reference genome for Atlantic sturgeon. Firstly, we have assembled Illumina sequencing reads testing multiple k-mer and bubble size combinations in order to find the optimal combination to achieve a contiguous and complete assembly. Long k-mers are more unique in the genome and should therefore result in longer contigs, however they might provide fewer connections

between reads spanning low coverage regions. On the other hand, the bubble size is the maximum length of ambiguities in a k-mer graph that the assembler will try to resolve based on coverage. Overall, the highest k-mer tested (63 nt) showed the best results and, when increasing of the bubble size from 50 bp to 250 bp, the contig N_{50} doubled without significant changes in the assembly size (data not shown). The final Illumina-based assembly was performed at k-mer 63 and bubble size 250 bp and resulted in an assembly size of 1.37 Gb divided over almost a million contigs with low N_{50} (3-fold smaller than the nanopore read N_{50}). Although only 240,230 Illumina contigs (24.6%) were longer than 1.5 kb, they contained up to 75.1% of the total assembled genome (**Figure S2**) and were used to extract seeds for the nanopore assembly. Nanopore reads had a lower quality score; however, the N_{50} was more than 85 times higher (13.13 kb) than that of the Illumina reads. We have exploited the advantages of each of the technologies to assemble the Atlantic sturgeon genome using up to 100-fold coverage of Illumina or nanopore sequencing data. Illumina contigs were fragmented into pieces of 1.5 kb in length to which the nanopore reads were aligned to build an assembly graph using the Tulipa-julia assembler, whose previous version was used to improve the contiguity of the European eel genome (Jansen et al. 2017). The nanopore assembly showed a 250-fold increase in the contig N_{50} to almost 1 Mb, a slightly smaller assembly size compared to the Illumina assembly and a BUSCO completeness score of 92.8%. Conversely, k-mer counting in Illumina reads (**Figure 3**) suggested an extremely high heterozygosity (5.36%) and a genome size of ~653 Mb, which is approximately 2 times smaller than the assembled size, and 4 times smaller than was estimated in previous studies using FIA and flow cytometry (Hardie and Hebert 2003). Although our flow cytometry results suggested a haploid genome size of 1.1 Gb for Atlantic sturgeon using the European eel as reference, species with larger and smaller genome sizes should have been included in order to increase the accuracy of our genome size estimations.

In order to investigate the coverage distribution of the nanopore assembly, we have aligned a set of Illumina reads to the genomic contigs. Considering the amount of Illumina reads used for the alignment (61.59 Gb) and the size of the assembled genome (1.28 Gb), the expected average contig coverage should be ~48-fold. Although the average contig coverage was indeed 48, the peak was located at a lower coverage (~36.5) and the distribution was skewed to the right, indicating that some contigs had higher coverage than others, possibly due to their higher repeat content (**Figure S5**).

Five out of 8 microsatellite loci had a disomic band pattern while tetrasomic allelic patterns were observed for 3 of them. There is evidence that ~120 chromosome sturgeon species can have a tetraploid origin with their diploid ancestor being extinct (Dingerkus and Howell 1976). Functional diploidization is considered as the switch from 4 alleles at a single locus (tetrasomic inheritance) to 2 alleles at each of the 2 distinct loci (disomic inheritance) (Wolfe 2001). Havelka et al. (2013) have previously investigated functionally diploid Acipenserids in paleotetraploid condition, showing a disomic allelic band in some microsatellite loci, whereas a residual tetrasomic pattern was observed in others (Havelka et al. 2013). Therefore, the coexistence of disomic and tetrasomic bands in Atlantic sturgeon is an indication that such rediploidization is still ongoing (Tagliavini et al. 1999; Rajkov et al. 2014). Delayed rediploidization has already been described in Atlantic salmon, in which

diploidy has not yet been fully reestablished after the recent Ss4R and 10% of the genome might still retain residual tetrasomy (Lien et al. 2016).

Non-duplicated genomes like chondrichthians and sarcopterygians usually contain 4 Hox clusters (Kim et al. 2000; Amemiya et al. 2013; Venkatesh et al. 2014). Non-teleost actinopterygians like the holostean Spotted gar (*Lepisosteus oculatus*) also contain 4 Hox cluster with 43 Hox genes (Braasch et al. 2016). As a result of the 3R, teleosts like the European eel (73 Hox genes) contain up to 8 Hox clusters (Henkel et al. 2012a), while others like zebrafish (53 genes) and medaka (46 genes) contain only 7 (zebrafish and medaka have lost Db and Cb clusters, respectively). The teleost Atlantic salmon, which has suffered a lineage-specific duplication event (Ss4R) contains 13 Hox clusters. We have found 76 Hox genes in the Atlantic sturgeon genome, distributed over 7 clusters including duplicated HoxA, HoxB and HoxD clusters and a single-copy HoxC cluster (**Figure 4**). All of the scrutinized Hox genes in the American paddlefish were found in the Atlantic sturgeon genome, which has only lost HoxA14, HoxB14 and HoxD5 compared to ghostshark. Atlantic sturgeon's Hox gene complement in each duplicate is the same as the Spotted gar, with the additional presence of HoxD14α. PG 14 genes, which contain three coding exons, were initially described in the coelacanth and have subsequently been found in the genome of two lamprey species, several cartilaginous fishes and other sarcopterygians like lungfish (Powers and Amemiya 2004; Feiner et al. 2011; Liang et al. 2011). To date, the American paddlefish and the Atlantic sturgeon are the only two actinopterygians for which the preservation of the PG 14 in the HoxD cluster has been reported, indicating their presence in the last common ancestor.

Interestingly, the contig coverage distribution results showed that duplicated Hox clusters (A, B and D) had an average coverage close to the coverage peak (31.4-fold), while the HoxC cluster, which is present as a single copy, showed a genomic coverage of 63.3-fold, twice as high as the duplicated Hox clusters. Highly heterozygous genomes tend to produce assemblies with a size larger than expected as a result of redundant information in the form of contigs that constitute haplotypes (Small et al., 2007). Considering the high heterozygosity (5.38%) and small genome size (~653 Mb) reported by the k-mer counting approach, the contig coverage results could indicate that the duplicated Hox clusters A, B and D corresponded to haplotypes which have been assembled separately as a result of high heterozygosity in these loci, while HoxC cluster might show less heterozygosity and resulted in a single contig.

Considering that the k-mer fitting model applied by GenomeScope is only suitable for diploids, another possible explanation is that Atlantic sturgeon is a paleotetraploid with delayed rediploidization, similar to the situation in Atlantic salmon (Lien et al. 2016; Robertson et al. 2017). Assuming this scenario, the k-mer peak identification performed by GenomeScope might be shifted: the surprisingly high error peak at coverage ~12.5 would correspond to the heterozygous k-mer peak, while the second peak, identified as the heterozygous peak, would correspond to the homozygous peak. Therefore, heterozygosity, which is calculated based on the homozygous and heterozygous peaks ratio, would be much smaller, and the genome size, calculated based on the homozygous peak, would be close to the assembly size. Therefore, the duplicated Hox clusters would be the result of a delayed rediploidization rather than high heterozygosity.

In order to test both hypotheses, we have verified the transcriptional activity of the annotated Hox genes in 4 stages of embryonic development, studied the synteny and phylogeny. Haplotypes should share synteny and, assuming that heterozygosity is evenly distributed throughout the genome, they are expected to differ as much as the heterozygosity rate (5.38%). On the other hand, paralogs should differ more than the heterozygosity rate and might or might not conserve the synteny depending on the age of the duplication, the length of the life cycle (and therefore number of meiosis) and the rate of evolution.

The Hox paralogs differed on average by 8.04%, which is higher than the divergence expected by heterozygosity but slightly smaller than the average divergence of Hox paralogs observed in the paddlefish (9.76%).

We have additionally identified polycistronic messengers in the HoxAα (5α-3α) and HoxAβ (3β-2β, 6β-5β) clusters of Atlantic sturgeon, and although most eukaryotic transcripts are typically monocistronic, Hox polycistronism has been previously observed in a variety of species (Simeone et al. 1988; Shiga et al. 2006; Michaut et al. 2011) and is believed to be involved in post-transcriptional regulation and maintenance of Hox clusters as a functional unit (Mainguy 2007; Boldogköi 2012).

Anterior (3') Hox genes B1α, B1β, D1α, D1β and C1 were highly expressed before hatching, of which only C1 maintained the expression levels after hatching. Overall, HoxB7 paralogs, which showed a sequence similarity of 90%, had differential expression during embryonic development: while HoxB7β was highly expressed after the onset of gastrulation, the HoxB7α paralog was not expressed at any of the studied timepoints. Among the Hox polycistronic transcripts, only HoxA2β and HoxA3β seemed to be expressed during the scrutinized embryonic development, with HoxA3β being more highly expressed before hatching and HoxA2β thereafter (**Figure 5**). A previous study in paddlefish larvae has shown the implication of HoxD13 paralogs in the rostrum transcriptome and, together with HoxD12, high expression levels in the barbel (Archambeault et al. 2014). However neither PG13 nor other 5' genes of the Hox clusters were expressed in any of the studied Atlantic sturgeon embryonic stages, which corresponded to an earlier developmental stage than the one studied in paddlefish.

Previous phylogenetic studies using cytochrome b amino acid sequences dated the split of paddlefish and sturgeons from 184 Mya (Peng et al. 2007), while more recent studies using transcriptome-derived amino acid sequences suggested that this event took place even earlier, around 204 Mya (Luo et al. 2019). Crown et al (2012) described the presence of duplicated HoxA and HoxD clusters in the American paddlefish and dated the duplication from ~ 41 Mya, based on amino acid sequence divergence (Crow et al. 2012). Therefore, it is assumed that the paddlefish duplication was specific to this family and has therefore been named paddlefish-specific WGD (PR). Additional work located the HoxA paralogs in two separate metacentric chromosomes of the paddlefish karyotype which, like the Atlantic and sterlet sturgeons, contains 2n=120 chromosomes (Symonová et al. 2017). Upon the identification of 7 Hox clusters in both sturgeons (Atlantic and sterlet), our phylogeny results revealed that Chondrosteans have suffered a speciation event followed by two independent WGD events: the already known PR, and the sturgeon-specific WGD (SR). Atlantic and sterlet sturgeons, which have largely been considered as functional diploids, belong to the sturgeon Sea and Atlantic clades respectively, which diverged ~145 Mya (Luo

et al. 2019). Therefore, the described SR event is assumed to have happened after the sturgeon speciation event and before the clade split (**Figure 6**), therefore between 204 and 145 Mya based on previous estimates (Luo et al. 2019).
Finally, we have performed synteny studies in both HoxA and HoxC clusters and compared it to other fish genomes. Neither paddlefish nor sterlet could be included in the study due to the short length of the Hox-containing scaffolds. The results showed that the Atlantic sturgeon HoxAα and HoxAβ duplicated clusters preserved a very conserved synteny compared to the Spotted gar, however 4 of the 18 identified neighboring genes showed asymmetric losses in the two copies. This suggests that they are paralogs derived from the SR event and located in different chromosomes, since their genomic sequences would likely not undergo homologous pairing during meiosis. On the other hand, HoxC coding-genes and neighboring genes were all present as single copy. Whether this happened before or after the speciation event or the SR duplication can't be determined due to scarce HoxC sequence information in the Order.

Overall, the k-mer and contig coverage results suggested that the Atlantic sturgeon assembled size was two times bigger than the genome size due to the presence of redundant contigs with half the coverage (e.g. HoxA, HoxB and HoxD paralogs) which might have been assembled separately as a result of the high heterozygosity (> 5%). However, nucleotide sequence divergence between the Hox paralogs was found to be higher than expected based on heterozygosity and phylogenetic analysis attributed this divergence to a WGD event. Additionally, the synteny of both HoxAα and HoxAβ clusters was not fully conserved. Due to their location in two separate chromosomes in the genome, HoxAα and HoxAβ were proved to be paralogs in the American paddlefish, sturgeon's closes relative. Hox paralogs amino acid sequences are more divergent in the American paddlefish than in the Atlantic sturgeon, indicating that paddlefish have undergone more evolutive processes (mutation, evolutive pressure, number of life cycles). Delayed rediploidization has been observed in Atlantic salmon after a recent (~80 Mya) WGD (Lien et al. 2016), while segmental paleotetraploidy was also evidenced by chromosome painting in sterlet, for which chromosomes underwent unequal rearrangements after the SR (Romanenko et al. 2015; Andreyushkova et al. 2017). Additionally supported by the microsatellite results, we provide in this work the first complete and contiguous genome sequence of Atlantic sturgeon. The genome has an estimated size of ~1.28 Gb and corresponds to a paleotetraploid in which rediploidization is still ongoing.

Availability of supporting data

Illumina genomic reads are deposited at NCBI Sequence Read Archive (SRA) SRR8119912.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Actions: Innovative Training Network "IMPRESS", Grant Agreement No. 642893.

Author Contributions

ESYP and RPD did the sampling. ESYP, HJJ and SAJR did the Illumina and nanopore sequencing libraries. SS performed the microsatellite analysis. HJJ and ESYP did the Illumina genome assembly. CVH developed the Tulia-julia software. AC and ESYP performed the evolutionary studies. ESYP, AC and SS wrote the first manuscript draft and all the authors contributed to the discussion of the results and manuscript revision.

Acknowledgements

The authors would like to thank Peter Gross, Vojta Kaspar and Karen Bosma for the help provided during sturgeon sampling and Elena Sarropoulou for her permission to use the Atlantic sturgeon embryonic reads.

Supplementary Material

Figure S1. Nanopore sequencing reads distribution after filtering. Histograms of read lengths in **A:** number of reads (in thousands) and **B**: in number of bases (in gigabases).
Figure S2. Size distribution of the Atlantic sturgeon Illumina assembly. Histograms of contig lengths in **A:** number of contigs (in thousands) and in **B:** number of bases (in kilobases).
Figure S3. Genome assembly completeness assessment using BUSCO against the vertebrata database. Complete single (light blue) and duplicate (dark blue) BUSCOS correspond to 92.8% of the total searched BUSCOs, while a small proportion were either fragmented (yellow) or missing (red).
Figure S4. Flow cytometry results of A) European eel and **B)** Atlantic sturgeon. The P5 peak corresponds to the cell population of interest (erythrocytes). The table below shows the number of registered events and the median fluorescence in the BLUE 695/40-A.
Figure S5. Coverage distribution in the final assembly. The Y-axis indicates the coverage frequency whereas the X-axis corresponds to the amount of coverage. Blue bars represent the observed coverage distribution, skewed towards high coverage contigs and peaking at coverage 38.
Table S1. Coverage distribution in the final assembly. The number of aligned reads is indicated for each contig in the nanopore assembly, indicating the size of each nanopore contig, which was used for the normalization of the coverage [(n. aligned reads * Illumina read length) / (size of nanopore contig)].
Table S2. List of annotated Hox paralogs. The table summarizes for each Hox paralog the acid length, number of exons, best blast hit (including e-value, identity and accession number) and the percentage of identity between paralogs.

Table S3. Number of embryonic reads per sample and replicate (A, B and c) mapped to the Hox genes expressed in raw values, RPKM and mean values per group including the standard deviation (SD). MG: mid-gastrula; **GC:** Gastrula complete; **H:** Hatching; **2DPH:** 2 days post-hatching.

Table S4. Protein accession numbers used for the phylogenetic studies.

File S1. Hox amino acid superalignments.

References

Amemiya, C.T., Alföldi, J., Lee, A.P., Fan, S., Philippe, H. et al. 2013. The African Coelacanth Genome Provides Insights into Tetrapod Evolution. *Nature* 496: 311–316. https://doi.org/10.1038/nature12027.

Amores, A. 1998. Zebrafish Hox Clusters and Vertebrate Genome Evolution. *Science* 282 (5394): 1711–1714. https://doi.org/10.1126/science.282.5394.1711.

Andrews, S. 2010. FASTQC. A Quality Control Tool for High Throughput Sequence Data. *Http://Www.Bioinformatics.Babraham.Ac.Uk/Projects/Fastqc/*. https://doi.org/v.0.11.5.

Andreyushkova, D.A., Makunin, A.I., Beklemisheva, V.R., Romanenko, S.A., Druzhkova, A.S., et al. 2017. Next Generation Sequencing of Chromosome-Specific Libraries Sheds Light on Genome Evolution in Paleotetraploid Sterlet (*Acipenser ruthenus*). *Genes*. https://doi.org/10.3390/genes8110318.

Archambeault, S., Taylor, J.A. and Crow, K.D. 2014. HoxA and HoxD Expression in a Variety of Vertebrate Body Plan Features Reveals an Ancient Origin for the Distal Hox Program. *EvoDevo* 5 (1): 1–10. https://doi.org/10.1186/2041-9139-5-44.

Betancur, R., Broughton, R.E., Wiley, E.O., Carpenter, K., López, J.A. et al. 2013. The Tree of Life and a New Classification of Bony Fishes. *PLOS Currents Tree of Life* 0732988: 1–45. https://doi.org/10.1371/currents.tol.53ba26640df0ccaee75bb165c8c26288.Abstract.

Blacklidge, K.H., and Bidwell, C.A. 1993. Three Ploidy Levels Indicated by Genome Quantification in Acipenseriformes of North America. *The Journal of Hereduty* 84 (6): 427–30.

Boldogköi, Z. 2012. Transcriptional Interference Networks Coordinate the Expression of Functionally Related Genes Clustered in the Same Genomic Loci. *Frontiers in Genetics* 3 (July): 1–17. https://doi.org/10.3389/fgene.2012.00122.

Börk, K., Drauch, A., Israel, J.A.,Pedroia, J., Rodzen, J. et al. 2008. Development of New Microsatellite Primers for Green and White Sturgeon. *Conservation Genetics* 9 (4): 973–79. https://doi.org/10.1007/s10592-007-9417-9.

Braasch, I., Gehrke, A.R., Smith, J.J., Kawasaki, K., Manousaki, T. et al. 2016. The Spotted Gar Genome Illuminates Vertebrate Evolution and Facilitates Human-Teleost Comparisons. *Nature Genetics* 48 (4): 427–37. https://doi.org/10.1038/ng.3526.

Bruce, A.E., Oates, A.C., Prince, V.E., and Ho, R.K. 2001. Additional Hox Clusters in the Zebrafish: Divergent Expression Patterns Belie Equivalent Activities of Duplicate HoxB5 Genes. *Evolution and Development* 3 (3): 127–44. https://doi.org/10.1046/j.1525-142X.2001.003003127.x.

Crow, Karen D., Christopher D. Smith, Jan Fang Cheng, G??nter P. Wagner, and Chris T. Amemiya. 2012. "An Independent Genome Duplication Inferred from Hox Paralogs in the American Paddlefish-a Representative Basal Ray-Finned Fish and Important Comparative Reference." *Genome Biology and Evolution* 4 (9): 937–53. https://doi.org/10.1093/gbe/evs067.

David, L., Blum, S., Feldman, M.W., Lavi, U. and Hillel, J. 2003. Recent Duplication of the Common Carp (*Cyprinus Carpio* L.) Genome as Revealed by Analyses of

Microsatellite Loci. *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msg173.

De Coster, W., D'Hert, S., Schultz, D.T., Cruts, M. and Van Broeckhoven, C. 2018. NanoPack: Visualizing and Processing Long-Read Sequencing Data. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bty149.

Dingerkus, G. and Howell, M.W. 1976. Karyotypic Analysis and Evidence of Tetrapolidy in the North American Paddlefish, Polydon Spathula. *Science* 194: 842–44.

Duboule, D. 2007. The Rise and Fall of Hox Gene Clusters. *Development* 134 (14): 2549–60. https://doi.org/10.1242/dev.001065.

Feiner, N., Ericsson, R., Meyer, A. and Kuraku, S. 2011. Revisiting the Origin of the Vertebrate Hox14 by Including Its Relict Sarcopterygian Members. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*. https://doi.org/10.1002/jez.b.21426.

Fontana, F., Lanfredi, M., Kirschbaum, F., Garrido-Ramos, M.A., Robles, F. et al. 2008. Comparison of Karyotypes of Acipenser Oxyrinchus and A. Sturio by Chromosome Banding and Fluorescent in Situ Hybridization. *Genetica* 132 (3): 281–86. https://doi.org/10.1007/s10709-007-9171-4.

Freyhof, J. and Brooks, E. 2011. *European Red List of Freshwater Fishes*. https://doi.org/10.2779/85903.

Gessner, J. and Bartel, R. 2000. Sturgeon Spawning Grounds in the Odra River Tributaries : A First Assessment. *Boletin Instituto Espanol de Oceanografia* 16 (1–4): 127–37.

Grunow, B.S, Noglick, S., Kruse, C. and Gebert, M. 2011. Isolation of Cells from Atlantic Sturgeon Acipenser Oxyrinchus Oxyrinchus and Optimization of Culture Conditions. *Aquatic Biology* 14: 67–75. https://doi.org/10.3354/ab00383.

Hardie, D.C. and Hebert, P.D.N. 2003. The Nucleotypic Effects of Cellular DNA Content in Cartilaginous and Ray-Finned Fishes. Genome / National Research Council Canada = Génome / Conseil National de Recherches Canada 46 (4): 683–706. https://doi.org/10.1139/g03-040.

Havelka, M., Hulák, M., Bailie, D.A., Prodöhl, P.A. and Flajšhans, M. 2013. Extensive Genome Duplications in Sturgeons: New Evidence from Microsatellite Data. *Journal of Applied Ichthyology* 29 (4): 704–8. https://doi.org/10.1111/jai.12224.

Havelka, M., Bytyutskyy, D., Symonová, R., Ráb, P. and Flajšhans, M. 2016. The Second Highest Chromosome Count among Vertebrates Is Observed in Cultured Sturgeon and Is Associated with Genome Plasticity. *Genetics Selection Evolution* 48 (1): 12. https://doi.org/10.1186/s12711-016-0194-0.

Henkel, C.V., Burgerhout, E., de Wijze, D.L., Dirks, R.P., Minegishi, Y. et al. 2012. Primitive Duplicate Hox Clusters in the European Eel's Genome." *PLoS ONE* 7 (2). https://doi.org/10.1371/journal.pone.0032231.

Hughes, L.C., Ortí, G., Huang, Y., Sun, Y., Baldwin, C.C. et al. 2018. Comprehensive Phylogeny of Ray-Finned Fishes (Actinopterygii) Based on Transcriptomic and Genomic Data. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1719358115.

Jansen, H.J., Liem, M., Jong-Raadsen, S.A., Dufour, S., Weltzien, F.A. et al. 2017. Rapid de Novo Assembly of the European Eel Genome from Nanopore Sequencing Reads. *Scientific Reports*. https://doi.org/10.1038/s41598-017-07650-6.

Kaitetzidou, E., Ludwig, A., Gessner, J. and Sarropoulou, E. 2017. Expression Patterns of Atlantic Sturgeon (*Acipenser oxyrinchus*) During Embryonic Development. *G3 Genes|Genomes|Genetics* 7 (2): 533–42. https://doi.org/10.1534/g3.116.036699.

Kim, C.B., Amemiya, C.T., Bailey, W.J., Kawasaki, K., Mezey, J.G. et al. 2000. Hox Cluster Genomics in the Horn Shark, Heterodontus Francisci. *Proceedings of the*

National Academy of Sciences of the United States of America 97 (4): 1655–60. https://doi.org/10.1073/pnas.030539697.

King, B.L., Gillis, J.A., Carlisle, H.R. and Dahn, R.D. 2011. A Natural Deletion of the HoxC Cluster in Elasmobranch Fishes. *Science* 334 (6062): 1517. https://doi.org/10.1126/science.1210912.

King, L., Lubinski, A. and Spidle, P. 2001. Microsatellite DNA Variation in Atlantic Sturgeon (Acipenser. *Conservation Genetics* 2 (Schultz 1980): 103–19.

Kondrashov, F.A., Rogozin, I.B., Wolf, Y.I. and Koonin, E.V. 2002. Selection in the Evolution of Gene Duplications. *Genome Biology* 3 (2): 1–9.

Kuraku, S. and Meyer, A. 2009. The Evolution and Maintenance of Hox Gene Clusters in Vertebrates and the Teleost-Specific Genome Duplication. *International Journal of Developmental Biology* 53 (5–6): 765–73. https://doi.org/10.1387/ijdb.072533km.

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

Liang, D., Wu, R., Geng, J., Wang, C. and Zhang, P. 2011. A General Scenario of Hox Gene Inventory Variation among Major Sarcopterygian Lineages. *BMC Evolutionary Biology*. https://doi.org/10.1186/1471-2148-11-25.

Lien, S., Koop, B.F., Sandve, S.R., Miller, J.R., Matthew, P. et al. 2016. The Atlantic Salmon Genome Provides Insights into Rediploidization. *Nature* 533 (6020): 200–205. https://doi.org/10.1038/nature17164.

Ludwig, A. 2008. Identification of Acipenseriformes Species in Trade. *Journal of Applied Ichthyology* 24: 2–19. https://doi.org/10.1111/j.1439-0426.2008.01085.x.

Ludwig, A., Belfiore, N.M., Pitra, C., Svirsky, V. and Jenneckens, I. 2001. Genome Duplication Events and Functional Reduction of Ploidy Levels in Sturgeon (Acipenser, Huso and Scaphirhynchus). *Genetics* 158 (3): 1203–15.

Ludwig, A., Debus, L., Lieckfeldt, D., Wirgin, I., Benecke, N. et al. 2002. When the American Sea Sturgeon Swam East. *Nature* 419 (6906): 447–48. https://doi.org/10.1038/419447a.

Luo, D., Li, Y., Zhao, Q., Zhao, L., Ludwig, A. et al. 2019. Highly Resolved Phylogenetic Relationships within Order Acipenseriformes According to Novel Nuclear Markers. *Genes*. https://doi.org/10.3390/genes10010038.

Lynchu, M. and Conery, J.S. 2000. The Evolutionnary Fate and Consequences of Duplicate Genes. *Science* 290: 1151–55. https://doi.org/10.1126/science.290.5494.1151.

Macqueen, D.J. and Johnston, I.A. 2014. A Well-Constrained Estimate for the Timing of the Salmonid Whole Genome Duplication Reveals Major Decoupling from Species Diversification. *Proceedings of the Royal Society B: Biological Sciences*. https://doi.org/10.1098/rspb.2013.2881.

Mainguy, G. 2007. Extensive Polycistronism and Antisense Transcription in the Mammalian Hox Clusters. *Plos One*, no. 4: 1–7. https://doi.org/10.1371/journal.pone.0000356.

Marçais, G. and Kingsford, C. 2011. A Fast, Lock-Free Approach for Efficient Parallel Counting of Occurrences of k-Mers. *Bioinformatics* 27 (6): 764–70. https://doi.org/10.1093/bioinformatics/btr011.

Martinez, P. and Amemiya, C.T. 2002. Genomics of the HOX Gene Cluster. *Comparative Biochemestry and Physiology* 133: 571–80. https://doi.org/10.2307/2851043.

McGinnis, W. and Krumlauf, R. 1992. Homeobox Genes and Axial Patterning. *Cell* 68 (2): 283–302. https://doi.org/10.1016/0092-8674(92)90471-N.

McQuown, E.C, Sloss, B.L., Sheehan, R.J., Rodzen, J., Tranah, G.J. et al. 2000. Microsatellite Analysis of Genetic Variation in Sturgeon: New Primer Sequences for Scaphyrhynchus and Acipenser. *Transactions of the American Fisheries Society* 129

(1): 1380–88. https://doi.org/10.1577/1548-8659(2000)129.
Meyer, A. and Schartl, M. 1999. Gene and Genome Duplications in Vertebrates: The One-to-Four (-to-Eight in Fish) Rule and the Evolution of Novel Gene Functions. *Current Opinion in Cell Biology* 11 (6): 699–704. https://doi.org/10.1016/S0955-0674(99)00039-3.
Michaut, L., Jansen, H.J., Bardine, N., Durston, A.J. and Gehring, W.J.. 2011. Analyzing the Function of a Hox Gene : An Evolutionary Approach. *Development, Growth & Differentiation*, 982–93. https://doi.org/10.1111/j.1440-169X.2011.01307.x.
Minguillón, C., Gardenyes, J., Serra, E., Castro, L.F.C., Hill-Force, A. et al. 2005. No More than 14: The End of the Amphioxus Hox Cluster. *International Journal of Biological Sciences* 1: 19–23.
Peng, Z., Ludwig, A., Wang, D., Dioso, R., Wei, Q. et al. 2007. Age and Biogeography of Major Clades in Sturgeons and Paddlefishes (Pisces: Acipenseriformes). *Molecular Phylogenetics and Evolution* 42 (3): 854–62. https://doi.org/10.1016/j.ympev.2006.09.008.
Pierre, R., Parauka, F., 2006. (U.S. Fish & Wildlife Service). *Acipenser oxyrinchus*. The IUCN Red List of Threatened Species 2006 e.T245A107, http://dx.doi.org/10.2305/IUCN.UK.2006.RLTS.T245A1.
Powers, T.P. and Amemiya, C.T. 2004. “Evidence for a Hox14 Paralog Group in Vertebrates [2].” *Current Biology*. https://doi.org/10.1016/j.cub.2004.02.015.
Prince, V.E., Joly, L., Ekker, M. and Ho, R.K. 1998. Zebrafish Hox Genes: Genomic Organization and Modified Colinear Expression Patterns in the Trunk. *Development* 125 (3): 407–20.
Rajkov, J., Shao, Z. and Berrebi, P. 2014. Evolution of Polyploidy and Functional Diploidization in Sturgeons: Microsatellite Analysis in 10 Sturgeon Species. *Journal of Heredity* 105 (4): 521–31. https://doi.org/10.1093/jhered/esu027.
Robertson, F.M., Gundappa, M.K., Grammes, F., Hvidsten, T.R, Redmond, A.K. et al. 2017. Lineage-Specific Rediploidization Is a Mechanism to Explain Time-Lags between Genome Duplication and Evolutionary Diversification. *Genome Biology*. https://doi.org/10.1186/s13059-017-1241-z.
Rodzen, J.A. and May, B. 2002. Inheritance of Microsatellite Loci in the White Sturgeon (*Acipenser Transmontanus*). *Genome* 45 (6): 1064–76. https://doi.org/10.1139/g02-083.
Romanenko, S.A., Biltueva, L.S., Serdyukova, N.A., Kulemzina, A.I., Beklemisheva, V.R. et al. 2015. Segmental Paleotetraploidy Revealed in Sterlet (Acipenser Ruthenus) Genome by Chromosome Painting. *Molecular Cytogenetics* 8: 90. https://doi.org/10.1186/s13039-015-0194-8.
Rosa, R., Greiner, J.K., Andreeva, T., Adoutte, A., Akam, M. et al. 1999. Hox Genes in Brachipods and Priapulids and Protostome Evolution 399 (June): 772–76.
Shiga, Y., Sagawa, K., Takai, R., Sakaguchi, H., Yamagata, H. et al. 2006. Transcriptional Readthrough of Hox Genes Ubx and Antp and Their Divergent Post-Transcriptional Control during Crustacean Evolution. *Evolution and Development*. https://doi.org/10.1111/j.1525-142X.2006.00114.x.
Simeone, A., Pannese, M., Acampora, D., D’esposito, M. and Boncinelli, E. 1988. At Least Three Human Homeoboxes on Chromosome 12 Belong to the Same Transcription Unit. *Nucleic Acids Research*. https://doi.org/10.1093/nar/16.12.5379.
Small, K.S., Brudno, M., Hill, M.M. and Sidow, A. 2007. A Haplome Alignment and Reference Sequence of the Highly Polymorphic Ciona Savignyi Genome. *Genome Biology* 8 (3). https://doi.org/10.1186/gb-2007-8-3-r41.
Stamatakis, A. 2014. RAxML Version 8: A Tool for Phylogenetic Analysis and Post-

Analysis of Large Phylogenies. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btu033.

Symonová, R., Havelka, M., Amemiya, C.T., Howell, W.M., Ko, T. et al. 2017. Molecular Cytogenetic Differentiation of Paralogs of Hox Paralogs in Duplicated and Re-Diploidized Genome of the North American Paddlefish (Polyodon Spathula). *BMC Genomics* 18: 19. https://doi.org/10.1186/s12863-017-0484-8.

Tagliavini, J., Williot, P., Congiu, L., Chicca, M., Lanfredi, M. et al. 1999. Molecular Cytogenetic Analysis of the Karyotype of the European Atlantic Sturgeon, Acipenser Sturio. *Heredity* 83 (Pt 5) (July): 520–25. https://doi.org/10.1038/sj.hdy.6886150.

Vaser, R., Sovic, I., Nagarajan, N. and Sikic, M. 2017. Fast and Accurate de Novo Genome Assembly from Long Uncorrected Reads. *Genome Research* 27 (5): 737–46. https://doi.org/10.1101/068122.

Vasil'ev, V.P., Vasil'eva, E.D., Shedko, S.V. and Novomodny, G.V. 2010. How Many Times Has Polyploidization Occurred during Acipenserid Evolution? New Data on the Karyotypes of Sturgeons (Acipenseridae, Actinopterygii) from the Russian Far East. *Journal of Ichthyology*. https://doi.org/10.1134/s0032945210100048.

Venkatesh, B., Lee, A.P., Ravi, V., Maurya, A.K., Lian, M.M., et al. 2014. Elephant Shark Genome Provides Unique Insights into Gnathostome Evolution. *Nature* 505 (7482): 174–79. https://doi.org/10.1038/nature12826.

Venter, J.C., Adams, M.A., Myers, E.W., Li, P.W., Mural, R.J. et al. 2001. The Sequence of the Human Genome. *Science (New York, N.Y.)* 291 (February): 1304–51. https://doi.org/10.1126/science.1058040.

Vurture, G.W., Sedlazeck, F.J., Nattestad, M., Underwood, C.J., Fang, H. et al. 2017. GenomeScope: Fast Reference-Free Genome Profiling from Short Reads. *Bioinformatics* 33 (14): 2202–4. https://doi.org/10.1093/bioinformatics/xxxxx.

Wang, J.T., Li, J.T., Zhang, X.F. and Sun, X.W. 2012. Transcriptome Analysis Reveals the Time of the Fourth Round of Genome Duplication in Common Carp (*Cyprinus Carpio*). *BMC Genomics*. https://doi.org/10.1186/1471-2164-13-96.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A. et al. 2014. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS ONE* 9 (11): e112963. https://doi.org/10.1371/journal.pone.0112963.

Waterhouse, R.M., Seppey, M., Simao, F.A., Manni, M., Ioannidis, P. et al. 2018. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Molecular Biology and Evolution* 35 (3): 543–48. https://doi.org/10.1093/molbev/msx319.

Welsh, A.B., Blumberg, M. and Bernie, M. 2003. Identification of Microsatellite Loci in Lake Sturgeon, Acipenser Fluvescens, and Their Variability in Green Sturgeon A. Medirostris. *Molecular Ecology Notes* 3: 47–55. https://doi.org/10.1046/j.1471-8286.

Wolfe, K.H. 2001. Yesterday's Polyploids and the Mystery of Diploidization. *Nature Reviews Genetics*. https://doi.org/10.1038/35072009.

Yebra-Pimentel, E.S., Gebert, M., Jansen, H.J., Jong-Raadsen, S.A. and Dirks, R.P.H. 2019. Deep Transcriptome Analysis of the Heat Shock Response in an Atlantic Sturgeon (*Acipenser oxyrinchus*) Cell Line. *Fish and Shellfish Immunology* 88. https://doi.org/10.1016/j.fsi.2019.03.014.

Yuan, J., He, Z., Yuan, X., Jiang, X., Sun, X. et al. 2010. Speciation of Polyploid Cyprinidae Fish of Common Carp, Crucian Carp, and Silver Crucian Carp Derived from Duplicated Hox Genes. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*. https://doi.org/10.1002/jez.b.21350.

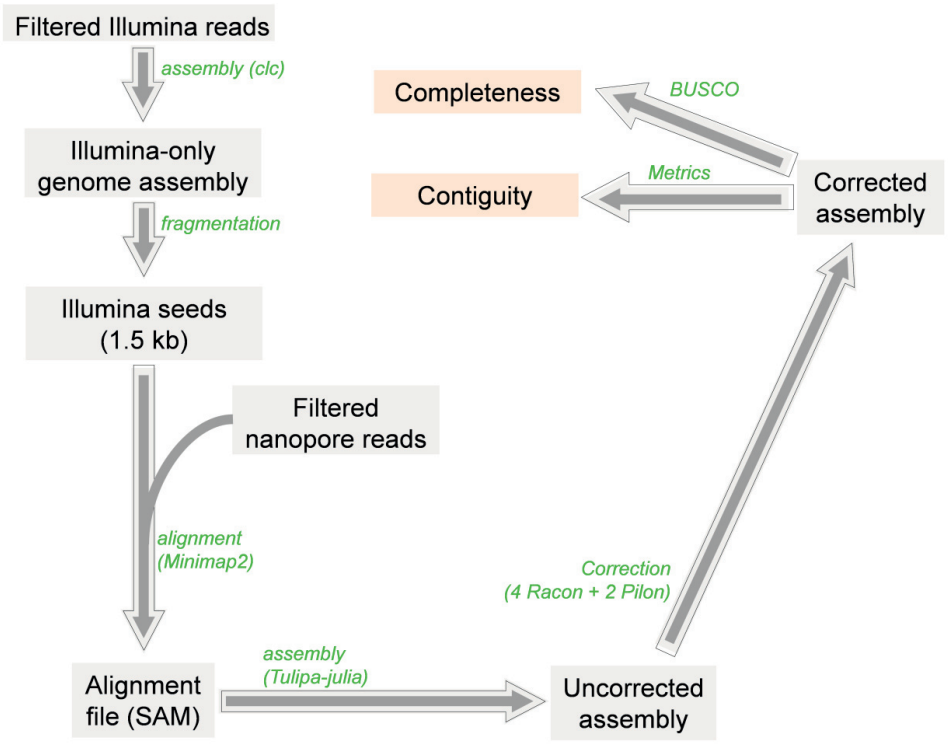


Figure 1. Overview of the genome assembly strategy followed for Atlantic sturgeon. Illumina reads were *de novo* assembled into contigs and fragmented in pieces of 1.5 kilobases (kb) to which the nanopore reads were aligned. The resulting alignment was used to assemble the Atlantic sturgeon genome with the Tulipa-julia assembly software, followed by four iterations of Racon correction and two Pilon correction rounds. The resulting assembly metrics were inspected and the completeness was evaluated using Benchmark Universal Single-Copy Orthologs (BUSCO).

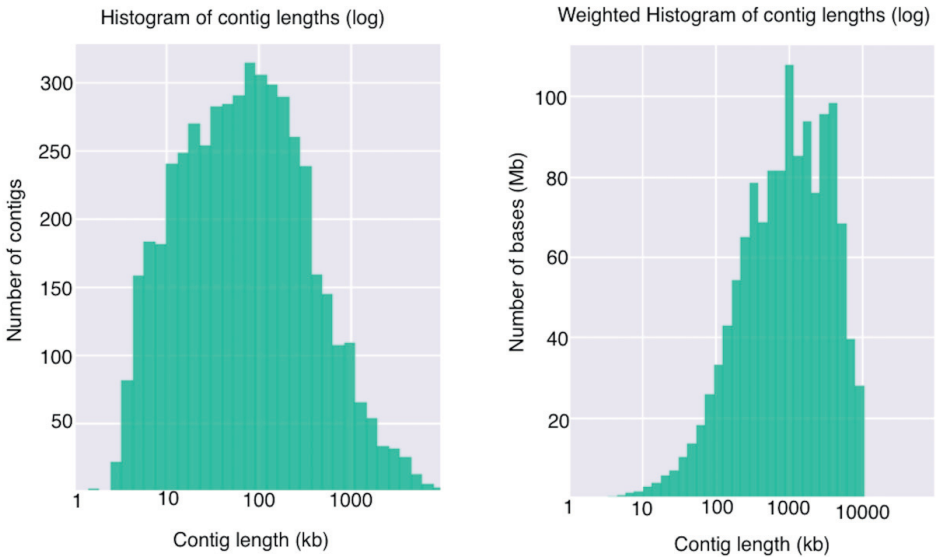


Figure 2. Size distribution of the Atlantic sturgeon assembly. Histograms of contig lengths in number of contigs (A) and in number of bases (B).

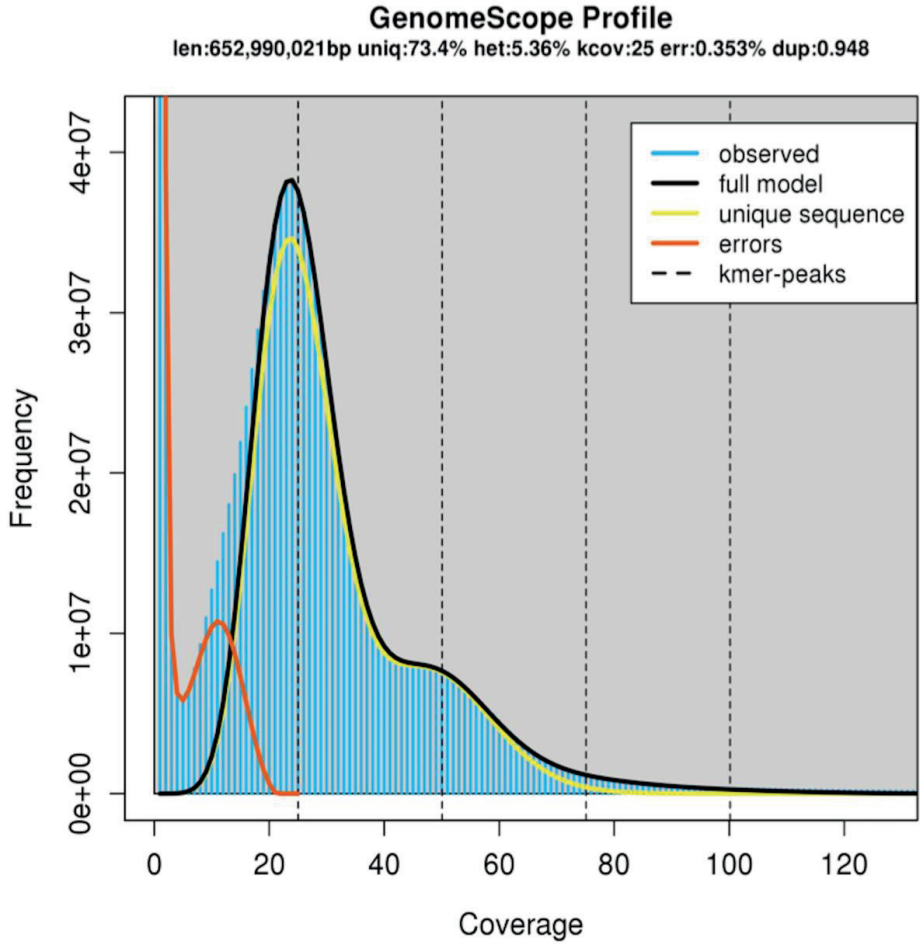


Figure 3. GenomeScope profile for k-mer 19. The Y-axis indicates the k-mer frequency whereas the X axis corresponds to the coverage of each k-mer. The blue bars represent the observed distribution, the black line represents the fitted model and the orange line represents the errors (k-mers with low coverage). The dotted vertical lines represent the k-mer peaks.

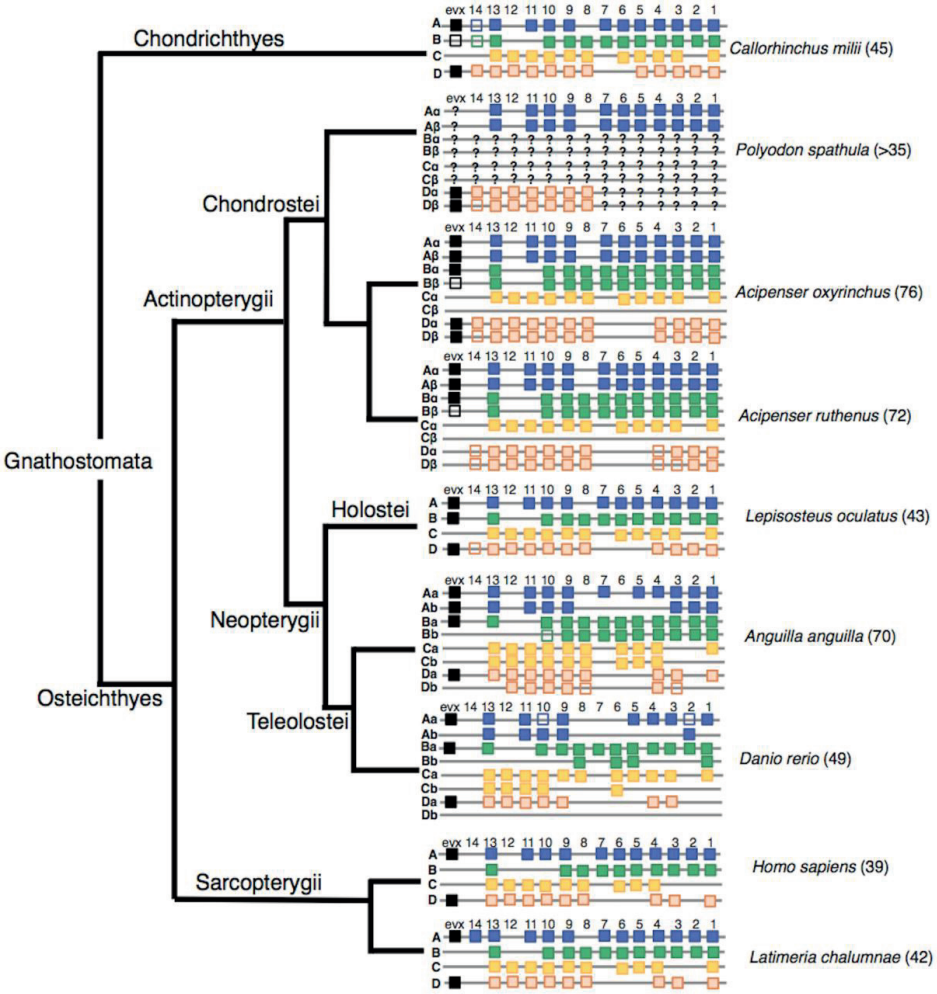


Figure 4. Model of the evolution of Hox clusters in gnathostomates. Schematic Hox clusters are depicted for each of the species, including the chondrichthian outgroup (*Callorhinchus milii*), sarcopterygians (*Latimeria chalumnae* and *Homo sapiens*), teleosteans (*Danio rerio* and *Anguilla anguilla*), holosteans (*Lepisosteus oculatus*) and chondrosteans (*Polyodon spathula, Acipenser oxyrinchus* and *Acipenser ruthenus*). Hollow boxes indicate pseudogenes or partial gene annotations. The number of Hox-coding genes is indicated in brackets.

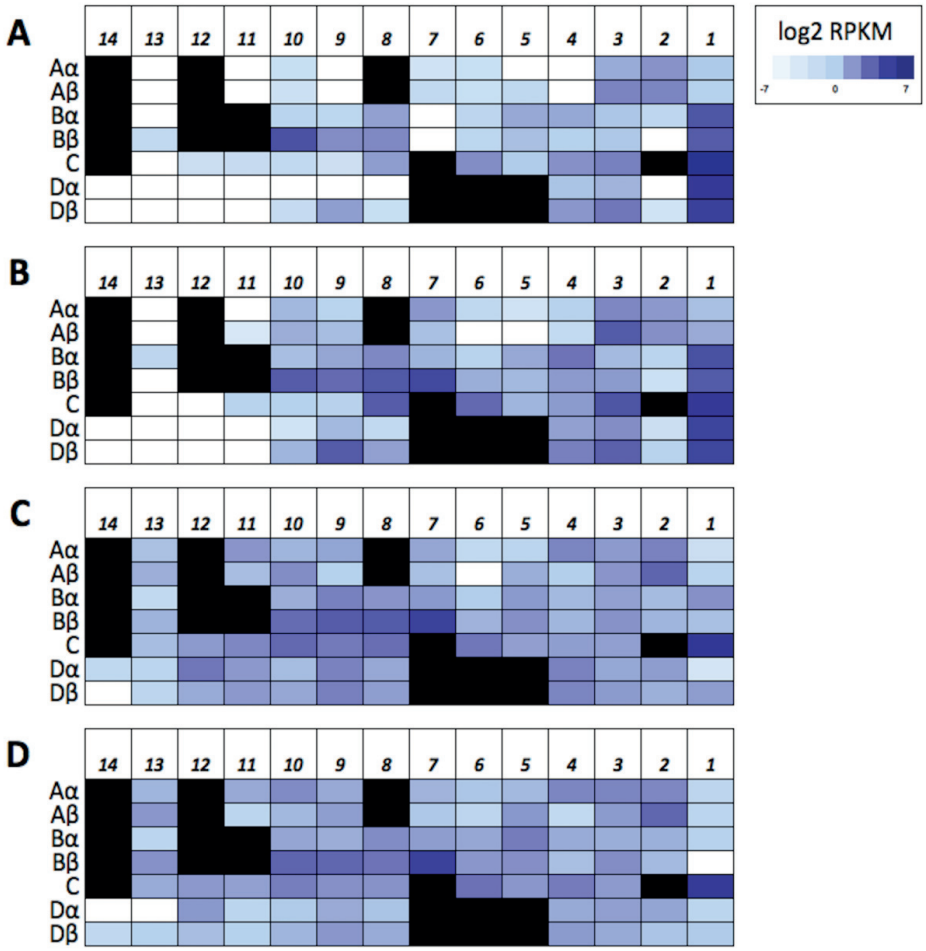


Figure 5. Expression of Hox clusters in Atlantic sturgeon embryos. Expression levels are depicted based on the log2 of the average RPKM (reads per kilobase per million) for each of the developmental stages. **A:** mid-gastrula; **B:** gastrula complete; **C:** hatching; **D:** 2 days post-hatching. Color key is expressed in log2 RPKM.

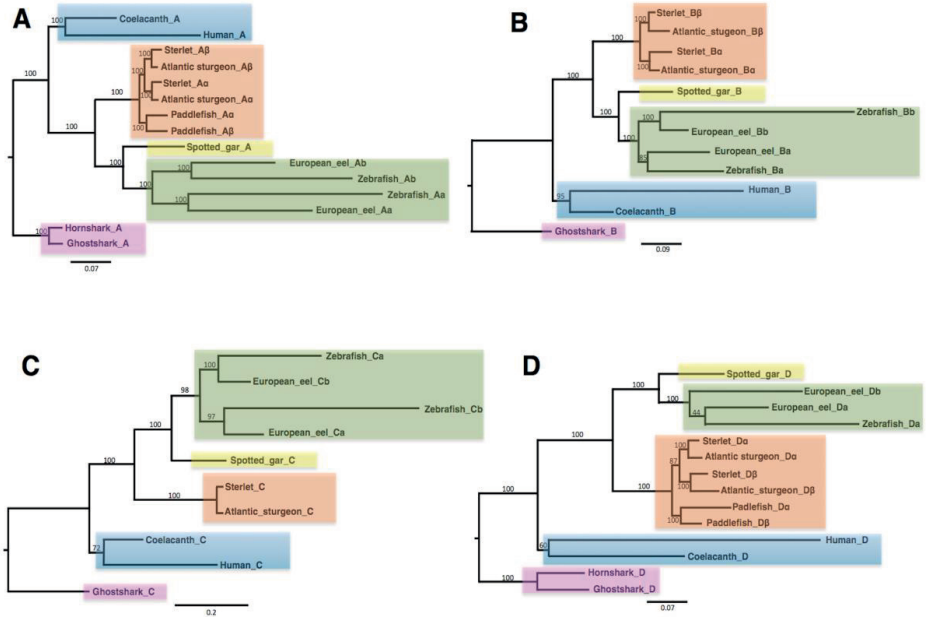


Figure 6. Consensus phylogenetic trees of the gnathostomates Hox cluster amino acid sequences for Hox clusters A: HoxA, **B:** HoxB, **C:** HoxC, **D:** HoxD. Phylogenetic trees were performed using the maximum likelihood method with 1,000 bootstrap replicates. The chondrichthian (pink) Ghostshark (*Callorhinchus milii*) was used as outgroup group for HoxB and HoxC clusters, while additionally Hornshark (*Heterodontus francisci*) was also used for clusters HoxA and HoxDa. The number shown at each branch node indicates the bootstrap value (%). The scale bar indicates the percentage of divergence in the amino acid sequence. The colored rectangles indicate different phylogenic divisions: sarcopterygians (blue), chondrosteans (orange), holosteans (yellow) and teleosteans (green).

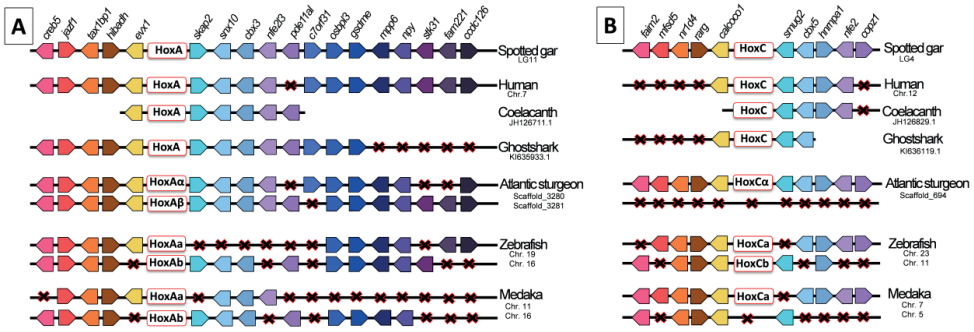


Figure 7. Syntenic region of Hox clusters in key gnathostomata species. A: HoxA clusters; **B:** HoxC cluster. Spotted gar (*Lepisosteus oculatus*) was used as a reference species for the synteny analyses. For each species, chromosome or scaffold is indicated.

The Atlantic sturgeon (Acipenser oxyrinchus) genome provides insights into genome evolution in paleotetraploids

Elena Santidrián Yebra-Pimentel,[*,1,2] Aurora Campo,[3] Sahana Shivaramu,[4] Christiaan V. Henkel,[2] Hans J. Jansen,[1] Susanne A. Jong-Raadsen, [1] and Ron P. Dirks[1]

Supplementary Material

Supplementary tables and files are available at:

https://www.dropbox.com/s/vejdqxatv5368l2/Article3_supplementary_Tables.zip?dl=0

Supplementary Tables:

- **Table S1. Coverage distribution in the final assembly.** The number of aligned reads is indicated for each contig in the nanopore assembly, indicating the size of each nanopore contig, which was used for the normalization of the coverage [(n. aligned reads * Illumina read length) / (size of nanopore contig)].
- **Table S2. List of annotated Hox paralogs.** The table summarizes for each Hox paralog the acid length, number of exons, best blast hit (including e-value, identity and accession number) and the percentage of identity between paralogs.
- **Table S3. Number of embryonic reads per sample and replicate (A, B and c) mapped to the Hox genes expressed in raw values, RPKM and mean values per group including the standard deviation (SD). MG:** mid-gastrula; **GC:** Gastrula complete; **H:** Hatching; **2DPH:** 2 days post-hatching.
- **Table S4. Protein accession numbers used for the phylogenetic studies.**

Supplementary Files:

- **File S1. Hox amino acid superalignments.**

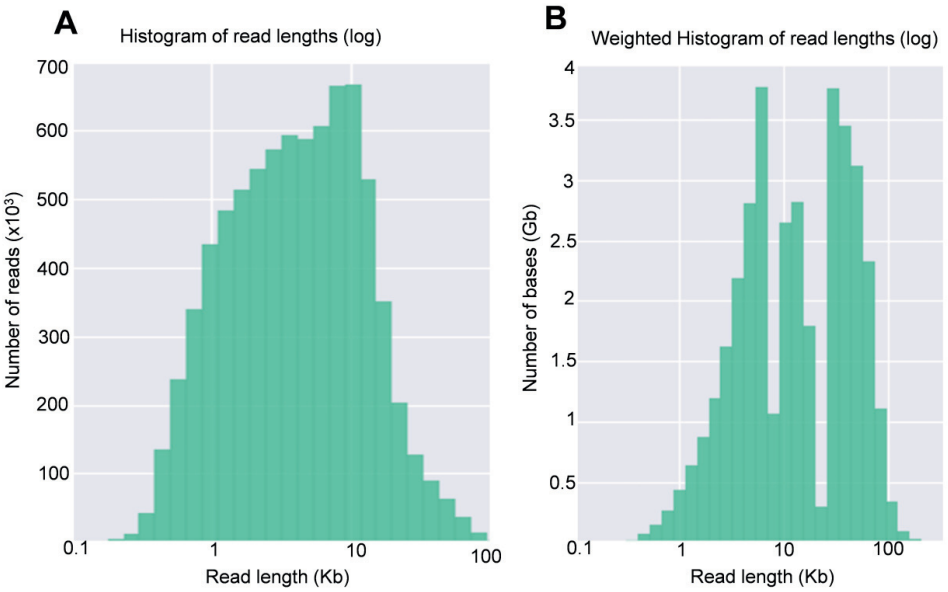


Figure S1. Nanopore sequencing reads distribution after filtering. Histograms of read lengths in **A:** number of reads (in thousands) and **B**: in number of bases (in Gigabases).

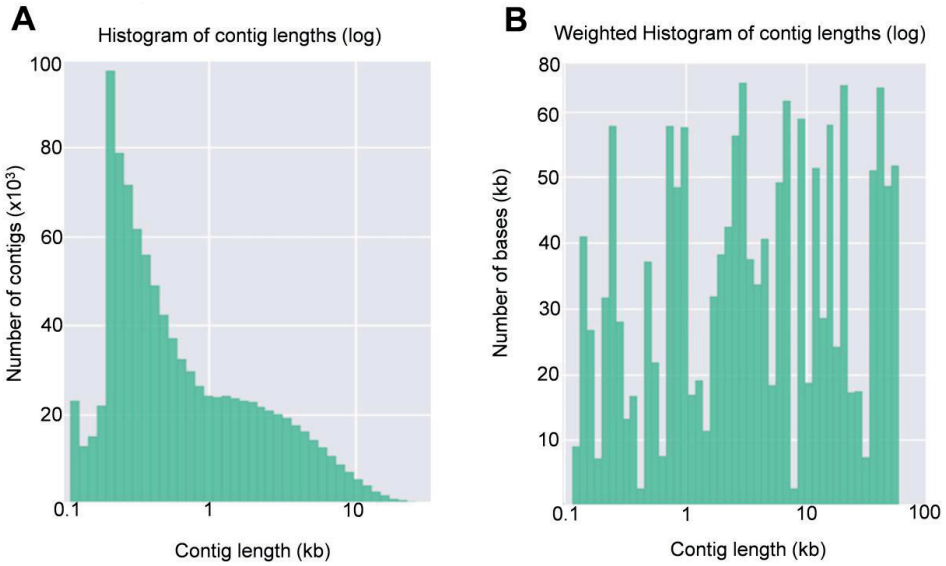


Figure S2. Size distribution of the Atlantic sturgeon Illumina assembly. Histograms of contig lengths in **A:** number of contigs (in thousands) and in **B:** number of bases (in kilobases).

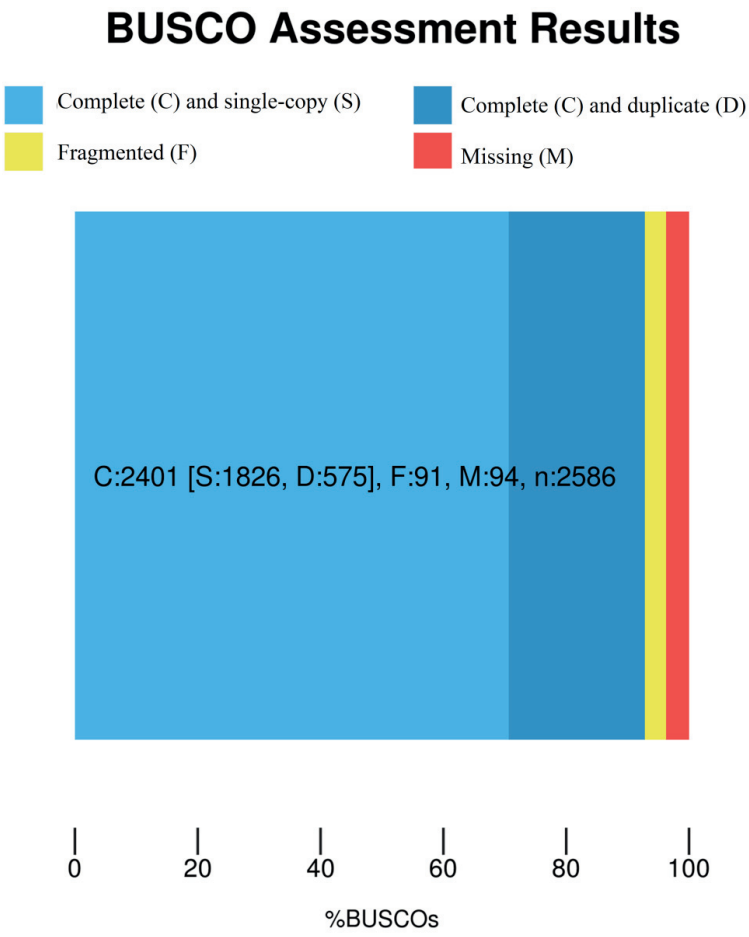
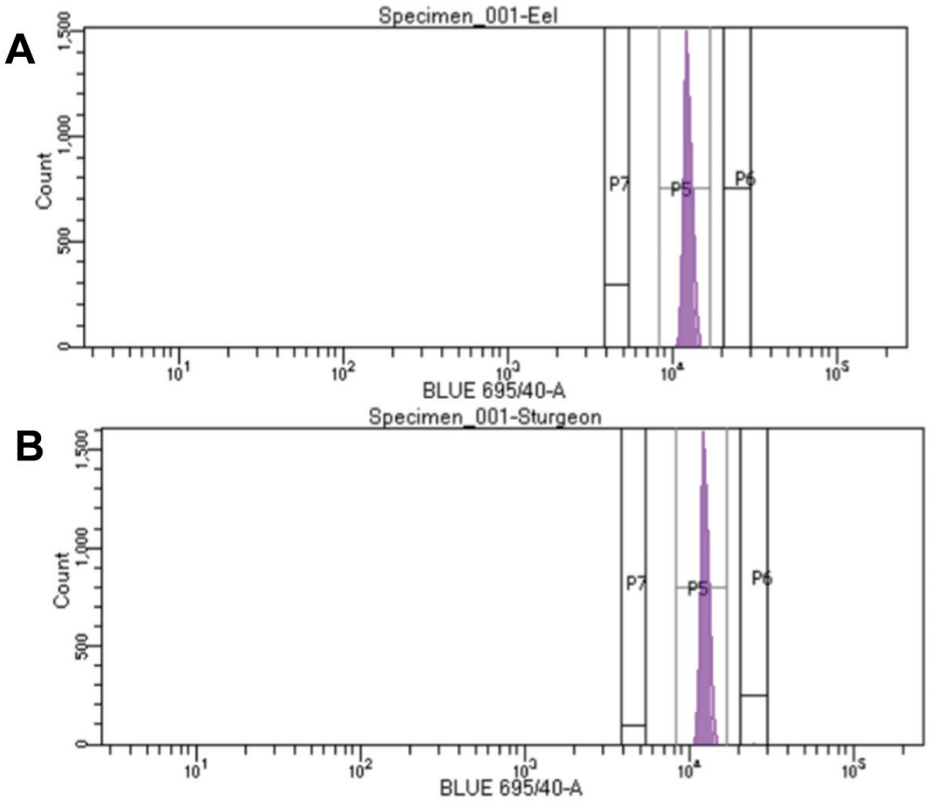


Figure S3. Genome assembly completeness assessment using BUSCO against the vertebrata database. Complete single (light blue) and duplicate (dark blue) BUSCOS correspond to 92.8% of the total searched BUSCOs, while a small proportion were either fragmented (yellow) or missing (red).



		Events	BLUE 695/40-A Median
Specimen_001-Eel	P5	4,150	12,135
Specimen_001-Sturgeon	P5	4,298	12,162

Figure S4. Flow cytometry results of A) European eel and **B)** Atlantic sturgeon. The P5 peak corresponds to the cell population of interest (erythrocytes). The table below shows the number of registered events and the median fluorescence in the BLUE 695/40-A.

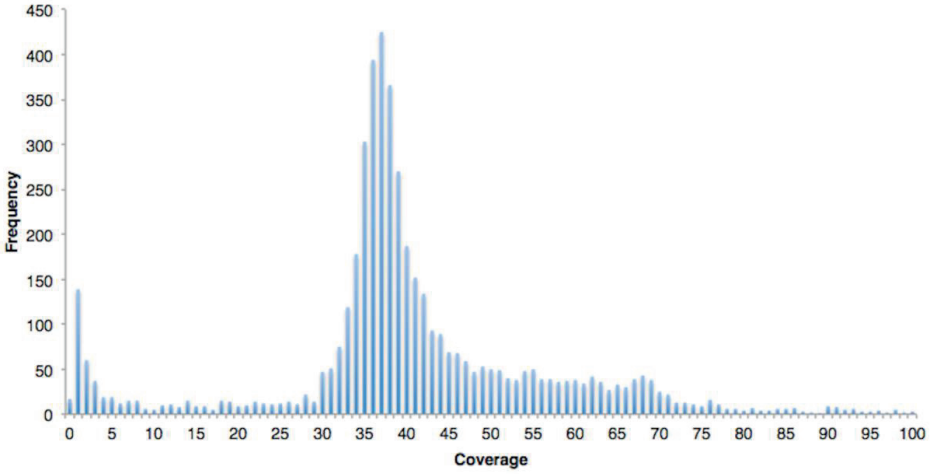


Figure S5. Coverage distribution in the final assembly. The Y-axis indicates the coverage frequency whereas the X-axis corresponds to the amount of coverage. Blue bars represent the observed coverage distribution, skewed towards high coverage contigs and peaking at coverage 38.

ISBN: 978-82-575-1649-9
ISSN: 1894-6402

Norwegian University
of Life Sciences

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