ADAPTIVE AND NEUTRAL GENETIC VARIATION IN SPRING- AND FALL-SPAWNING HERRING (*CLUPEA HARENGUS* L.) IN THE NORTHWEST ATLANTIC

by

Angela Patricia Fuentes Pardo

Submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
June 2019

To the memory of my dear father Luis Fuentes Cordoba and the wonderful women in my life, my mother Edith Pardo, sister Gloria Fuentes Pardo and grandmother Oliva Rivera. To them, infinite thanks for teaching me the value of love, strength, and commitment to achieve life goals

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Understanding the factors influencing the spatial distribution of intraspecific diversity is both a fundamental goal in evolutionary biology and necessary for the design of robust management strategies for harvested species. Intraspecific diversity can now be assessed using novel genomic techniques that enable the high-density screening of neutral and adaptive genetic variation. In this thesis, I used whole-genome resequencing of pooled DNA of Atlantic herring (*Clupea harengus*) to (*i*) identify patterns, genomic regions, evolutionary processes and environmental variables involved in the origin and maintenance of population divergence and local adaptation in the northwest Atlantic, and (*ii*) develop a diagnostic genetic tool that can inform conservation and fisheries management. I found fine-scale population structure in herring at putatively adaptive loci, despite low differentiation at neutrally evolving loci. Populations were distinguishable by spawning time and along a latitudinal cline defined by winter sea-surface temperature. Divergent selection leading to adaptation to seasonal reproduction and spawning at different latitudes is likely maintaining molecular divergence patterns for these traits in this environment. Each pattern was underpinned by thousands of outlier SNPs distributed in specific genomic regions spanning several candidate genes, some with a known role in the timing of reproduction (i.e. *TSHR*). Many spawning time-associated SNPs were shared with populations across the ocean, suggesting such adaptation predates the last glacial maximum. Finally, I developed and evaluated the efficacy of two cost-effective SNP-panels diagnostic of spawning season and latitudinal origin. Individual genotypes at these loci confirmed temporal stability of genetic differences among northern populations and between reproductive strategies despite their mixing outside of the spawning season, suggesting spawning time and site fidelity in ecological time scales. Admixture between reproductive and latitudinal components is unrestricted, and the proportion of admixed individuals vary across aggregations. Some individuals with intermediate admixture levels spawned in either season, suggesting that spawning time is not completely fixed. The analysis of mixture samples revealed the dynamic composition of aggregations outside of the breeding season and demonstrated the utility of the SNP-panels for mixed stock assessment. Altogether, these results contribute to the hypothesis that selection influences spatial distributions of genetic variation, highlighting the need to manage or conserve ecologically important adaptive variation in nature.

LIST OF ABBREVIATIONS USED

ABBREVIATION	DESCRIPTION
bp	Base pair
BQSR	Base quality score recalibration
CALM	Calmodulin
CNVs	Copy number variations
DFO	Department of Fisheries and Oceans Canada
DNA	Deoxyribonucleic acid
ESR2A	Estrogen receptor 2
F_{ST}	Fixation index
GB	Giga bytes
GNL	Gulf of St. Lawrence/Newfoundland/South Labrador
GO	Gene Ontology term
GSL	Gulf of St. Lawrence
hrWGR	High-coverage haplotype-resolved individual WGR
huWGR	High-coverage haplotype-unresolved individual WGR
INDELS	Insertions and deletions
IR	INDEL recalibration
Kb	Kilo bases
lcWGR	Low-coverage individual whole-genome resequencing
Mb	Mega bases
MYHC	Myosin heavy chain
NE	Northeast
NW	Northwest
Pool-seq	Whole-genome resequencing of pooled DNA
QC	Quality control
RAD-seq	Restriction site associate DNA sequencing
RAM	Random access memory
RDA	Redundancy analysis
RF	Random forest

RNA-seq	Sequencing of cDNA obtained from mRNA
RRS	Reduced Representation Sequencing
SBT	Sea bottom temperature
SD	Standard deviation
SFM	Scotian shelf/Bay of Fundy/Gulf of Maine
SFS	Site frequency spectrum
SLR-seq	Synthetic long-read sequencing
SMRT-seq	Single-molecule real-time sequencing
SNP	Single nucleotide polymorphism
SOX11	SOX11 transcription factor
SSS	Sea surface salinity
SST	Sea surface temperature
SV	Structural variants
TSHR	Thyroid-stimulating hormone receptor
VCF	Variant Call Format
WES	Whole-Exome sequencing
WGR	Whole genome resequencing

ACKNOWLEDGEMENTS

First and foremost, I would like to thank the members of my PhD committee, specially my supervisor, Daniel Ruzzante, for accepting me in his lab and for all his support and teachings during these years, and Paul Bentzen, Ian Bradbury, and Robert Beiko, for their enriching feedback; your guidance was instrumental for completing this dissertation. Thanks to Sean Rogers, my external examiner, for taking the time to be involved in my thesis defence. My deepest gratitude to the agencies that sponsored my graduate studies in Canada, the Killam Trust, the Vanier Canada Graduate Scholarship, the President's Award of Dalhousie University, the Nova Scotia Graduate Scholarship, the Natural Sciences and Engineering Research Council (NSERC) Strategic Grant awarded to my supervisor, and the Lett Fund; and to the agencies that funded this research, NSERC and the Department of Fisheries and Ocean Canada (DFO).

Completing this dissertation would not have been possible without the help of many people from near and far. Many thanks to those who provided fish and tissue samples, in particular to DFO staff Doug Swain, Hugues Benoît, Christina Bourne, Rabindra Singh, Caroline Vanier, Francois Gregorie, Alain Mallet, Derek Knox, Jennifer McDermid, Kim Emond, to fisherman Donald and Crystal Kent, and to Comeau's Seafood, Alma Duffy, and Clearwater seafood. Thanks to Greg McCracken for field, lab work support, and funny moments in the office, to Ian Patterson for advice with all about lab stuff, and to Connor Booker, Quentin Kerr, and James Kho for their help in DNA extraction. All my gratitude to our collaborators in the Andersson's Lab in Uppsala University, Sweden, professor Leif Andersson and researchers Sangeet Lamichhaney, Nima Rafati, Alvaro Martinez and Ulla, for willing to collaborate with us and for introducing me to the world of genomics and bioinformatics.

I also want to thank the many colleagues that generously devoted some of their time to help me figure out bioinformatics puzzles or made fixes to their computer packages so they could handle my large datasets, Michael Blum's Lab (*pcadapt* developer team), Mathieu Gautier (*poolfstat* developer team), Erik Garrison (*freebayes* developer), Thibaut Jombart and Zhian Kamvar of *adegenet* developer team, Sheila and Geraldine Van de Auwera of *GATK* support team of the Broad Institute, Gavin Douglas, Simon Fior, Hannes Svardal, Britta Meyer, Ann-Marie Waldvogel, Emmanuel Russet,

Arne Jacobs, Franscesco Del Grande, and to all the anonymous peers that solve questions in computer forums such Stackoverflow and R-bloggers. I owe you a ton.

Special thanks to Emma Sylvester, Sarah Lehnert, Nick Jeffery, and Ryan Stanley for their data analysis advice and generosity to share some code; thanks so much Ryan for obtaining the oceanographic data used in chapter 4! Thanks a lot to Compute Canada for providing the computing infrastructure required to complete my thesis, and to their staff, Rob Dickson (Acenet), Hui Zhong Lu, and Minh-Nghia Nguyen (Calcul Quebec) for their assistance using the cluster computers. Thanks to Claudio Slamovits, and to Morgan Languille and lab members Akilesh and André, for allowing me to use their lab equipment.

Thanks to my dear friends from around the world who gave me an ear, a hug, and a laugh when I needed it the most, fellow students at Dal: Sarah Toast, James, Quentin, Annie, Emilie R., Njal, Dave K., Rebekah O., Paul D., Nancy R., Myriam, Nadine, Christophe, Devon, Chris K., Kristina B., Marion, Scott M., Lea, and Ana C.; friends I met while in Uppsala, Sangeet, Nima, Neda, Dora, Anna, Sergei, Gosia, Andres C., Sergio T., and Martin; friends I met in Haligonia, Mirjam and Chris, Patricia and Bryan, "El combo latino": Laurie Baker, Ivan and Anahi, Camilo and Lauri, Francisco and Mariana, Jorge and Angela, Diego and Laura, Noelia and Axel, Sergio L., Daniela C., Daniela G., Dayana S.; and to my dear roomies Lisette, Francisca, Mailen, and Sergio. I cherish our moments together and feel so lucky to count you among my friends for life.

Finally, infinite thanks to all my blood and extended family, my father Luis, who would have loved seen me graduated, to my mom Edith, my sister Gloria, my grandma Oliva, my best friend Carolina Paredes, and to Cathy and John Campbell and family; there are not enough words to express my gratitude for your love, kindness, generosity and support, I am here because of you. Thanks Scott Campbell, for the love, laughter, and dance moves you bring to my life, thanks for your patience and unconditional support, I cannot wait to share time with you again.

CHAPTER 1. INTRODUCTION

1.1 Background

Species persistence over time relies on the capacity of populations to respond and adapt to shifting environments. Such capacity is largely determined by the amount of genetic variation existing among individuals within populations and between populations within a species (i.e. intraspecific variation). Intraspecific variation results from the interplay of four evolutionary forces: mutation, genetic drift, gene flow, and natural selection. Mutation, natural selection and genetic drift promote tend to increase genetic divergence among populations, whereas gene flow has a homogenizing effect counteracting natural selection and local adaptation by spreading new genetic variants between populations (Slatkin, 1987). Yet, it is not well understood how these evolutionary forces shape the contemporary distribution of genetic variation throughout a species' geographic range. One of the main difficulties to address this matter resides in the overlapping signatures left in the genome by diverse evolutionary processes. For example, high differentiation in allele frequencies is often interpreted as a signature of natural selection; however, such pattern of genetic divergence can result from neutral evolutionary processes such as genetic drift leading to increased population structuring and/or allele surfing during range expansions (Excoffier, Foll, & Petit, 2009; Travis et al., 2007) as well as from divergent selection along an environmental gradient (Hoban et al., 2016). Another challenge is that, until the recent development of new sequencing technologies, it was not possible to assess neutral and adaptive genetic variation in a high genomic resolution using traditional genotyping methods (Allendorf, 2016). This was particularly limiting for the estimation of population structure in species characterized by high gene flow (Waples, 1998) or in which natural selection played an important role. Thus, with the increased capacity to assess neutral and adaptive genetic variation with an unprecedented marker density, we are now in a new era of discoveries that promises solving many of the long standing questions in evolutionary biology, such how population divergence can lead to new species, what the genetic basis of local adaptation is, or how population divergence arises in the sea.

Given the general perception that there are no evident physical barriers to gene flow in the ocean and the commonly observed high dispersal potential, marine species were usually assumed to be genetically homogeneous or minimally structured (Hauser & Carvalho, 2008; Palumbi, 1994). Recent genomic studies however, are challenging this view by revealing fine-scale structuring never seen before in marine organisms (Benestan et al., 2015; Bradbury et al., 2013; Martinez Barrio et al., 2016). From a conservation perspective, this discovery is relevant and has several practical implications (Conover, 1998). For example, fishery management practices that ignore species' biocomplexity (i.e. diversity of reproductive strategies or locally adapted populations), could risk the species' long-term persistence and of the economic activities derived from its harvesting (Ruzzante et al., 2006). The diversity of reproductive strategies and of locally adapted populations are some of the most crucial components of intraspecific diversity that should be protected in wild fish stocks. Such components are directly linked to adult reproductive success and offspring survival in stochastic environments, which ultimately determine the recovery capacity of populations to fishing pressure (Schindler et al., 2010). When biologically relevant components of a species remain undetected or untraceable and are not accounted for in fish stock delineation, then they become vulnerable to overfishing. For instance, uninformed fishing targeting mixed stocks could remove biologically relevant biodiversity, compromising the species' evolutionary potential (Frankham, 2010; Ruzzante et al., 2000). Therefore, in the case that reproductive strategies are genetically distinguishable, it would be ideal to have a genetic tool that allow the identification of reproductive components outside of the breeding season, which remains challenging using traditional methods. However, knowledge on the population structure and adaptive variation of most commercially harvested marine species remains limited.

Local adaptation arises when organisms have increased reproductive success in their local environment than elsewhere (Blanquart, Kaltz, Nuismer, & Gandon, 2013; Savolainen, Lascoux, & Merilä, 2013). Consequently, local adaptation plays an important role in generating and maintaining biological diversity (Gavrilets, 2003). Genome scans using modern sequencing techniques are helping to identify and characterize genetic regions involved in local adaptation, commonly identified by elevated allele frequency differences (Jones et al., 2012; Tavares et al., 2018). Nevertheless, the still high cost of sequencing many individuals, commonly required in population genomics studies, together with the scarcity of genomic resources (i.e. reference genome and gene annotations) available for non-model species, are restrictive. The study of local adaptation in the sea presents additional challenges. For instance, the genetic basis of adaptive traits remains largely unknown (Barrett & Hoekstra, 2011); it is difficult to disentangle genomic signatures of selection from signatures of demographic history (Hoban et al., 2016); and it is not well understood how population divergence and local adaptation arises in the presence of high gene flow (Feder, Egan, & Nosil, 2012; Tigano & Friesen, 2016). An alternative to expand our understanding of ecological adaptation in the sea, and overcome some of the limitations previously mentioned, is the study of widely distributed and abundant marine species. In species with extensive geographic distribution ranges, populations can be exposed to diverse ecological habitats and selective pressures, which opens up opportunities for local adaptation to take place (Yeaman & Whitlock, 2011). When, in addition, such species are very abundant and exhibit high levels of diversity (high effective sizes), the effect of genetic drift is negligible, meaning that patterns of population structure are likely due to natural selection.

Several attributes then, make Atlantic herring (*Clupea harengus* L.) an ideal candidate species for the study of population divergence and adaptation in the sea: *i*) it is a highly abundant (average population size estimated to the order of 10^6, Martinez Barrio et al., (2016)) marine schooling pelagic fish, implying the role played by genetic drift in shaping patrons of population divergence is minor; *ii*) it is highly migratory and a broadcast spawner, for which its populations may be close to random mating; *iii*) it is widely distributed throughout diverse environments across the North Atlantic Ocean (including open ocean and the brackish waters of the Baltic Sea), which leaves opportunities for ecological adaptation; and *iv*) it has available an annotated reference genome that facilitates the identification and characterization of the genomic basis of local adaptation (Martinez Barrio et al., 2016).

The life history of herring is recognized by its high level of complexity and plasticity (McQuinn, 1997; Ruzzante et al., 2006; Geffen, 2009; Stephenson et al., 2009), especially in its reproductive system. In the Northwest (NW) Atlantic spawning takes place near the coast and in offshore banks from Cape Cod to northern Newfoundland, in predictable times and locations (Iles & Sinclair, 1982). Spawning occurs from April to October, mainly in spring (April to June) and fall (1st July to October) (Leblanc et al., 2010), with a higher abundance of spring spawners in the north and fall spawners towards the south of the range, and a coexistence of both strategies in the Gulf of St. Lawrence (Melvin, Stephenson, & Power, 2009). Mature individuals (3-4 years old) spawn in schools once a year. Tagging data indicate that herring often return to the spawning ground they have used previously; however it is still not clear whether or not herring actually show natal homing (Geffen, 2009; Ian H. McQuinn, 1997; Melvin et al., 2009). Large females deposit around 360,000 eggs on gravel or rocks, which stay on the bottom until hatching (Messieh, 1988). Time for hatching varies depending of the spawning season and the geographic location of the spawning ground. For example, in the Gulf of St. Lawrence, where the two spawning types coexist, eggs deposited by spring spawners hatch after 30 days at 5°C, while eggs released by fall spawners hatch after 10 days at 15°C. In Nova Scotia, where fall spawners dominate, eggs of fall spawners hatch in 11 days at 10°C (Jean, 1956 cited at Scott & Scott, 1988). One can infer from this that special adaptations should have evolved for the survival of eggs in these two contrasting environmental conditions. Larvae remain aggregated and dispersion is influenced by oceanographic conditions, food concentrations and light intensity (Geffen, 2009). Juveniles and adults migrate annually among spawning, overwintering, and feeding areas. Adult overwintering and feeding aggregations generally consist of individuals of mixed origin.

Atlantic herring is a key forage species in the marine ecosystem, feeding on plankton and being prey of numerous marine fishes, birds and mammals. Additionally, a profitable fishery is sustained by this species throughout the North Atlantic (FAO, 2019). In the last century though, some fish stocks have experienced significant collapse and others, signs of recovery, urging the implementation of better management practices that protect the biological complexity of the species. However, the resolution of the population structure of Atlantic herring remains challenging due to its high biocomplexity (Iles & Sinclair, 1982; Ruzzante et al., 2006).

Multiple attempts have been made to resolve the population structure of herring using a variety of genetic markers and at different spatial scales, mostly in the northeast (NE) Atlantic. These studies have commonly reported low levels of population differentiation at neutral loci (Andersson, Ryman, Rosenberg, & Ståhl, 1981; André et al., 2011; Jorgensen, Hansen, Bekkevold, Ruzzante, & Loeschcke, 2005). Most recently, by screening thousands of single nucleotide polymorphisms (SNPs) using novel sequencing approaches, significant genetic differentiation was detected at putatively adaptive loci that appear to respond to environmental gradients (Guo, Li, & Merilä, 2016; Lamichhaney et al., 2012; Limborg et al., 2012). Additionally, with the recent publication of a reference genome sequence for the Atlantic herring (Martinez Barrio et al., 2016), it is now possible to assess genetic variation at millions of SNPs, this being a major advance in the possibility to study the genetic basis of ecological adaptation in this species.

In this thesis I assess neutral and adaptive genetic variation at the whole genome level in spawning herring aggregations throughout the NW Atlantic. Specifically, I addressed the following questions:

i) What are the spatial scale and pattern of population structuring in herring?

ii) What is the genetic basis of such structuring?

iii) What is the potential functional effect of variant sites underlying population divergence?

iv) Which evolutionary processes and environmental variables are associated with population structure patterns?

v) Is it possible to develop a highly informative and cost-effective genetic tool for the diagnosis of adaptive components in aggregations of presumed mixed origin?

Ultimately, my work aims to contribute to an increased understanding of the patterns, genetic basis, and mechanisms underpinning population divergence in the sea, which ideally, would help in the implementation of effective conservation and sustainable fisheries management practices.

1.2 Thesis structure

My thesis consists of six chapters. Chapter 1 is the present introduction. In Chapter 2, I present a comprehensive review of whole-genome resequencing approaches, discuss their advantages and limitations, and provide recommendations for their application in conservation biology. In Chapter 3, I compare whole-genome resequencing data of pooled DNA of individuals (Pool-seq) of Atlantic herring populations. In total, 6 from the northwest (NW) and 19 from the northeast (NE) Atlantic Ocean were examined to elucidate shared and unique genomic patterns of differentiation between the spring and fall reproductive components, and among geographically close and distant populations. In Chapter 4, I focus on the northwest Atlantic region and analyze Pool-seq data of 14 spawning aggregations distributed across the reproductive range of the species in North America to investigate fine-scale patterns of genomic divergence, their genomic basis, and their association with oceanographic variables. Exploring the practical applications of the findings of previous chapters, in Chapter 5, I developed two highly informative and reduced single nucleotide polymorphisms (SNPs) panels for NW Atlantic herring to examine spatial and temporal variation in allele frequencies. I genotyped 993 individuals collected from 30 locations, including spawning aggregations and inshore and offshore mixed aggregations. Finally, in Chapter 6, I discuss the general implications of my work and present some general conclusions. Specifically, I discuss the implications of my work for fisheries management as well as its limitations and some future potential directions.

1.3 Statement of co-authorship

This thesis consists of a critical literature review and three data chapters, each of them corresponding to a manuscript written for publication in a scientific journal. All co-authors contributed to these manuscripts by providing crucial tissue samples or contributing in funding, experimental design, data analysis, interpretation of results, and writing of manuscripts. The publication status of each chapter is as follows:

Chapter 2:
This chapter was published as "**Fuentes-Pardo, A.P.** and D.E. Ruzzante. Whole-genome sequencing approaches for conservation biology: advantages, limitations, and practical recommendations" in Molecular Ecology. 2017;26:5369–5406.
https://doi.org/10.1111/mec.14264

Chapter 3:
This chapter was published as "Lamichhaney, S.*, **Fuentes-Pardo, A.P.***, Rafati, N., Ryman, N., McCracken, G.R., Bourne, C., Rabindra, S., Ruzzante, D.E. and L. Andersson. Parallel adaptive evolution of geographically distant herring populations on both sides of the North Atlantic Ocean" in PNAS. 2017;26:5369–5406.
https://doi.org/10.1073/pnas.1617728114
* equal contribution

Chapter 4:
This chapter was submitted for publication as "**Fuentes-Pardo, A. P.**, Bourne, C., Rabindra, S., Emond, K., Pinkham, L., McDermid, J. L., Andersson, L. and D. E. Ruzzante. Adaptation to seasonal reproduction and thermal minima-related factors drives fine-scale divergence despite gene flow in Atlantic herring populations" to Molecular Ecology on March 2019, it is under revision and it is currently available as a preprint in BioRxiv, https://doi.org/10.1101/578484

Chapter 5:
This chapter is a manuscript and it is not published yet.

1.4 References

Allendorf, F. W. (2016). Genetics and the conservation of natural populations: Allozymes to genomes. *Molecular Ecology*, *38*(1), 42–49. https://doi.org/10.1111/mec.13948

Andersson, L., Ryman, N., Rosenberg, R., & Ståhl, G. (1981). Genetic variability in Atlantic herring (Clupea harengus harengus): description of protein loci and population data. *Hereditas*, *95*(1), 69–78. https://doi.org/10.1111/j.1601-5223.1981.tb01330.x

André, C., Larsson, L. C., Laikre, L., Bekkevold, D., Brigham, J., Carvalho, G. R., … Ryman, N. (2011). Detecting population structure in a high gene-flow species, Atlantic herring (Clupea harengus): direct, simultaneous evaluation of neutral vs putatively selected loci. *Heredity*, *106*(2), 270–280. https://doi.org/10.1038/hdy.2010.71

Barrett, R. D. H., & Hoekstra, H. E. (2011). Molecular spandrels: tests of adaptation at the genetic level. *Nature Review Genetics*, *12*(11), 767–780. https://doi.org/10.1038/nrg3015

Benestan, L., Gosselin, T., Perrier, C., Sainte-Marie, B., Rochette, R., & Bernatchez, L. (2015). RAD genotyping reveals fine-scale genetic structuring and provides powerful population assignment in a widely distributed marine species, the American lobster (Homarus americanus). *Molecular Ecology*, *24*(13), 3299–3315. https://doi.org/10.1111/mec.13245

Blanquart, F., Kaltz, O., Nuismer, S. L., & Gandon, S. (2013). A practical guide to measuring local adaptation. *Ecology Letters*, *16*(9), 1195–1205. https://doi.org/10.1111/ele.12150

Bradbury, I. R., Hubert, S., Higgins, B., Bowman, S., Borza, T., Paterson, I. G., … Bentzen, P. (2013). Genomic islands of divergence and their consequences for the resolution of spatial structure in an exploited marine fish. *Evolutionary Applications*, *6*(3), 450–461. https://doi.org/10.1111/eva.12026

Conover, D. O. (1998). Local adaptation in marine fishes -evidence and implications for stock enhancement. *Bulletin of Marine Science*, *62*(2), 477–493.

Excoffier, L., Foll, M., & Petit, R. J. (2009). Genetic Consequences of Range Expansions. *Annual Review of Ecology, Evolution, and Systematics*, *40*(1), 481–501. https://doi.org/10.1146/annurev.ecolsys.39.110707.173414

FAO. (2019). Species Fact Sheets: Clupea harengus (Linnaeus, 1758). Retrieved from http://www.fao.org/fishery/species/2886/en

Feder, J. L., Egan, S. P., & Nosil, P. (2012). The genomics of speciation-with-gene-flow. *Trends in Genetics*, *28*(7), 342–350. https://doi.org/10.1016/j.tig.2012.03.009

Frankham, R. (2010). Challenges and opportunities of genetic approaches to biological conservation. *Biological Conservation*, *143*(9), 1919–1927. https://doi.org/10.1016/j.biocon.2010.05.011

Gavrilets, S. (2003). Perspective: Models of speciation: what have we learned in 40 years? *Evolution*, *57*(10), 2197–2215. https://doi.org/10.1111/j.0014-3820.2003.tb00233.x

Geffen, A. J. (2009). Advances in herring biology: from simple to complex, coping with plasticity and adaptability. *ICES Journal of Marine Science*, *66*(8), 1688–1695. https://doi.org/10.1093/icesjms/fsp028

Guo, B., Li, Z., & Merilä, J. (2016). Population genomic evidence for adaptive differentiation in the Baltic Sea herring. *Molecular Ecology*, *25*(12), 2833–2852. https://doi.org/10.1111/mec.13657

Hauser, L., & Carvalho, G. R. (2008). Paradigm shifts in marine fisheries genetics: ugly hypotheses slain by beautiful facts. *Fish and Fisheries*, *9*(4), 333–362. https://doi.org/10.1111/j.1467-2979.2008.00299.x

Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., … Whitlock, M. C. (2016). Finding the Genomic Basis of Local Adaptation: Pitfalls, Practical Solutions, and Future Directions. *The American Naturalist*, *188*(4), 379–397. https://doi.org/10.1086/688018

Iles, T. D., & Sinclair, M. (1982). Atlantic Herring: Stock Discreteness and Abundance. *Science*, *215*(4533), 627–633. https://doi.org/10.1126/science.215.4533.627

Jean, Y. (1956). A Study of Spring and Fall Spawning Herring (Clupea Harengus L.) at Grande-Rivière, Bay of Chaleur, Québec. *Department of Fisheries Québec Constribution*, *49*, 76p.

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., … Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, *484*(7392), 55–61. https://doi.org/10.1038/nature10944

Jorgensen, H. B. H., Hansen, M. M., Bekkevold, D., Ruzzante, D. E., & Loeschcke, V. (2005). Marine landscapes and population genetic structure of herring (Clupea harengus L.) in the Baltic Sea. *Molecular Ecology*, *14*(10), 3219–3234. https://doi.org/10.1111/j.1365-294X.2005.02658.x

Lamichhaney, S., Barrio, A. M., Rafati, N., Sundstrom, G., Rubin, C.-J., Gilbert, E. R., … Andersson, L. (2012). Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. *Proceedings of the National Academy of Sciences*, *109*(47), 19345–19350. https://doi.org/10.1073/pnas.1216128109

Leblanc, C., Swain, D., MacDougall, C., & Bourque, C. (2010). Assessment of the NAFO Division 4T southern Gulf of St. Lawrence herring stocks in 2009. *DFO Canadian Science Advisory Secretariat*, (Research Document 2010/059), 143 p.

Limborg, M. T., Helyar, S., De Bruyn, M., Taylor, M. I., Nielsen, E. E., Ogden, R., … Bekkevold, D. (2012). Environmental selection on transcriptome-derived SNPs in a high gene flow marine fish, the Atlantic herring (Clupea harengus). *Molecular Ecology*, *21*(15), 3686–3703. https://doi.org/10.1111/j.1365-294X.2012.05639.x

Martinez Barrio, A., Lamichhaney, S., Fan, G., Rafati, N., Pettersson, M., Zhang, H., … Andersson, L. (2016). The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. *ELife*, *5*(MAY2016), 1–32.

https://doi.org/10.7554/eLife.12081

McQuinn, I. H. (1997). Metapopulations and the Atlantic herring. *Reviews in Fish Biology and Fisheries*, *7*(3), 297–329. https://doi.org/10.1023/A:1018491828875

Melvin, G. D., Stephenson, R. L., & Power, M. J. (2009). Oscillating reproductive strategies of herring in the western Atlantic in response to changing environmental conditions. *ICES Journal of Marine Science*, *66*(8), 1784–1792. https://doi.org/10.1093/icesjms/fsp173

Messieh, S. N. (1988). Spawning of Atlantic Herring in the Gulf of St. Lawrence. *American Fisheries Society Symposium*, *5*, 31–48.

Palumbi, S. R. (1994). Genetic Divergence, Reproductive Isolation, and Marine Speciation. *Annual Review of Ecology and Systematics*, *25*(1), 547–572. https://doi.org/10.1146/annurev.es.25.110194.002555

Ruzzante, D. E., Mariani, S., Bekkevold, D., André, C., Mosegaard, H., Clausen, L. A. W., … Carvalho, G. R. (2006). Biocomplexity in a highly migratory pelagic marine fish, Atlantic herring. *Proceedings of the Royal Society B: Biological Sciences*, *273*(1593), 1459–1464. https://doi.org/10.1098/rspb.2005.3463

Ruzzante, D. E., Taggart, C. T., Lang, S., Cook, D., Applications, E., & Aug, N. (2000). Mixed-stock analysis of Atlantic cod near the Gulf of St. Lawrence based on microsatellite DNA. *Ecological Applications*, *10*(4), 1090–1109.

Savolainen, O., Lascoux, M., & Merilä, J. (2013). Ecological genomics of local adaptation. *Nature Reviews. Genetics*, *14*(11), 807–820. https://doi.org/10.1038/nrg3522

Schindler, D. E., Hilborn, R., Chasco, B., Boatright, C. P., Quinn, T. P., Rogers, L. A., & Webster, M. S. (2010). Population diversity and the portfolio effect in an exploited species. *Nature*, *465*(7298), 609–612. https://doi.org/10.1038/nature09060

Scott, W. B., & Scott, M. G. (1988). *Atlantic fishes of Canada. Canadian Bulletin of Fisheries and Aquatic Sciences, bulletin 219*. Toronto, CA: University of Toronto Press.

Slatkin, M. (1987). Gene flow and the geographic structure of natural populations. *Science*, *236*(4803), 787–792. https://doi.org/10.1126/science.3576198

Stephenson, R. L., Melvin, G. D., & Power, M. J. (2009). Population integrity and connectivity in Northwest Atlantic herring: a review of assumptions and evidence. *ICES Journal of Marine Science*, *66*(8), 1733–1739. https://doi.org/10.1093/icesjms/fsp189

Tavares, H., Whibley, A., Field, D. L., Bradley, D., Couchman, M., Copsey, L., … Coen, E. (2018). Selection and gene flow shape genomic islands that control floral guides. *Proceedings of the National Academy of Sciences*, *115*(43), 11006–11011. https://doi.org/10.1073/pnas.1801832115

Tigano, A., & Friesen, V. L. (2016). Genomics of local adaptation with gene flow. *Molecular Ecology*, n/a-n/a. https://doi.org/10.1111/mec.13606

Travis, J. M. J., Munkemuller, T., Burton, O. J., Best, A., Dytham, C., & Johst, K. (2007). Deleterious Mutations Can Surf to High Densities on the Wave Front of an Expanding Population. *Molecular Biology and Evolution*, *24*(10), 2334–2343. https://doi.org/10.1093/molbev/msm167

Waples, R. S. (1998). Separating the wheat from the chaff: patterns of genetic differentiation in high gene flow species. *Journal of Heredity*, *89*(5), 438–450. https://doi.org/10.1093/jhered/89.5.438

Yeaman, S., & Whitlock, M. C. (2011). The genetic architecture of adaptation under migration-selection balance. *Evolution*, *65*(7), 1897–1911.

CHAPTER 2. WHOLE-GENOME SEQUENCING APPROACHES FOR CONSERVATION BIOLOGY: ADVANTAGES, LIMITATIONS, AND PRACTICAL RECOMMENDATIONS

2.1 Abstract

Whole-genome resequencing (WGR) is a powerful method for addressing fundamental evolutionary biology questions that have not been fully resolved using traditional methods. WGR includes four approaches: the sequencing of individuals to a high depth of coverage with either unresolved (huWGR) or resolved haplotypes (hrWGR), the sequencing of population genomes to a high depth by mixing equimolar amounts of unlabelled-individual DNA (Pool-seq), and the sequencing of multiple individuals from a population to a low depth (lcWGR). These techniques require the availability of a reference genome. This, along with the still high cost of shotgun sequencing and the large demand for computing resources and storage, has limited their implementation in non-model species with scarce genomic resources and in fields such as conservation biology. Our goal here is to describe the various WGR methods, their pros and cons, and potential applications in conservation biology. WGR offers an unprecedented marker density and surveys a wide diversity of genetic variations not limited to single nucleotide polymorphisms (e.g. structural variants and mutations in regulatory elements), increasing their power for the detection of signatures of selection and local adaptation as well as for the identification of the genetic basis of phenotypic traits and diseases. Currently though, no single WGR approach fulfills all requirements of conservation genetics, and each method has its own limitations and sources of potential bias. We discuss proposed ways to minimize such biases. We envision a not distant future where the analysis of whole genomes becomes a routine task in many non-model species and fields including conservation biology.

2.2 Introduction

Over the last 40 years, genetics has emerged as an important tool in the conservation of threatened species. Based on the analysis of genetic variation of individuals and populations, genetics has provided insights on diverse areas in conservation biology including species identification, hybridization, kinship, evolutionary history, effective population size (N_e), population substructure, population connectivity, adaptive genetic variation, local adaptation, and inbreeding (Haig et al., 2016; Hedrick & Miller, 1992; von der Heyden et al., 2014).

Genetic variation has traditionally been examined using from a single to a handful (12-24) of molecular markers including allozymes, mitochondrial DNA, and microsatellites (see review by Allendorf (2017)). Most of these markers target a few neutral positions in the genome, limiting the ability to estimate genome-wide parameters (Avise, 2010). The development of high-throughput sequencing (HTS) technologies over a decade ago revolutionized the way genetic variation is assessed (Goodwin, McPherson, & McCombie, 2016). These technologies allow the massive sequencing of thousands to millions of loci in a short time for an affordable cost, resulting in a much higher marker density than experienced with past technologies. Today, individual research groups have the option of sequencing the reference genome of their focal species and re-sequencing genomes of individuals and populations for the detection of both, neutral and adaptive variation (Ellegren, 2014). The extraordinary increase in number of markers available with genomic approaches has sparked much expectation within the conservation community, reflected in several recent review papers on this topic (Allendorf, Hohenlohe, & Luikart, 2010; Angeloni, Wagemaker, Vergeer, & Ouborg, 2012; Avise, 2010; L. M. Benestan et al., 2016; Frankham, 2010; Funk, McKay, Hohenlohe, & Allendorf, 2012; Garner et al., 2016; McMahon, Teeling, & Höglund, 2014; Ouborg, Pertoldi, Loeschcke, Bijlsma, & Hedrick, 2010; Primmer, 2009; Shafer et al., 2015; Steiner, Putnam, Hoeck, & Ryder, 2013). The hype is a reflection of the promise of increased statistical power in population genetics tests, but most importantly, of the possibility of addressing long standing questions in conservation biology not fully resolved with traditional methods. Some of these questions are: What is the phylogenetic relationship between unresolved taxa? What are the loci responsible for speciation, for local adaptation, for interactions among species, or for inbreeding depression? What is the genetic basis of traits related to fitness? (Allendorf et al., 2010; McMahon et al., 2014; Ouborg et al., 2010).

The advances achieved with HTS promise an exciting time for genomics-based research, although these developments have their own limitations. For example, short-read sequences (~100 base pairs (bp) long), that are commonly obtained with current sequencing technologies, are problematic for genome assembly and detection of large structural variants. The relatively high error rate of existing sequencing platforms makes it necessary to obtain high depth of coverage for the correct identification of variants (Goodwin et al., 2016) and sequencing cost is still high for population studies that require analysing multiple individuals. Currently, some alternatives to overcome the cost limitation are: (*i*) using reduced-representation sequencing (RRS) methods that screen a fraction of the genome (da Fonseca et al., 2016), (*ii*) obtaining whole-genome resequencing (WGR) data from pooled DNA of individuals per population to a high coverage [known as Pool-seq, (Schlötterer, Tobler, Kofler, & Nolte, 2014)], or (*iii*) low-coverage WGR data of individuals from a population [known as lcWGR, (Nielsen, Paul, Albrechtsen, & Song, 2011)]. These approaches have successfully screened multiple loci genome-wide in several species, and have been instrumental in addressing a variety of questions in molecular ecology (Foote et al., 2016; Hohenlohe et al., 2010; S. Lamichhaney et al., 2017). These methods however, have their own restrictions and sources of bias and error that should be minimized for the correct inference of population parameters (Anderson, Skaug, & Barshis, 2014; Lowry et al., 2017a).

Today, when the genome of virtually any species can be sequenced, it is pertinent to ask, when is the analysis of whole-genome data justified in conservation biology? What are the limitations of current WGR methods, and how could they be overcome? These questions are particularly important for three main reasons: *1*) traditional molecular methods can solve some of the questions in conservation for a small fraction of the cost and effort relative to genomic approaches (e.g. dozens of polymorphic microsatellites generate acceptable estimates of population structure, gene flow, N_e, kinship) (Allendorf, 2016; McMahon et al., 2014); *2*) RRS methods generate thousands of molecular markers genome-wide, increasing the power of statistical tests for a lower cost compared to whole-genome approaches (Andrews, Good, Miller, Luikart, &

Hohenlohe, 2016); *3*) current short-read sequence data presents some restrictions that limit the kind of analysis that can be performed (Goodwin et al., 2016).

Given the increased interest in the use of genomics in conservation biology, in this review we first provide a general background on sequencing technologies and whole-genome sequencing (Box 1). We then describe the various WGR approaches used in population genomics (Box 2), discuss their limitations and potential solutions (Box 3 and 4), and compare WGR to RRS methods (Table 2.1). We also discuss limitations of genome scans for detecting selection and inferring adaptation from genomic data (Box 5). We subsequently present case studies for the areas of conservation biology that can in principle be benefited by WGR analysis (Table 2.2). Finally, we provide guidelines for choosing between RRS and WGR methods depending on the type of genetic variation of interest and the expected haplotype block size, and explore recent innovations that promise overcoming the limitations of current methodologies.

2.3 Genome sequencing techniques

The improved understanding of the complexity of the genome architecture during the Human Genome Project (Human Genome Research Institute (NIH), https://www.genome.gov/12011239/), coupled with advances in molecular techniques and equipment set the basis for the beginning of the 'genomic era' in the last two decades. The development of sequencing technologies in particular, has revolutionized the way we examine and comprehend the genome. Three major sequencing generations have taken place thus far. Sanger-sequencing (or 'chain-termination method'), considered the first generation, was introduced in 1977 (Sanger, Nicklen, & Coulson, 1977). This method provides high per-base accuracy (99,999%, Shendure & Ji (2008)) and medium-read length (~1000 bp) but has the main limitations of low-throughput and relatively high cost per base. Genome assembly is achieved via sequencing of bacterial artificial chromosome libraries containing pieces of the whole-genome. Using specialized software, the sequence of each fragment is assembled into a contiguous sequence (NIH, https://www.genome.gov/12011239/). The first genome sequences of several model species (e.g. yeast, *Drosophila melanogaster, Caenorhabditis elegans*, *Arabidopsis* *thaliana*) including humans were obtained with this technology, which incidentally, also led to the expansion of genetics research overall (Goodwin et al., 2016; Heather & Chain, 2016; Pettersson, Lundeberg, & Ahmadian, 2009).

The second-generation of sequencing technology, which is based on 'sequencing-by-synthesis' and innovative high-throughput systems (i.e. 454-pyrosequencing, Illumina, Ion-Proton), appeared between 2005 and 2010. These technologies have a higher error rate (accuracy >99.5%) and produce shorter sequences (75-300 bp Illumina, <400 bp Ion-Proton, <700 bp 454-pyrosequencing) than Sanger-sequencing, but the massive parallel sequencing of fragments of sheared DNA significantly increased throughput. *De novo* genome assembly is achieved with data of paired-end short-reads (~100 bp) of various libraries with different insert size (350 bp to 40 Kilobases - Kb) to maximize genome coverage. The consensus sequence results from the computational assembly of short-reads. First, contigs [i.e. sequences resulting from the joint of overlapping smaller sequences with no gaps or runs of more than 10 ambiguous bases (Ns)] and scaffolds [i.e. larger sequences formed by joining contigs with no sequence overlap but gaps can be present] (https://www.ncbi.nlm.nih.gov/assembly/agp/AGP_Specification/) are obtained, and these are subsequently linked and ordered with data of long-insert-size short-read libraries and long-reads. Short-reads however, are problematic for the assembly of repetitive sequences; interruptions (gaps) are thus common in the consensus sequence. Similarly, the detection of large structural variants (SVs) and the ability to estimate haplotypes (i.e. 'haplotype phasing') are limited due to the short and fragmented nature of these sequences. The main contributions of second-generation technologies have been the substantial drop in sequencing cost and the exponential increase in throughput, unlocking genome-, exome-, transcriptome-, and epigenome- sequencing approaches to non-model organisms and in doing so, revolutionizing medicine, agriculture, and biological research (Goodwin et al., 2016; Heather & Chain, 2016; Pettersson et al., 2009).

The third-generation appeared between 2011 and 2014 with sequencing technologies that produce reads of unprecedented length (average ~2–10Kb). Currently, long-reads can be obtained with two methods, 'single-molecule real-time' sequencing

(SMRT-seq) [i.e. Pacific Biosystems (PacBio), Oxford Nanopore Technologies (ONT)] and 'synthetic long-read' sequencing (SLR-seq) [i.e. Illumina synthetic long-reads, 10X Genomics]. SMRT-seq methods produce long-reads of single DNA molecules, whereas the synthetic approaches do not; in the latter, long sequences are computationally assembled from barcoded short-reads coming from the same DNA molecule (Bleidorn, 2016; Goodwin et al., 2016; Lee et al., 2016). Throughput, error rate, and cost vary between long-read approaches. For example, within SMRT-seq methods, throughput of PacBio is lower than any second-generation technique but nanopore's sequencing competes with Illumina HiSeqX. In contrast, the throughput of SLR-seq methods is the same as in Illumina systems. Error rate is much higher in SMRT-seq methods than in any second-generation technique (15-20% in PacBio, although 99,99% accuracy can be achieved with ~50x coverage (Berlin et al., 2015); 30-40% in nanopore-sequencers). In SLR-seq, error rate is the same as in Illumina. In terms of cost, PacBio is pricey (~USD$ 1000/Giga byte (GB), (Goodwin et al., 2016)) and nanopore-sequencing promises low cost (~USD$ 20/hour) though it has not yet been disclosed (Bleidorn, 2016). In SLR-seq methods, cost can be high as it includes short-read Illumina sequencing to very high coverage (~1000x (Lee et al., 2016)) and library preparation that incorporates barcodes. In 10X Genomics additional equipment is required. Genome assembly can be achieved with only long-read data (>50x), or with a combination of long- and short-reads. Long-reads help resolve complex stretches of the genome (i.e. repetitive sequences and SVs) that are poorly handled by short-reads, significantly improving quality of genome assembly. With long-reads it is also possible to sequence entire transcripts, enhance metabarcoding and metagenomics, and perform direct haplotype assignment (i.e. phasing) of genomes (Bleidorn, 2016) that otherwise can only be limitedly inferred from population-level short-read data (Snyder, Adey, Kitzman, & Shendure, 2015). A haplotype refers to groups of genetic variants that are located in the same chromosome (Snyder et al., 2015)). The composition and extension of haplotypes constitute valuable information for many analyses including, demographic history, linkage disequilibrium, GWAS, genealogical tracing of mutations, and allele-specific expression, among others (Browning & Browning, 2011; Lee et al., 2016; Snyder et al., 2015). For a more detailed description of sequencing techniques, see Goodwin *et al.* (2016); Lee et al. (2016); Bleidorn (2016).

2.3.1 Comparison of whole-genome resequencing and de novo sequencing

Whole-genome sequencing can be classified in two categories (Fig. 2.1): (*1*) *de novo* whole-genome sequencing (WGS); and (*2*) whole-genome resequencing (WGR). (*1*) The goal of WGS is the assembly of a genome sequence for the first time. This can be a demanding task depending on size and genome complexity, desired level of completeness, computing resources, and bioinformatics experience. Programming skills and understanding of assembly algorithms are fundamental for optimal results (Ekblom & Wolf, 2014). (*2*) The objective of WGR is instead, to compare genomic variability among individuals or populations. This approach requires previous availability of the reference genome for read mapping and variant identification. Sequencing is dedicated to obtaining reads from the genome of individuals or populations to a particular coverage depending on the application.

The absence of the reference genome of the species of interest likely constitutes the main limitation faced by conservation geneticists when implementing a WGR approach, justifying the use of the genome sequence of a closely-related species (Dennenmoser, Vamosi, Nolte, & Rogers, 2017; Lamichhaney et al., 2012). Caution is however warranted with this procedure as differences in genomic organization (e.g. copy number variation, structural variants) can exist, even between closely related species (Ekblom & Wolf, 2014). The use of a reference genome of another species restricts the mapping of short reads to conserved regions between the two taxa. The power of WGR could be diminished as potentially informative variation present uniquely in the focal species is likely to be missed following this procedure. Additionally, the genomic differences between taxa can affect the accuracy of both, read mapping and SNP calling (Nevado, Ramos-Onsins, & Perez-Enciso, 2014). Thus, when possible, it is preferable to use the genome of the focal species for WGR analysis, unless the research question can be addressed examining conserved regions alone. A brief overview of genome assembly guidelines and completeness status of genomes sequenced to date is provided in Box 1.

The steady decrease in sequencing cost promised by new technologies suggests that access to the genome sequence may in the near future no longer be an unsurmountable obstacle for non-model species (Goodwin et al., 2016). Prove of this are the multiple international initiatives that are collaboratively sequencing genomes of various taxa including fungi (Grigoriev et al., 2014), invertebrates (GIGA, 2014), arthropods (Evans et al., 2013), birds (Zhang, 2015), fishes (Malmstrøm, Matschiner, Tørresen, Jakobsen, & Jentoft, 2017; The FAASG Consortium, 2016), mammals (Fontanesi et al., 2016), vertebrates (Koepfli, Paten, Genome 10K Community of Scientists, & O'Brien, 2015), among others.

2.3.2 Comparison of whole-genome resequencing approaches for population genomics

A population genomics study can be based on the analysis of individual sequences (individual-based approaches) or on the analysis of the sequences of a group of individuals as a whole (population-based approaches). In individual-based approaches the goal is obtaining high quality individual genotypes, required for analysis on population demographic history and N_e estimation, and genealogical tracing of mutations, among others. There are currently two techniques: (*i*) high-coverage haplotype-unresolved individual WGR (huWGR) and (*ii*) high-coverage haplotype-resolved individual WGR (hrWGR). In both techniques, high read depth (>30-50x depth) is targeted for achieving accurate SNP, short INDEL (>50bp), and genotype calling, as multiple reads (observations) help distinguish true variation from sequencing error (Nagasaki et al., 2015). (*i*) In huWGR, short-read data per individual results in unphased individual genotypes that are used for subsequent analyses. If haplotype information is required, phasing can be indirectly achieved using statistical methods that rely on genotypes of several related or unrelated individuals. Such methods are then limited by the need for large sample sizes and by the extend of linkage disequilibrium blocks that vary across the genome (Browning & Browning, 2011). (*ii*) In hrWGR, the goal is to directly obtain haplotype-resolved genomes of single individuals using specific experimental procedures and short- and/or long-read sequencing, implying large sample sizes are not required (reviewed in Snyder *et al.* 2015).

In population-based approaches, the goal is obtaining population-level genomic data (e.g. allele frequencies or genotype likelihoods) from several individuals analyzed as a whole and sequenced to a high or a low depth (Buerkle & Gompert, 2013), to offset the cost of obtaining high-coverage per individual. Such population-level data can be used for inference of population structuring, detection of outlier loci and signatures of selection, among others. Two methods can be identified (Fig. 2.2): (*i*) Pool-seq, or the sequencing at a high coverage (>50x) of pooled DNA in equimolar concentration of unlabelled individuals from a population (Futschik & Schlötterer, 2010; Schlötterer et al., 2014), and (*ii*) lcWGR, or low-coverage individual whole-genome resequencing of multiple barcoded individuals from a population (~2-4x per individual) (Durbin et al., 2010; Nielsen et al., 2011). The general workflow for data acquisition with these approaches is presented in Box 2, and a comparison of requirements, technical aspects, and expected outcomes is shown in Table 2.1.

These two methods have several pros and cons. The main advantage of Pool-seq is the cost reduction achieved from the preparation of a single sequencing library per pooled DNA instead of one library per individual. This allows using large sample sizes per population (Fig. 2.2). Also, pooling equal amounts of DNA of multiple individuals facilitates the sequencing of a few chromosomes several times, leading to an improvement in SNP allele frequency estimates (Ferretti, Ramos-Onsins, & Pérez-Enciso, 2013; Gautier et al., 2013; Schlötterer et al., 2014). In Pool-seq the detection of variant sites and the estimation of population-level allele frequencies per SNP are derived from the relative proportion of read counts of each allele within a pool (Fig. 2.2). Pool-seq has three main limitations: First, individual genotypes are missed after mixing DNA samples in a pool. This makes it impossible to track technical errors during library preparation. Second, allele frequency estimation is susceptible to multiple factors including uneven representation of individual DNA in a pool, and sequencing and mapping errors. Finally, rare alleles are likely to be underrepresented in this kind of datasets, which can lead to a truncated distribution of allele frequencies or site frequency spectrum (SFS). Pool-seq data is thus mostly biased toward the detection of frequent and large-effect alleles (Cutler & Jensen, 2010; Raineri et al., 2012). Potential solutions to these limitations are discussed in Box 3. Obtaining a complete SFS is important in population genetics, as this metric synthesizes all the sequence variation at unlinked sites in a sample. Its shape varies with different evolutionary processes including bottlenecks or range expansions (Gutenkunst, Hernandez, Williamson, & Bustamante, 2009), and natural selection (Bustamante, Wakeley, Sawyer, & Hartl, 2001; Ronen, Udpa, Halperin, & Bafna, 2013), and is used to infer several metrics including Tajima's D and F_{ST} (Durrett, 2008; Han, Sinsheimer, & Novembre, 2015).

The main advantage of lcWGR is the low sequencing depth targeted per individual (~1-4x) that facilitates the analysis of a large number of samples per population, however one library needs to be prepared for each individual (Fig. 2.2). Individual library preparation is still cost restrictive for large sample sizes given the current high cost of commercial library preparation kits (Table 2.1). To overcome this limitation, the use of smaller reaction volumes and cheaper reagents has been advocated achieving a 6-10 times cost reduction per sample as shown for microbial genomes (<15 Mb) (Baym et al., 2015) and a teleost fish genome (~730Mb) (Therkildsen & Palumbi 2016). Despite the significant cost savings with this procedure, lcWGR is still slightly more expensive than Pool-seq for an equivalent sample size and sequencing depth (assuming 1x coverage per individual). The overall cost of Pool-seq and lcWGR is fairly equivalent, though, when ~50 individuals are included (~$USD 280 more in lcWGR as June 2017). Just like in Pool-seq, some of the disadvantages of lcWGR are that individual genotypes cannot be called: the low depth per individual impedes a reliable variant and genotype calling. Instead of read counts, this method detects variant sites and calculates genotype likelihoods (GLs) per site based on the accumulated sequence data of multiple individuals in a sample using a probabilistic framework that incorporates the uncertainty of the data due to sequencing, alignment and SNP calling errors. Based on GLs obtained across sites, a sample allele frequency per site is calculated from which other statistics are inferred, including the SFS (Korneliussen et al., 2014; Nielsen, Korneliussen, Albrechtsen, Li, & Wang, 2012; Nielsen et al., 2011). From the sample and population allele frequency likelihood, SNP and genotype calls can be obtained using a likelihood ratio test (Kim et al., 2011; Korneliussen et al., 2014). In theory, because SNP calling is avoided, all alleles present in a sample are considered in the GL calculation, resulting in the SFS potentially being less biased than in Pool-seq data. Another disadvantage of this method is that GLs values can vary depending on several factors, and currently a few programs accept GLs. Potential solutions to these limitations are discussed in Box 4.

2.3.3 Comparison of WGR with Reduced-Representation sequencing (RRS)

RRS is a general category of techniques that sequence a subset of the genome following different strategies. These techniques can be classified in three major groups: (*i*) RAD-seq (Restriction site Associate DNA sequencing) (Andrews et al., 2016), (*ii*) RNA-seq (sequencing of cDNA obtained from mRNA) (Ozsolak & Milos, 2011), and (*iii*) WES (Whole-Exome sequencing) (Warr et al., 2015).

(*i*) RAD-seq refers to a group of methods (e.g. traditional RAD, ddRAD, ezRAD, RAD-cap, among others) that evaluate the genetic variation present around restriction cut sites. The selection of restriction enzyme (frequent or rare cutter) ultimately determines the resulting marker density (i.e. number of loci sampled per physical genomic distance unit), making these methods flexible and customizable. These methods typically examine thousands of low-density genome-wide SNPs located in neutral and putatively functional loci that can be genotyped by sequencing in multiple individuals and populations for a relatively low cost (reviewed by Andrews et al. (2016)). (*ii*) RNA-seq focuses on genetic variants in parts of the genome that are being transcribed at the time of sampling. RNA-seq is thus mostly used for gene expression quantification but also for the comparison of variants at genes being transcribed in a particular time/tissue (reviewed by Ozsolak and Milos (2011)). (*iii*) WES explores genetic variants in exons of protein-coding genes using capture probes usually developed from a well annotated reference genome (reviewed by Warr et al. (2015)). A reference genome is however, not indispensable, since capture probes can also be designed from PCR products of targeted loci, from *de novo* assembly of RNA-seq transcriptomes or expressed sequence tags (ESTs), RAD-seq or WGR data, and from the genome of a closely-related species as functional elements are usually located in highly conserved regions (reviewed by Jones & Good (2016)).

The three RRS methods share the characteristic that they typically evaluate only a small fraction (~1–5 %) of the genome, which translates into reduced sequencing cost, computing resources and storage requirements compared to WGR approaches (Ozsolak & Milos, 2011; Warr et al., 2015). RRS techniques provide hundreds to thousands of genome-wide SNPs, that in RAD-seq, for example, are plenty for robust population genetics analysis based on neutral SNPs (Andrews et al., 2016). Functional SNPs can also be detected from RAD-seq data, although this depends on experimental design. The target marker density should ensure sampling of loci that contribute to a trait which relates to the average linkage disequilibrium decay (Gagnaire & Gaggiotti, 2016) (more on this is discussed in Box 5).

RNA-seq also can help in the identification of functional SNPs, although care needs to be taken during sampling, sequencing and SNP calling. This is because multiple factors can affect gene expression at the time of sampling, producing technical and biological variability that needs to be accounted for including replicates. Moreover, high read coverage is necessary to detect rare transcripts, and the intrinsic complexity of the transcriptome (e.g. alternative splicing) could make challenging read alignment for SNP detection (Conesa et al., 2016; Ozsolak & Milos, 2011).

WES constitutes a cost effective alternative to WGR for functional SNPs identification, as it screens protein-coding regions that in humans, for example, represent less than 2% of the total genome (a 100-times reduction of the amount of data for the same coverage) and contains 85% of mutations related to diseases in Mendelian disorders (Rabbani, Tekin, & Mahdieh, 2014). WGR is, however, more robust than WES for the detection of exome variants as it provides a more homogeneous sequence read coverage and a better sequencing quality overall (Belkadi et al., 2015). A potential problem with WES is that the exon capture/PCR amplification steps can produce low coverage (limiting variant detection) when probes are poorly designed (span exon boundaries) and fail to bind to the target region (Bi et al., 2012; Jones & Good, 2016).

Probably the most convenient advantage of RRS approaches is that they do not rely on a reference genome for SNP calling, since it can be accomplished by local assembly of reads. This has facilitated the broad use of RRS methods in population genomics studies of non-model organisms. The number and quality of markers, however, can be significantly improved when the reference genome is available (Andrews et al., 2016; Jones & Good, 2016; Ozsolak & Milos, 2011; Warr et al., 2015). The well-documented pipelines, open access software, and affordability of RAD-seq and RNA-seq methods have made them the most popular RRS approaches in non-model species studies nowadays (Andrews et al., 2016; De Wit, Pespeni, Ladner, Barshis, & Palumbi, 2012). Nevertheless, as experience on these techniques accumulates, so have realizations on limitations and potential sources of bias and error. For example, in RAD-seq, potential bias can be introduced during library preparation, sequencing, and data analysis (reviewed in Cariou *et al.* 2016; Shafer *et al.* 2016). Also, the small fraction of the genome that is often interrogated and the variable linkage disequilibrium (LD) block size usually seen in nature may limit the power of this method for the detection of adaptive variation (Hoban et al., 2016; Lowry et al., 2017a, 2017b). Some researchers argue that it is possible to achieve with RAD-seq the high genome coverage required for detecting signatures of selection owing to the flexibility of protocols that allow fine-tuning of marker density (Catchen et al., 2017; McKinney, Larson, Seeb, & Seeb, 2017). Whereas every contribution of genomic information for species with little genomic knowledge counts and should be promoted, it is also important to realize methodological limitations to avoid false expectations and misinterpretation of results (Lowry et al., 2017b) (discussion expanded in Box 5).

Unlike RRS, WGR approaches provide the highest marker density of the current genomic methods, facilitating the characterization of neutral and functional genetic variation as well as the discovery of the genetic basis of phenotypic traits (Ellegren, 2014). The proportion of the genome screened with WGR depends on read depth and completeness of the reference sequence. A comparison of the expected relative proportion of the genome covered by WGS, WES, and RAD-seq is shown in Fig. 2.3, and a comparison of requirements and outcomes of RAD-seq, Pool-seq and lcWGR is presented in Table 2.1. Another advantage of WGR approaches is that they examine multiple types of genetic variation including structural variations (SVs) (i.e. deletions, insertions, chromosomal rearrangements, copy number variation, (Alkan, Coe, & Eichler, 2011)) and mutations in regulatory elements (REs) (i.e. non-coding regions that regulate gene expression and function, (Wray, 2007)). In contrast, RRS techniques are mostly restricted to one base changes (i.e. SNPs), and RNA-seq and WES are to variation within coding sequences.

2.4 Applications of WGR in conservation and management

Below, we describe contributions that WGR analysis can make to some of the main areas of interest in conservation. For each area we provide study cases that illustrate the type of questions that can be addressed with WGR. Table 2.2 lists key aspects of the experimental design of these studies. Within parentheses, we denote in the title of each conservation area, which WGR approach (huWGR, hrWGR, Pool-seq or lcWGR) applies.

2.4.1 Phylogenomics, hybridization and taxonomical species resolution (approach: huWGR, hrWGR, lcWGR)

The successful implementation of conservation plans relies on the correct identification of the taxonomic status of organisms target of protection (Mace, 2004). Whole-genome data constitutes a complete record of a species evolutionary process. By comparing large portions of the genome rather than the sequence of a few genes, as has traditionally been done, a more robust reconstruction of the evolutionary relationships among species can be achieved. This is the aim of phylogenomics (Chan & Ragan, 2013; Delsuc, Brinkmann, & Philippe, 2005). Recent studies have provided evidence of the power of whole-genome data for the reconstruction of the tree of life and for the detection of species hybridization events. More work needs to be done, however, to resolve algorithm limitations associated with the analysis of such large amount of data, and to overcome intrinsic genomic challenges such as protein-coding sequence convergence, genome rearrangements, lateral gene transfer, incomplete lineage sorting, among others (Chan & Ragan, 2013; Delsuc et al., 2005).

huWGR and lcWGR approaches have been used in phylogenomics (Table 2.2) and in the detection of species hybridization. huWGR provides high genome coverage and taxonomic resolution whereas lcWGR offers a fragmented coverage that still can be useful for the development of phylogenetic markers (e.g. organellar genome assembly, ortholog genes, repetitive elements). For example, using huWGR, Jarvis et al. (2014) compared 48 modern bird species for the reconstruction of their phylogeny. They obtained a highly resolved tree that discriminates close relationships, identified the first divergence of Neoaves, but found difficulties attempting to resolve deep branches. Zhou et al. (2014) sequenced and assembled the genome of snub-nosed monkey and compared it with those of three related monkey species. They found evidence of functional evolution and leaf-eating dietary adaptations in this group, and were able to reconstruct species-specific demographic histories more precisely than previous attempts. vonHoldt et al. (2016) sequenced 28 wolf genomes to demonstrate that two endemic species of the North American wolf (the red wolf and the eastern wolf) are actually hybrids of coyote and gray wolf. With lcWGR data, Straub *et al.* (2011) characterized for the first time phylogenetic markers for the common milkweed (*Asclepias syriaca* L.), including the complete chloroplast genome, a partial mitochondrial genome sequence, and some single copy ortholog genes. Blischak *et al.* (2014) used ultra-low coverage genome data (~0.005x–0.007x) for annotation and gene prediction of more than 10,000 contigs for primer design of phylogenetic markers in the plant genus *Penstemon*. Wall *et al.* (2016) analyzed lcWGR data of yellow baboons (*Papio cynocephalus*), Anubis baboons (*P. anubis*), and their hybrids in the Amboseli ecosystem of Kenya. They found genetic differentiation between parent taxa and enough evidence to infer a complex admixture history involving intermittent but multiple hybridization events that did not indicate fitness reduction in hybrids. All these studies indicate the comparison of individual genomes can make significant contributions to conservation biology by helping resolve the phylogenetic status of species of concern and by identifying genomic regions that can be used for the development of cost-effective tools for species and hybrid identification. As phylogenomic inference relies on haplotypes, this kind of analysis cannot be performed using Pool-seq data but will be benefited by advances in hrWGR methods with long-read data.

2.4.2 Demographic history and historical effective population size (approach: huWGR, hrWGR, lcWGR, Pool-seq)

The study of a species' demographic history, including bottlenecks, migration patterns, range expansion, and changes in historical effective population size (hN_e), is of great interest in conservation as it helps understand past historical events and their influence on the genetic makeup of contemporary populations. Such studies also allow for the testing of hypotheses regarding the effect of dated environmental events (e.g. appearance of barriers to gene flow, anthropogenic disturbance, climate change) on historical demographic processes that may have left a genetic imprint.

huWGR and lcWGR approaches have been used for the reconstruction of the demographic history of different species (Table 2.2). For example, Zhao *et al.* (2012) analyzed huWGR data of 34 pandas finding genetic evidence of multiple demographic events such as population expansion, bottlenecks, and divergence, and inferred that human activities most likely contributed to their decline in the last 3000 years. Foote *et al.* (2016) reconstructed the ancestral demography of five distinct ecotypes of the killer whale (*Orcinus orca*) based on lcWGR data of 48 individuals and huWGR data of 2 individuals. They discovered that patterns of differentiation between pairs of contemporary allopatric and sympatric ecotypes are most likely the consequence of ecological divergence and genetic drift resulting from bottlenecks experienced during past founder events.

hN_e has been estimated from huWGR data. For example, using pairwise sequentially Markovian coalescent (PSMC) analyses, Nadachowska-Brzyska *et al.* (2016) obtained hN_e estimates of four species of Western Palearctic black-and-white flycatchers of the genus *Ficedula* based on whole-genome data of 200 individuals from 10 European populations. The hN_e curves indicated the most recent common ancestor of the four species dates back to 1–2 million years (Mya) and each species followed separate evolutionary paths involving population growth, decline (~100–200 thousand years ago (Kya)) and expansion. Authors suggest a mean genome coverage of ≥18X per individual, a per-site filter of ≥10 reads and no more than 25% of missing data are required for a proper inference of demographic history using PSMC and huWGR data (Nadachowska-Brzyska *et al.* 2016). Boitard *et al.* (2016) estimated N_e using an Approximate Bayesian Computation Approach and whole genomes of 15-25 individuals from each of four cattle breeds (Angus, Fleckvieh, Holstein, Jersey). They found evidence that historical domestication and modern breeding events were related to population decline. More recent statistical models promise the possibility to estimate hN_e from lcWGR and Pool-seq data, the first based on inbreeding Identity By Descent (IBD) tracts (Vieira, Albrechtsen, & Nielsen, 2016), and the latter on allele frequency changes between two temporal samples while correcting for the potential inflation of variance in allele frequency due to the two sampling steps involved in Pool-Seq experiments (i.e. during the sampling of individuals and the sequencing of pools) (Jonas, Taus, Kosiol, Schlotterer, & Futschik, 2016). In summary, the studies above illustrate how huWGR and lcWGR can facilitate inference on hN_e fluctuations and the tracing of historical demographic events that can help understand patterns of genetic diversity and structure in contemporary populations. These kinds of analyses can be extended to species of conservation interest. Haplotype-resolved genomes obtained with hrWGR approaches promise enhanced accuracy in demographic history and effective population size estimates as older long identical-by-descent portions of the genome can be assessed (Schiffels & Durbin, 2014; Snyder et al., 2015).

2.4.3 Population structure and admixture (approach: huWGR, Pool-Seq, lcWGR)

One of the main goals in conservation biology is to maintain high genetic diversity in vulnerable species. Natural populations are commonly structured in local subpopulations. Genetic differences can arise among subpopulations over time as a result of the interplay between gene flow (e.g. reduced due to geographical distance or presence of barriers to dispersal), genetic drift, and local adaptation (Allendorf, Luikart, & Aitken, 2013). Traditionally, the partitioning of genetic diversity within and among populations has been inferred using F-statistics, with F_{ST} being an estimate of the genetic differentiation among subpopulations (Holsinger & Weir, 2009). The assumption of this approach is that the average effect of neutral processes (e.g. gene flow and genetic drift) acting equally throughout the whole genome, can be estimated based on the average allele frequency at several neutral loci within subpopulations and over all subpopulations. A genomics approach, where multiple high-density loci are examined, allows instead the detection of the effect of different evolutionary forces (e.g., drift, selection) along the genome through the estimation of genetic diversity using a sliding window procedure (Allendorf, 2016). As the analysis of whole genomes provides the highest marker density, it allows the simultaneous evaluation of genome-wide patterns in neutral loci that act as a record of demographic and historical events, as well as locus-specific effects that can be associated to natural selection, fitness, and adaptation (Allendorf, 2016; Allendorf et al., 2013). This new perspective for the comparison of genetic diversity among populations is providing novel insights on how different evolutionary forces have affected particular loci and how differentiation could arise among natural populations despite gene flow. Below we describe some studies that estimated population structure and admixture based on neutral loci surveyed with WGR (Table 2.2).

Parejo *et al.* (2016) used huWGR data of 151 haploid drones to assess the degree of admixture between native European dark honeybee (*Apis mellifera mellifera*) and two introduced honey bee subspecies (*A. m. carnica* and buckfast) in four conservation areas of *A. millifera millifera* in Switzerland and one in France. They found genetic differentiation between subspecies that coincided with geography and admixed individuals in protected areas. With the 50 most informative loci, they created a SNP panel for the genetic identification and monitoring of native and introduced bees. Fischer *et al.* (2017) compared genetic diversity and population differentiation estimates from Pool-seq and microsatellite data of 9 wild populations of the plant *A. halleri* in south-eastern Switzerland and northern Italy. They found no concordance of expected heterozygosity (H_e) estimates between marker types, and microsatellite allelic richness was a better descriptor of genome-wide diversity than H_e. They found that a few thousand SNPs can provide a better estimate of genetic diversity and genetic differentiation among their populations than the 19 microsatellite loci tested. Velasco *et al.* (2016) conducted a genome-wide analysis of the effects of domestication and mating system on genetic diversity of almond (*Prunus dulcis*) and peach (*P. persica*). With lcWGR data of 13 individuals from each species, they found that the genome-wide nucleotide diversity was ~7-fold higher in almond than in peach, an excess of rare alleles in both species likely consistent with a recent population expansion event, no evidence of population bottleneck related to domestication, and a strong genetic differentiation between species. Overall these examples demonstrate WGR data is useful for the estimation of population structure and admixture in a variety of species.

2.4.4 Signatures of selection, genetic basis of phenotypic traits, and local adaptation (approach: hrWGR, huWGR, Pool-seq, lcWGR)

The identification of genomic regions involved in adaptation to local environmental conditions (local adaptation) is one of the main goals in evolutionary biology. This knowledge is crucial for conservation biology because of the importance of functional genetic diversity potentially linked with persistence in novel environments (Allendorf et al., 2010). Establishing the connection between genotype, phenotype and fitness is usually difficult though, and requires additional testing to verify the effect of presumably adaptive loci on fitness (Barrett & Hoekstra, 2011; Nielsen, 2009). We thus refer as "putatively" adaptive variants the parts of the genome that exhibit genetic signatures of selection for which the effect on fitness has not yet been tested.

There are three general strategies for the identification of loci under selection: (1) forward genetics [includes QTL mapping and Genome Wide Association Studies (GWAS)], when the phenotypic traits that underpin adaptation are known; (2) reverse genetics [includes genome scans via Genetic-Environment Association (GEA) analyses and outlier loci tests], when the adaptive phenotype is unknown; and (3) candidate genes examination. A complete explanation of these methods is reviewed elsewhere (Barrett & Hoekstra, 2011; Pardo-Diaz, Salazar, & Jiggins, 2015; Vitti, Grossman, & Sabeti, 2013). WGR is usually classified as reverse genetics as the traits under selection are generally unknown for non-model species. However, when there is particular interest in comparing contrasting phenotypes (e.g. ecotypes), a forward genetics approach following a GWAS-type comparison is possible, as is the directed screening of candidate genes discovered via genome scans. Genome scans are probably the most common method to detect signatures of selection in genomic data. Despite their proven power for this purpose, numerous considerations need to be accounted for when designing a genome scan experiment (see Box 5).

The advantage of WGR over other genomic approaches for the detection of loci under selection relies on the possibility to screen neutral and functional polymorphisms in high genomic resolution. Such high marker density is crucial for the identification of genetic signatures of selection such as, reduced nucleotide diversity, extended linkage disequilibrium, and high homozygosis (Ellegren, 2014). WGR has been used for the discovery and mapping of the genetic basis of phenotypic traits with adaptive importance (Table 2.2), for instance, the beak shape of Darwin finches (huWGR; Lamichhaney *et al.* 2015a) or the age at maturity in Atlantic salmon (huWGR: Barson *et al.* 2015; Pool-seq: Ayllon *et al.* 2015). Other examples, all based on Pool-seq data, include: the red beak colour in canaries (Lopes et al., 2016), circadian timing in midges (Kaiser et al., 2016), genes affecting brain and neuronal development associated with domestication of rabbits (Carneiro et al., 2014), genes associated with breeding related traits of pathogen resistance and reproductive ability in two highly inbred chicken lines (Fleming et al., 2016), candidate genes potentially driving the morphological, life history and salt tolerance differences between ecotypes of the yellow monkey-flower plant (Gould, Chen, & Lowry, 2017), and immune related genes in Atlantic salmon (Kjærner-Semb et al., 2016). Similarly, studies based on lcWGR data include the detection of a recent partial barrier (large inversion) to gene flow between subgroups of the mosquito *Anopheles gambiae* s.l. (Crawford et al., 2016), and the identification of loci presumably involved in adaptation to high altitude and arid environments in native sheep (Yang et al., 2016).

Additionally, WGR allows the examination of SVs and mutations in regulatory elements (REs) (i.e. non-coding DNA sequences that control expression of neighboring genes), providing a more complete genome-wide spectrum of the amount and distribution of genetic variation. Both, SVs and mutations in REs have been shown to play an important role in the determination of phenotypic diversity, some of which could affect fitness (Wittkopp & Kalay, 2012), and the development of diseases (Melton, Reuter, Spacek, & Snyder, 2015). For example, a large deletion identified with Pool-seq data causes skeletal atavism in Shetland ponies (Rafati et al., 2016). A large chromosomal inversion discovered from huWGR data underlies the complex male mating morph diversity exhibited by the bird ruff (Küpper et al., 2015; Sangeet Lamichhaney, Fan, et al., 2015). And a chromosomal inversion is also responsible for the individual wing-pattern of diverse mimetic morphs in butterflies (Joron et al., 2011). Inversions play an important role because they reduce recombination, thus preventing the disruption of co-adapted gene complexes (Hoffmann & Rieseberg, 2008). Mutations in REs can affect gene expression having functional consequences in phenotypic traits (Wittkopp & Kalay, 2012; Wray, 2007). For example, mutations in the regulatory sequence of genes related to pigmentation produce different coloration patterns in cuticle, wings, and abdomen of fruit flies (*Drosophila melanogaster*) (Wittkopp & Kalay, 2012). Changes in regulatory elements most likely are responsible for pelvic structure reduction in three-spine sticklebacks (*Gasterosteus aculeatus*) (Shapiro et al., 2006) and for some human limb malformations (VanderMeer & Ahituv, 2011). Haplotype information obtained with hrWGR can help in the detection of haplotype-specific mutations and in the association of epigenetic factors and gene expression in tumors (Adey et al., 2013).

With WGR it is also possible to perform GEA analyses for the identification of genetic variation associated with adaptation to local conditions. For example, using Pool-seq data, Fischer *et al.* (2013) and Rellstab *et al.* (2016) evaluated the association of natural populations of *Arabidopsis halleri* with environmental factors in two time periods. For the first period, they analyzed 5 populations and detected two million SNPs. They found 175 genes to be highly associated with some of the five environmental factors tested. For the second period, they extended the study to 18 populations covering a larger geographic area. Only 11 genes were found with the same association in both time periods, which could be a result of the alpine environment heterogeneity for which selection may be acting at the population level. Martinez Barrio *et al.* (2016) compared Pool-seq data of 20 Atlantic herring populations across the brackish Baltic Sea and the northeast Atlantic Ocean finding significant allele frequency differences at multiple loci between brackish-water and oceanic populations and between spring and fall spawning populations, contrary to the common expectation of overall small genetic divergence within European populations previously observed in studies that only considered low density loci (Larsson, Laikre, André, Dahlgren, & Ryman, 2010; Limborg et al., 2012; Teacher, André, Jonsson, & Merilä, 2013; Teacher, André, Merilä, & Wheat, 2012). Similarly, Lamichhaney *et al.* (2017) observed a pattern of differentiation between herring populations on both sides of the North Atlantic that comprised minute genetic differences at neutral loci but significant allele frequency differences between spring and autumn spawning populations at 6 333 SNPs, some of which are most likely associated with spawning time regulation. A total of 25% of such loci where shared between the American and European populations, and the unique loci found on each side of the ocean presumably result from local adaptation. In summary, the WGR approach is a powerful tool for the detection of signatures of selection, for uncovering the genetic basis of phenotypic traits and diseases, and for the identification of signatures of local adaptation.

huWGR, Pool-seq, lcWGR approaches can contribute to the understanding of the genetic variants and mechanisms underlying adaptive traits.

2.4.5 Inbreeding depression, conservation breeding and restoration (approach: huWGR, potential contributions of Pool-seq and lcWGR)

Understanding the genetic basis and effects of inbreeding depression (defined as fitness reduction of the offspring resulting from the mating between closely related individuals) is a major goal in conservation biology as it affects the long-term viability of small isolated populations, whose persistence depends on targeted breeding, purging, and restoration programs (Allendorf et al., 2013). Numerous studies have tried to reveal the genetic basis underlying inbreeding depression in wild populations however, the major obstacle for this has been the limitation to estimate the degree of individual inbreeding following traditional methods, as they require parental analysis over several generations. WGR analysis can solve these limitations by providing a large amount of genomic data per individual, which relaxes the need for parental analysis (reviewed by Kardos *et al.* (2016b) and Hedrick *et al.*(2016)). For instance, Xue *et al.* (2015) obtained huWGR data of 44 wild individuals representing four subspecies of gorilla in Africa. They observed on average 34% homozygosis in individual genomes which indicates extensive inbreeding most likely as a result of severe recent population decline. They also found very low genetic diversity in two of the four subspecies likely resulting from steady population declines over the past 100 000 years. Myburg *et al.* (2014) compared huWGR data on one outbred *Eucalyptus grandis* parent tree and 28 offspring obtained through self-fertilization. The progeny retained high and different heterozygosity percentage (52% to 79%, average 66%), in disagreement with an expectation of 50% homozygosis Identical-By-Descent (IBD) produced by selfing without selection. Hedrick et al. (2016) analysed the same dataset finding that pseudo-overdominance most likely explained the observed inbreeding depression, which could be underlined by 100 or more genes of large effect associated with viability.

To our knowledge no study has thus far (June 2017) used the lcWGR approach for the study of inbreeding depression in wild populations, however, this method has been successful in the identification of causal mutations of three phenotypic traits in inbred rice varieties (Wang et al., 2016) and the estimation of individual inbreeding coefficients (Vieira et al., 2016; Vieira, Fumagalli, Albrechtsen, & Nielsen, 2013). Thus, we envision lcWGR data is likely to be useful for this purpose in the near future. Inbreeding depression cannot be estimated from Pool-seq data as individual information is lost. However, this type of data can help in the identification of the genetic basis of phenotypic traits associated with inbreeding depression, as it has been used for the characterization of genetic variation underlying diseases (Rafati et al., 2016) and phenotypic traits in inbred organisms (Fleming et al., 2016). In conclusion, the examples presented here demonstrate WGR can help in the characterization of the genetic basis of inbreeding depression. This genetic information can be used for early diagnosis of inbreeding depression, assist in the planning of breeding programs so as to avoid the inclusion of individuals carrying deleterious mutations that can affect recovery of captive or wild populations, and for the prediction of purging efficacy (Allendorf et al., 2010).

2.4.6 Units of conservation, mixed stock analysis, and genetic monitoring

Genetic patterns obtained from the analysis of neutral and adaptive genetic variation are useful for the delineation of conservation and management units (Funk et al., 2012). As previously shown, WGR approaches generate a large amount of neutral and putatively adaptive genetic markers. A subset of these loci can be used for the development of cost-effective genotyping tools suitable for the assessment of diverse aspects of interest in conservation and management (e.g. taxonomic status, hybrids, sex, carriers of genetic diseases, population structure, individual assignment and population of origin, among others). These tools can be incorporated in conservation plans of threatened species (Fussi et al., 2016; Grossen, Biebach, Angelone-Alasaad, Keller, & Croll, 2017; Ivy, Putnam, Navarro, Gurr, & Ryder, 2016; Muñoz et al., 2015; Norman, Street, & Spong, 2013; Stetz et al., 2016; Vandergast, 2017), and in management plans of commercially valuable species (Aykanat, Lindqvist, Pritchard, & Primmer, 2016; Bekkevold et al., 2015a; Bradbury et al., 2015; Habicht et al., 2012; Martinsohn & Ogden, 2009; Sinclair-Waters, 2017).

2.5 Concluding remarks and future directions

Our review synthesizes the advantages of WGR for addressing central questions in evolutionary biology that have not been fully answered using traditional techniques. The power of WGR resides in two main features: (*i*) it assesses neutral and functional genetic variation (including regulatory regions) at the highest genomic resolution among of current methods; and (*ii*) examines a wide variety of genetic variation, from one base changes to structural variants. The method thus facilitates the examination of the genetic basis of phenotypic traits and diseases as well as the detection of the genetic signatures of natural selection, some of which can be related to local adaptation.

The scarcity of genomic resources for most species under conservation concern coupled with the still high cost of high-throughput sequencing and the elevated demand for computing resources, however, have limited the implementation of WGR in conservation biology. Some questions in conservation biology can be reasonably addressed using traditional or RRS approaches. The question then arises of when is a WGR approach justifiable? The answer depends on the research question, knowledge on the biological system, genomic resources available, the genetic architecture of phenotypic traits and ultimately on funding. If the research focus is the analysis of neutral processes, then WGR would not be necessary since RAD-seq methods would excel for an affordable price. If a highly accurate reconstruction of the species historical demography is sought, WGR would be justifiable since the estimation of coalescent events benefits from the information provided by haplotype-resolved genomes. A good approximation to the species historical demography can also be achieved with the data generated by RRS, though a larger sample size would be necessary (Manthey, Campillo, Burns, & Moyle, 2016). The major motivation for using WGR is thus the detection of signatures of selection and the characterization of the genetic basis of phenotypic traits and diseases. RAD-seq can also be used for this purpose at the fraction of the genome screened, although its success may depend on the proportion of the genome covered. Ideally this proportion should match the extension of linkage disequilibrium blocks, but this is usually unknown. Previous knowledge of the system and genomic resources can assist in the choice between RNA-seq, WES, and WGR. For example, when there is a presumption that selection is operating on a specific tissue/life stage/time, then RNA-seq would be appropriate for assessing genetic variation in the genomic regions expressed at time of sampling. If the genes of interest are already described, then target capture and sequencing is the best strategy. When no candidate genes are known, a higher density screening such as WES or WGR would be preferable. When there is high confidence that selection is acting mostly on protein-coding parts of the genome, WES would be a cost-effective approach compared with WGR. When there is a notion that selection could be acting in regulatory elements or could be mediated by large structural variations, then WGR is likely the best choice as it provides the highest marker density and diversity in genetic variants assessed. Given that in general we do not know how selection is acting on a particular species, life stage, tissue or part of the genome, WGR should be considered as a starting point for the exploration of genomic diversity assuming sufficient funding and a reference genome are available. In the absence of reference genome, RAD-seq is an affordable alternative for the screening of neutral and putatively adaptive variation in a fraction of the genome (Catchen et al., 2017; McKinney et al., 2017) with some limitations (Hoban et al., 2016; Lowry et al., 2017a, 2017b). Fig. 2.4 summarizes the rationale for the selection of genomic approach as a function of expected linkage block size and type of genetic variation of interest. Once regions of interest are detected and mapped, then a more affordable and scalable genotyping approach can be developed for massive individual genotyping on such loci.

The success of genome scans to detect adaptive variation depends on multiple factors including genetic architecture, effect size, sample size, percentage of the genome covered, effective population size and genetic drift, etc. (Box 5). Such information is usually unknown, making it hard to predict the success of RRS to reveal loci underlying a specific trait. Genome completeness, effect size and genetic architecture of a trait, and sampling design determine whether the entire genome is assessed, whether the genetic basis of a trait is traceable using genomics tools, and whether relevant individuals are included in the analysis, respectively. Also, genetic patterns resulting from demographic processes and drift may resemble those of local adaptation (Hoban et al., 2016). Therefore, outlier loci detected with genome scans should be treated as working hypothesis for further functional testing. Experimental evaluation of the effect of mutations on phenotypic traits and fitness is required to confirm and understand their adaptive nature. Such experimentation is unlikely to be performed on threatened species, although model species can be used instead, as many biochemical pathways are conserved across taxa (Andersson et al., 2012), and advances in genome-editing tools (i.e. CRISPR/Cas9) may facilitate functional testing (Varshney et al., 2015). Despite its limitations, WGR has proven to be an alternative for revealing adaptive loci and the genetic architecture of traits in a variety of organisms (Table 2.2), including humans (Auer & Lettre, 2015; Durbin et al., 2010; Field et al., 2016; Nielsen et al., 2017). Additionally, genomic data alone may not fully explain phenotypic variation. Epigenetic mechanisms (i.e. modifications of gene expression not due to changes in DNA sequences) play an important role in phenotype determination (Bossdorf, Richards, & Pigliucci, 2008; Richards, Bossdorf, & Pigliucci, 2010) suggesting a holistic approach would be ideal for better understanding phenotypic diversity and evolution.

Currently there is no single WGR method fulfilling all requirements in conservation geneticists, and each method has its own limitations including sources of potential error and bias (Box 3 and 4). However, the implementation of good practices can control and minimize such biases, resulting in informative and reliable datasets that can be used for population genomics inference (details in Box 2; Fracassetti *et al.* 2015; Wang *et al.* 2016a; Martinez Barrio *et al.* 2016). The field of genomics is rapidly changing, bringing new technologies and computing algorithms that promise solutions to present restrictions. For example, short-read sequences are of limited assistance for genome assembly, haplotype-phasing, and detection of large SVs. Long-reads from third-generation sequencing can help overcome these limitations by resolving difficult parts of the genome (i.e., repetitive sequences and SVs) and by allowing the direct phasing of haplotypes. The relative low throughput, high error rate and cost have however, restricted the use of third-generation sequencing platforms (Bleidorn et al. 2016); though improvements promising even higher throughput and lower cost and error rate are underway. This implies that in the future lower coverage per individual will likely be needed, high-throughput sequencing will likely be cheaper making it more accessible than otherwise, and larger sample sizes could be screened. Similarly, new computer tools and paradigms are being created, for example, graph-based genomes (Paten, Novak, Eizenga, & Garrison, 2017; The Computational Pan-genomics Consortium, 2016) aim to overcome the current limitations of genome assembly which produce haplotype genome sequences, excluding a great amount of genomic variation and limiting variant detection.

For some conservation areas described above the benefit of a dense and large number of markers is not clear. Effective population size estimation is a case in point (Waples, Larson, & Waples, 2016). Genomic information gathered via WGR can make important contributions to conservation planning and management of commercially exploited species, for instance, by helping in the delimitation and monitoring of evolutionary and/or management units and in the prioritization of imperiled populations. At the fast pace of computational and sequencing development, we can envision a not very distant future where simplified procedures, analysis, and interpretation will make genomic tools accessible to managers (Garner et al., 2016; Shafer et al., 2015), the analysis of genomes will be performed in the field (Quick et al., 2016), and genomic analysis will become a routine task in many non-model species and fields.

2.6 Box 1: Genome assembly and completeness of genomes sequenced to date

The assembly of a genome consists in joining sequences of the DNA of one or several conspecific individuals into a single sequence. DNA is first fragmented to a particular length and sequenced to a certain coverage depending on the sequencing platform. DNA sequences are then assembled using specialized computer algorithms. As per the current assembly algorithms, a haploid sequence is generally obtained, implying species genome diversity and assembly accuracy could be compromised (Baker, 2012; Paten et al., 2017).

A genome project generally has the goal of obtaining a contiguous and complete genome sequence with annotated genes (Veeckman, Ruttink, & Vandepoele, 2016). To achieve this using current sequencing technologies of second- and third-generation, it is necessary to get high sequence depth (>50-60x) evenly distributed across the genome, counteracting the relatively high sequencing error rate (Bleidorn, 2016; Goodwin et al., 2016; H. Lee et al., 2016). As mentioned, repetitive sequences and large structural variants are difficult to assemble using just short sequence reads. Thus, the combined assembly of short and long reads (or long reads only (Bickhart et al., 2017; Chakraborty, Baldwin-Brown, Long, & Emerson, 2016)) is common practice (Ekblom & Wolf, 2014).

The quality of a genome assembly (how complete and accurate it is) has traditionally been assessed using different metrics such as N50 and L50. These two metrics assess contiguity, N50 estimates the contig/scaffold length at which 50% of the total bases fall in in a given assembly, and L50 is the number of contigs/scaffolds that are longer than or equal to the N50 length, including 50% of the total bases of a given assembly (https://www.ncbi.nlm.nih.gov/assembly/help/). However, these metrics have several limitations (Gurevich, Saveliev, Vyahhi, & Tesler, 2013; Salzberg et al., 2012) for which new ones have been proposed, for example NG50 and NA50 (Bradnam et al., 2013; Earl et al., 2011). Guidelines for achieving high quality *de novo* genome assembly of non-model species are presented in Ekblom and Wolf (2014) and in Koepfli *et al.* (2015), and recent advances in genome assembly are addressed in a special issue of the journal Genome Research (Phillippy, 2017).

Genomes assembled to date are publicly available online in GenBank of the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/genbank/), the European Nucleotide Archive (ENA) (http://www.ebi.ac.uk/ena) and the DNA DataBank of Japan (DDBJ) (http://www.ddbj.nig.ac.jp/), institutions that constitute the International Nucleotide Sequence Database Collaboration. Ongoing genome projects are listed in GenBank in the Bioprojects page (https://www.ncbi.nlm.nih.gov/bioproject/) and in the Genomes Online Database (GOLD) (https://gold.jgi.doe.gov/projects) (Mukherjee et al., 2017).

With the advances in sequencing technology and computation achieved with the Human Genome Project (HGP) (1990-2003) (Human Genome Sequencing Consortium, 2004), there has been an exponential growth in the number of genomes published in GenBank per year (Box 1 Fig. 2.1A). The great majority of genomes correspond to viruses and prokaryotes, and within eukaryotes, fungi, animals and protists follow in representation. Within animals, mammals and insects have the highest number of genomes sequenced, whereas amphibians and reptiles have the lowest. Within plants, land plants have the highest representation (Box 1 Fig. 2.1B). Contrary to expectations, most prokaryotic and eukaryotic genome sequences to date are incomplete (Box 1 Figure C and D), as they are assembled to the scaffold level. In general, it is only prokaryotes, viruses, and a few model eukaryotic species (e.g. yeast, 12.1 Mega bases - Mb, and fruit fly, 175Mb) with relatively small, simple or only moderately complex genomes that have their sequence complete (Box 1 Figure C and D). Within animals, mammals, fishes, insects, and birds have their genomes assembled to the chromosome level, whereas mammals, insects, flat and round worms followed by fishes and birds have their genomes at the contig level. Land plants have also a great proportion of genomes at the chromosome and contig level.

Within eukaryotes there is a great diversity in genome size, complexity, and proportions of repetitive sequence content [e.g., Atlantic salmon (*Salmo salar*): 2.97 Giga bases (Gb) with ~60% repetitive elements (Lien et al., 2016); loblolly pine tree (*Pinus taeda*): 23.2 Billion bases (Bb) (largest genome sequenced to date with 82% repetitive content (Neale *et al.* 2014)]. Genome complexity increases with the occurrence of duplicated genes (or paralogs), long repeat sequences, polymorphic genes (e.g. MHC, Trowsdale & Knight 2015), GC content, and ploidy (Treangen & Salzberg, 2011). The complexity of a genome, and especially its repetitive content (Goodwin et al., 2016; Reinert, Langmead, Weese, & Evers, 2015), imposes significant challenges for sequence assembly (Baker, 2012; Ekblom & Wolf, 2014; Ellegren, 2014; Treangen & Salzberg, 2011). This could largely explain the varying degree of completeness observed in the genomes sequenced to date.

How complete and accurate a genome assembly is will determine its suitability for posterior analyses. For example, a very incomplete genome can still be helpful for the identification of SNPs but it would fail in the detection of large structural variation (da Fonseca et al., 2016). Despite its usefulness in some applications, the general consensus is that incomplete draft genomes bring more problems than solutions, especially for the accuracy of SNP calling (Li & Wren, 2014).

2.7 Box 2: General workflow for whole-genome resequencing data acquisition

Schematic illustration of the general workflow is shown in (Box 2 Fig. 2.1).

2.7.1 Wet-lab procedures

Tissue sampling and preservation

Damage of DNA should be avoided. Examples of tissue sampling protocols used in the 10K vertebrates genome project (Koepfli et al., 2015) are described in Wong *et al.* (2012).

DNA extraction, quality and quantity

Quality of DNA can be assessed with ~0.8-1% agarose gel electrophoresis and a 25Kb molecular weight ladder. A single high molecular weight band (~23Kb) indicates good quality DNA. High purity DNA is necessary, which corresponds to 260/280nm absorbance ratio of ~1.8-2.0. Highly fragmented DNA should be avoided as it cannot be quantified accurately using fluorometric-based methods, typically recommended for accurate double-strand DNA quantification (Sedlackova, Repiska, Celec, Szemes, & Minarik, 2013). For Pool-seq this is particularly important as the even contribution of individual DNA in a pool relies on accurate quantification. The amount of starting DNA depends on the library preparation kits' input requirements described in Table 2.1.

Standardization of DNA concentration across samples (for Pool-seq and lcWGR)

Each DNA sample is diluted or concentrated to a desired standard value (ng/μl). The diluting liquid should stabilize and protect DNA from damage (e.g. lowTE). A liquid handling robot is recommended for this step to eliminate the potential for pipetting error.

DNA pooling (Pool-seq)

Pooling consists on mixing equimolar amounts of DNA of several individuals from a population. When the interest is to identify the genetic basis of a trait, pools should comprise individuals sharing the same trait (not necessarily form the same population) and extreme trait categories have increased potential to lead to clearer genetic signals. A minimum of 50 individuals is recommended per pool, but including more (>100), (assuming proportional increase in sequencing depth) can help minimize slight unevenness in the representation of few individuals leading to more accurate allele frequency estimates (Gautier et al., 2013; Schlötterer et al., 2014). Individual DNA is then diluted to a standard concentration and verified through a quantification step. Once normalized, the same amount of DNA from individual samples can be pooled into a single tube.

Sequencing library preparation

Several kits for library preparation are available commercially. They differ in cost/sample, the need for a sonicator, the incorporation of a DNA amplification step using PCR, and the amount of input DNA. For current price and DNA input requirements of Illumina kits see Table 2.1. DNA amplification is convenient for low DNA amounts, but PCR can introduce biases (e.g. underrepresentation of GC-rich fragments, preferential amplification of short fragments, and duplicates) that can lead to uneven coverage in some loci. Some biases can be minimized by adjusting the PCR protocol (Aird et al., 2011) and duplicates can be removed *in silico* using Picard tools, http://broadinstitute.github.io/picard, or SAMtools (Li et al., 2009). Small structural variants (INDELs and CNVs) can be detected from short-reads of standard libraries (~350-550bp insert size). For large structural variants detection (spanning Mbs) a Mate-pair library is required (~2-20Kb insert size). Additional considerations are discussed in Head *et al.* (2015).

High-throughput sequencing of DNA libraries

Currently the most popular technology for short-read NGS is Illumina, though new technologies are being developed (Goodwin et al., 2016). Illumina offers an overall accuracy >99.5%, which is high relative to other platforms, but still restrictive as it is difficult to distinguish true genetic variation from technical artifacts (Laehnemann, Borkhardt, & McHardy, 2016). The suggested minimum coverage for huWGR is >30x/individual (Sims, Sudbery, Ilott, Heger, & Ponting, 2014), for Pool-seq it is >50x/pool (Schlötterer et al., 2014), though a much higher coverage should be targeted

(>100-200x) for rare allele detection (Jingwen Wang et al., 2016), and for lcWGR it is 1-4x/individual (Buerkle & Gompert, 2013; Nielsen et al., 2011). The number of Illumina lanes needed depends on the trade-off between genome size, target coverage per sample/pool, and flow-cell yield. Illumina sequencing is potentially prone to lane-to-lane variation (Ross et al., 2013), a problem that can be minimized by distributing barcoded libraries across multiple lanes (Sergio Pereira TCAG DNA pers. comm.).

2.7.2 Computer procedures

Quality control of raw sequences

Raw sequence data comes from the sequencer in the FASTQ format (Cock, Fields, Goto, Heuer, & Rice, 2009). To control for sequencing errors, low quality bases (PHRED quality scores <20) and adapter sequences are trimmed off using Trimmomatic (Bolger, Lohse, & Usadel, 2014) or Cutadapt (M. Martin, 2011) after an initial sequence quality assessment with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

Read mapping to a reference genome

High-quality reads are mapped to a genome based on sequence similarity. Multiple algorithms for short-read mapping exist and have been reviewed elsewhere (Fonseca, Rung, Brazma, & Marioni, 2012; Hatem, Bozdağ, & Çatalyürek, 2013; Reinert et al., 2015; Ye, Meehan, Tong, & Hong, 2015). Some of the most commonly used free aligners are BWA (Li, 2013; Li & Durbin, 2009a, 2010) (Table 2.2) and Bowtie2 (Langmead & Salzberg, 2012). Alignment artifacts could arise due to multiple factors, including misalignments around INDELs and divergence between the subject reads and the reference genome. It is thus important to understand how the various algorithms work to make informed decisions on how to optimize running parameters (see Box 3). The final product of read mapping is a SAM (Sequence Alignment/Map) file (several Gb in size), format that contains a line for each read and fields with associated information including read position and mapping quality score (MAPQ or MQ) (Li, Ruan, & Durbin, 2008) that can be used for SNP filtering. A BAM file, the compressed light-weighted binary version of the SAM file, is obtained using Picard tools (http://broadinstitute.github.io/picard), and it is the format commonly preferred as input file by other programs. Other steps to prepare a BAM file for variant calling are described in (Van der Auwera et al., 2013).

Quality control of mapped reads

A visual exploration of the BAM file with the Integrative Genomics Viewer (IGV) (Robinson et al., 2011; Thorvaldsdóttir, Robinson, & Mesirov, 2013) can help identify regions with extremely high or low coverage, strand bias, misalignments around INDELs and repetitive regions, among others. As alignment errors can occur, it is important to verify that reads mapped evenly and correctly to minimize false variant calls. Evaluation of mapping quality can be complemented with summary statistics (e.g. average depth of coverage; insert size distribution, and number of mapped reads, properly paired reads, singletons, and ambiguous mappings) that can be obtained with SAMtools, ea-utils (http://expressionanalysis.github.io/ea-utils/), or Qualimap (Okonechnikov, Conesa, & García-Alcalde, 2015).

Indel realignment (depending on SNP caller)

Punctual mapping artifacts around INDELs may not be resolved by optimizing overall mapping parameters. Local INDEL realignments are a necessary prerequisite when using a site-based SNP calling algorithm like SAMtools (Li et al., 2009) or GATK-*UnifiedGenotyper* (McKenna et al., 2010). This step is not needed when using haplotype-based callers like FreeBayes (Garrison & Marth, 2012) or the GATK-*HaploypeCaller* (http://gatkforums.broadinstitute.org/gatk/discussion/7847). INDEL realignment can be done with specific functions in GATK (McKenna et al., 2010) (tutorial: https://software.broadinstitute.org/gatk/guide/article?id=7156). A file with known INDELs can help defining targets for realignment (Van der Auwera et al., 2013), but in its absence, INDELs identified during read mapping can be used instead (default mode)(https://software.broadinstitute.org/gatk/events/slides/1504/GATKwr7-X-3-Non_human.pdf).

Base recalibration (optional but recommended)

Per-base quality scores obtained from sequencers often present errors. Because SNP calling and genotype likelihood algorithms consider such quality scores, they should be corrected. This can be achieved using the base quality score recalibration (BQSR) implemented in GATK (DePristo et al., 2011; Van der Auwera et al., 2013). A known set of variants is required, but in its absence, an iterative bootstrapping approach can be implemented (Snyder-Mackler et al., 2016; Tung, Zhou, Alberts, Stephens, & Gilad, 2015)

Detection of variant sites

Specific software exists for the detection of the different types of genetic variants (i.e. SNPs and INDELs, SVs, and CNVs). Such algorithms implement particular models of variation and sources of information for the discovery of polymorphisms from short-read data. Variant positions are detected differently in huWGR, Pool-seq, and lcWGR data. In the first two, polymorphic site detection is based on per-site read coverage and quality per individual or population, respectively, whereas in the latter, it is based on coverage and quality of all the reads covering a site from several individuals of a given sample. SNPs are not called in lcWGR, instead, per-site genotype likelihoods are calculated using software like ANGSD (Korneliussen et al., 2014). In huWGR and Pool-seq, SNPs are called using software like GATK-*HaplotypeCaller*, SAMtools, or FreeBayes (Table 2.2). A comprehensive review of SNP calling using NGS data can be found in Nielsen *et al.* (2011b) and (2012), and for structural variants in Alkan *et al.* (2011). Each SNP calling algorithm makes a series of assumptions that can lead to different results. Thus, a good practice is to compare the SNPs detected by at least two algorithms (O'Rawe et al., 2013). The product of variant calling is a VCF (Variant Call Format) file containing raw polymorphisms and annotations (Danecek et al., 2011).

The selection of a SNP calling algorithm for Pool-seq data requires consideration of whether it handles ploidies larger than 2. In theory, *Pool ploidy=Ploidy per individual* x *Number of individuals*. Assuming 50 diploid individuals are mixed, pool ploidy is 100. Such large ploidies, however, deplete system memory and multiply runtime (in GATK-*HaplotypeCaller* https://software.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_gatk_tools_walkers_haplotypecaller_HaplotypeCaller.php, and FreeBayes https://github.com/ekg/freebayes/commit/576bc703c246035342538a0feeecd1, accessed March 2017).

Using the default ploidy (2) leads the software to call only the 2 most common alleles in a pool, as ploidy assumes 50/50 allele frequency (http://gatkforums.broadinstitute.org/gatk/discussion/6551/what-if-ploidy-is-set-to-2-for-pooled-dna-sequencing-experiment, accessed March 2017). This might not be an issue when calling SNPs among closely related samples as SNPs are considered biallelic, but it would limit the overall number of alleles detected when comparing more distantly related samples. Use of large ploidies is now partially solved by establishing the maximum number of alternative alleles to be considered. In GATK v.3.7 this can be set with the flag *–maxGenotypeCnt* (https://software.broadinstitute.org/gatk/blog?id=8692), and in FreeBayes with *--use-best-n-alleles* and setting a pooled mode (*--pooled*). These settings make the algorithms run faster at the expense of missing low-frequency alleles in multiallelic loci (https://github.com/ekg/freebayes).

Quality control of raw variants

SNPs with low support should be removed from the final dataset as most likely are false calls. This can be achieved either using variant quality score recalibration (VQSR) or applying hard filters. The first, is generally preferred as it is an unbiased filtering based on a large number of validated variants that train the algorithm (Van der Auwera et al., 2013). The latter, is usually used when such known variants are not available and includes SNPs removal based on annotation parameters assigned to each SNP during read-mapping and variant calling. Common filters include: Low complexity, maximum depth, allele balance, double strand, Fisher strand, and quality filter (Li & Wren, 2014; Van der Auwera et al., 2013), as well as mapping quality (MQ) (Li et al., 2008). Each mapping algorithm calculates the MQ score differently (Ruffalo, Koyutürk, Ray, & LaFramboise, 2012) for which scores should not be compared between programs. The application of hard filters, however, can bias the Site Frequency Spectrum by excluding low-frequency variants and is limited by the absence of guidelines to select which annotations or cut-off values should be applied to a given data. The appropriate choice of cut-off values is a function of the data. The recommendation is thus to test different parameter combinations and thresholds to optimize these filters. This forum, https://software.broadinstitute.org/gatk/guide/article?id=6925, can offer some insight on hard filtering using GATK. Additionally, SNPs within low-complexity regions should be removed as these regions are troublesome for read mapping and SNP calling (Li & Wren, 2014). The final VCF file after quality control will be ready for downstream analyses.

Variant annotation

Sequence ontology terms can be annotated to variants in a VCF file using for instance, Vcfanno (Pedersen et al., 2016), ANNOVAR (Yang & Wang, 2015), or SNPeff (Cingolani et al., 2012a) programs.

Variant validation

Variants detected from WGR data should be treated as putative polymorphisms, especially in Pool-seq and lcWGR. SNP genotyping PCR-based methods can be used for SNP validation. PCR amplification and Sanger sequencing can be used for SVs validation.

For additional guidelines on how to obtain high-quality variants from high-throughput sequencing data see Van der Auwera *et al.* (2013); Pfeifer (2017).

2.8. Box 3: Limitations, sources of error and bias and potential solutions of the Pool-seq approach

Individual genotypes are missed

Pool-seq has been used for the estimation of genetic differences among populations, the detection of outlier loci, and for mapping genetic variation underlying phenotypic traits (see Table 2.2). This method, however, prevents the identification of individual reads as only one barcoded-library is prepared for the pooled DNA of individuals (Schlötterer et al., 2014). This has four important implications: (*1*) errors during library preparation or shotgun sequencing that could affect the homogeneous contribution of individual DNA to the final dataset cannot be detected; (*2*) as individual genotypes are lost, presence of migrants in the sample cannot be evaluated; (*3*) individual haplotypes and linkage disequilibrium (LD) cannot be assessed (e.g. the LD method for the estimation of effective population size cannot be used); (*4*) only total allele frequencies can be calculated for a given pooled DNA. These factors can bias the population-level allele frequency estimates and limit population genetics analyses, leading to concern about the suitability of the Pool-seq method for population genomic studies (Anderson et al., 2014; Therkildsen & Palumbi, 2017).

Several ways to mitigate these concerns have been proposed. Knowledge of population structure can help avoid accidentally pooling individuals of different origin. Mixed aggregations, however, can also be indirectly detected from Pool-seq data, and conveniently excluded of posterior analysis if required, as they exhibit extremely large branch lengths in a phylogenetic tree based on pairwise genetic distances (see Lamichhaney *et al.* (2017) Figs 2 and 5: WFB location). In the absence of prior information on population structure, ecological and biological knowledge (e.g. timing of reproduction, maturity stage at time of sampling, size and age, etc.) can help infer the local origin of individuals. Uneven representation of individual DNA samples in the final pool can be minimized following stringent laboratory procedures (Box 2). The limitations above concerning the lack of haplotype and hence individual genotypes and linkage disequilibrium information can be overcome by the selection of a subset of informative SNPs identified from the Pool-seq data and the subsequent genotyping of a number of individuals per population at these informative SNPs using PCR-based or NGS-based genotyping methods. This SNP validation step should be performed anyways as part of a good practice protocol for the Pool-seq workflow (Box 2).

lcWGR consists of individually barcoded DNA samples that allow the control of most of the concerns raised above for Pool-seq, except for the fact that individual genotypes cannot be reliably called. Instead, the raw sequence data is used to calculate per-site genotype likelihoods (GLs) based on the reads of multiple individuals in a sample. The sample allele frequency per site is calculated from the GLs and is used for posterior population genetics analyses (Korneliussen et al., 2014; Nielsen et al., 2012, 2011).

Allele frequency estimation is susceptible to multiple factors

A main weakness of Pool-seq is the uneven distribution of reads due to technical and computational artifacts (Anderson et al., 2014). Depth of coverage is the basis for variant detection where read counts are used for the estimation of allele frequencies in individual SNPs. Therefore, uneven coverage not resulting from biological processes can greatly increase false SNP calls or leave out informative SNPs, biasing downstream analysis and interpretation (Sims et al., 2014). Read depth of coverage can be affected by several factors including (*1*) duplicates and GC bias produced by PCR during library preparation (Sims et al., 2014); (*2*) amplification bias of NGS sequencing technologies (Goodwin et al., 2016); (*3*) incompleteness of the reference genome especially at repetitive regions that could produce misalignments and false calls (Li & Wren, 2014); (*4*) structural variations (e.g. CNVs, chromosomal inversions, transposable elements) can inflate allele frequency estimates (Schlötterer et al., 2014); and (*5*) poor read mapping can result in read misplacement or missing of divergent reads (Kofler, Langmuller, Nouhaud, Otte, & Schlotterer, 2016; Kofler, Orozco-terWengel, et al., 2011; Schlötterer et al., 2014). Factor (*1*) can be controlled by using PCR-free sequencing library preparation kits when plenty DNA is available; if this is not the case, then PCR duplicates can be easily removed using bioinformatic tools. Factor (*2*) is out of researchers' control but to minimize lane-to-lane variation, DNA libraries should be spread in different lanes (Ross et al., 2013). Factor (*3*) can be minimized by including an indel realignment step before variant calling (Van der Auwera et al., 2013), or by excluding variants detected in and around INDELs (Kofler, Orozco-terWengel, et al., 2011) and repetitive regions. The effect of factor (4) can be minimized by excluding structural variants of variant calling (Kofler, Orozco-terWengel, et al., 2011), though this type of variation should be included in genome scans for the detection of outlier loci or the characterization of the genetic basis of traits. And factor (*5*) requires the optimization of read mapping by selecting an adequate algorithm and running parameters that minimize misalignments.

Mapping short reads against a reference sequence is a difficult task considering the large number of reads to process and the computational challenge of determining their exact position in the genome. For instance, repetitive sequences are difficult to map because their sequence can be present in multiple positions, and the genome could also have assembly errors. Moreover, there may be sequence differences between subject reads and the reference genome, because the latter is usually not representative of the species genetic diversity (usually a genome sequence is built from DNA of one or a few individuals). (reviewed by Treangen & Salzberg (2012), Phan *et al.* (2014), Laehnemann *et al.* (2016)). Thus, the optimal choice of a mapping algorithm depends on the sequence data structure (e.g. repetitive content) and degree of divergence between the reads and the reference genome. Multiple mapping algorithms exist which are optimized for different levels of sequence divergence. For example, BWA excels in the mapping of reads with low divergence (<2%) (Li, 2013), however, the running parameters can be modified for mapping divergent reads (Kofler *et al.* 2011). In contrast, Stampy is better for mapping reads with high-divergence as the user can input a substitution rate value (Lunter & Goodson, 2011). Default settings of mapping algorithms are usually optimized for specific datasets and highly curated genomes (e.g. humans), thus running a mapper with default parameters may not be ideal for all datasets as this can produce multiple spurious outlier loci and false positive SNPs (Kofler et al., 2016). Time investment into optimizing read mapping parameters before SNP calling is thus recommended. Unfortunately, there is no golden rule applicable to all datasets but, in general, the idea is to experiment with parameters (e.g. mismatch and gap opening penalties), conduct read mapping, compare alignment statistics, and so on, following a multi-dimensional optimization process. The ideal parameter combination will maximize the number of properly paired-end reads while minimizing presence of discordant mates, singletons, and ambiguous mappings. This rationale is explained in detail for RAD-seq data by Jonathan Puritz (https://github.com/jpuritz/Winter.School2017/blob/master/Exercises/Day%201/Mapping%20Exercise.md).

Despite the various factors potentially affecting depth of coverage in Pool-seq data, numerous studies have demonstrated the method can produce reliable population-level allele frequency estimates (Fracassetti et al., 2015; S. Lamichhaney et al., 2017; Martinez Barrio et al., 2016; Jingwen Wang et al., 2016).

Rare and low-frequency variants are hard to detect

It is generally assumed that Pool-seq only allows for the discovery of common variants of large effect as different factors can affect read coverage, making it difficult to distinguish low frequency variants from sequencing error. Recent studies by Wang *et al.* (2016), however, challenge this idea as they recovered rare variants from high-depth Pool-seq data of Bull Terrier dogs (average 130x) and humans (average 150x). Minor allele frequency errors were evaluated using three variant calling programs, SAMtools, GATK (ploidy setting) and Freebayes (ploidy setting). A good proportion of rare SNPs identified from the pooled data were validated through individual genotyping of several samples using the MassARRAY System (Agena Bioscience, U.S.) and the Illumina SNP array (Illumina, U.S.) systems. Thus, Pool-seq can be a fast and affordable initial approach for the assessment of rare variants in large-scale association studies.

2.9 Box 4: Limitations, sources of error and bias and potential solutions of the lcWGR approach

Genotype likelihood (GL) values can vary

Genotype likelihoods are the foundation of the statistical framework for population genetics inference from low-coverage sequencing data (Buerkle & Gompert, 2013; Nielsen et al., 2012, 2011). GLs and additional analyses for this type of data are implemented in the programs ANGSD (Korneliussen et al., 2014; Nielsen et al., 2012) and NgsTools (Fumagalli, Vieira, Linderoth, & Nielsen, 2014). GLs per polymorphic site are estimated based on the observed sequence reads covering the site from a given sample of individuals, and the reads' PHRED quality scores. The base-error rate estimation method differs among GL models and error rates can be fixed or estimated from the quality scores or the sequence data. The 4 models for GLs and the 2 models for base-error rate calculations implemented in ANGSD are described by Korneliussen *et al.* (2014). Previous studies indicate the GL models can generate different results in some circumstances (Korneliussen *et al.* 2014). Thus, the choice of model can potentially introduce bias (http://www.popgen.dk/angsd/index.php/Genotype_Likelihoods)(Korneliussen *et al.* 2014) but models have not been compared nor, to our knowledge, has the procedure for model selection been discussed in the literature thus far. The base-error rate methods differ in how they model the error structures in the data, which is important in the calculation of GLs and downstream analyses. If the modelling misses error sources, then the GLs are likely to be biased and verification from mathematical derivations that the proper error structure in the data has been correctly incorporated is not straightforward.

Genotype likelihoods can be affected by: (*1*) accuracy of base-calling and quality score (Fumagalli et al., 2013), (*2*) read coverage distribution and filtering, (*3*) sample size and individuals included in the sample, (*4*) how accurately model assumptions are met, including (*5*) the assumption that markers are diallelic and organisms are diploid. (*1*) Sequencing error is extensive in available sequencing platforms (Goodwin et al., 2016), and the per-base quality scores obtained directly from the sequencing machines also have errors. Therefore, lcWGR data should be obtained using sequencing platforms that offer the lowest error possible, and the per-base quality scores need to be recalibrated before data analysis. It is also important to consider that the GL calculation assumes independence among reads, which may be violated in the presence of alignment error or PCR artifacts (Nielsen et al., 2011). (*2*) It is currently almost impossible to obtain an even read coverage distribution across the genome and among individuals. This is a result of the sequencing chemistry itself (usually following a Poisson distribution), and of laboratory procedures that can skew the representation of individual DNA samples added to a flow cell (errors in pipetting or DNA quantification, and variability in fragment sizes during library preparation). In addition, a very low sequencing depth per individual (<2x) could also limit the possibility of sequencing both alleles in a diploid organism (Fumagalli, 2013). Similarly, the joint effect of read filtering (including *in silico* coverage cut-offs across individuals in a sample), reference genome quality and completeness, and read mapping can also bias the individual read representation for a particular locus in the final dataset. Thus, a varying proportion of missing data per individual could be expected in lcWGR datasets, which implies polymorphic sites are covered by reads from a varying proportion of individuals in a sample. This could be a problem when reads of a small number of individuals are supporting a particular site because they may not be representative, potentially biasing GL estimates and downstream analyses. An excess of missing data can also bring convergence problems that impact the accuracy of many calculations including the individual admixture analysis implemented in NGSadmix (Skotte, Korneliussen, & Albrechtsen, 2013). In GWAS in humans, missing genotypes in extremely low-coverage sequencing data (0.1–0.5x/sample for 909 individuals) have been treated using imputation methods that rely on the availability of a set of known haplotypes (Pasaniuc et al., 2012). For non-model species such set is commonly absent, implying that imputation is not an option, unless individuals are highly inbred (Wang et al., 2016). In conclusion, lcWGR studies require a relatively even read coverage distribution among individuals (see Therkildsen & Palumbi, 2017). The removal of sites with large amounts of missing data (>80%) (Skotte et al., 2013) and a target read depth that assures both alleles are sequenced in a diploid organism (>2x) are recommended (Fumagalli, 2013). Similarly, (*3*) sample size and the actual individuals included in a sample determine the alleles and sites assessed as well as the population structure and admixture estimates. Simulation studies have shown good accuracy in population genetics parameters using a low depth per individual (~1-2x) as long as the number of individuals in the sample is large (Buerkle & Gompert, 2013; Fumagalli, 2013). A minimum number of individuals per sample or population has, however, not been proposed. Presumably this is because such number depends on several factors including the species genomic architecture (e.g. LD decay, recombination rate, mutation rate) and effective population size, and the funding available, among other variables. However, it would be useful to have at least some reference obtained from real data that can be used in the design of sampling programs. The genetic makeup of individuals composing a sample would determine the alleles and sites evaluated, therefore it is important to ensure individuals are not closely related to avoid inflated estimates of genetic differentiation. When comparing populations, it is fundamental to verify the degree of individual admixture before performing GLs calculation as the presence of mixed individuals will likely bias the alleles represented and thus, subsequent analyses and interpretation. Individual admixture can be verified using the program NgsAdmix (Skotte et al., 2013). (*4*) The accuracy of GLs and subsequent metrics depend on the fulfilment of the assumptions made in the mathematical models, for instance, independence among reads for GLs calculation (Nielsen et al., 2011), independence among sites and Hardy-Weinberg Equilibrium for calculation of the likelihood function for the site frequency spectrum (Nielsen et al., 2012), and independence among individuals for the estimation of allele frequencies (Kim et al., 2011). Finally, (*5*) the current four GL models were developed for diallelic markers in diploid organisms implying that lcWGR cannot currently be applied to non-diploid species or pooled DNA data. The method is thus limited to the assessment of genetic variation in SNP loci; INDELs are included in the models but not used for posterior analyses (Korneliussen et al., 2014). Putative structural variants could be detected, however, from high-density SNPs as they facilitate the identification of sweeps to a fixed allele, as observed in a cryptic subgroup of *Anopheles gambiae* s.l. (GOUNDRY) where ~500 SNPs allowed the detection of a putative large inversion (1.67-Mb) on the X chromosome (Crawford et al., 2016).

Few programs accept GLs

Many traditional population genetics software packages require individual genotypes as input data, limiting the possibility of use for the analysis of lcWGR data. To solve this problem, the software ANGSD includes several genotype callers (Korneliussen et al., 2014).

It is a relatively new approach

lcWGR was implemented in the 1000 Genomes Project (2008-2015) (Auton et al., 2015), where the initial statistical models, file formats, and programs were developed. Its use has been restricted mostly to humans and, more recently extended to agricultural and other non-model species. However, only few laboratories have used this approach thus far, explaining perhaps the scarcity of software available for this type of data. As this approach gains popularity, new computer packages are likely to be developed.

2.10 Box 5: Considerations and limitations of genome scans for detecting selection and inferring local adaptation

Genome scans are currently one of the most popular methods for the detection of selection from genomic data (Barrett & Hoekstra, 2011; Hoban et al., 2016; Martin & Jiggins, 2013; Pardo-Diaz et al., 2015; Vitti et al., 2013), in particular for the detection of directional selection.

Genome scans encompass a comprehensive survey of the genome of individuals and populations to identify molecular patterns that presumably result from selection, for example, increased linkage disequilibrium (LD) (i.e. long haplotype blocks) and reduced variation around beneficial mutations, abundance of rare alleles in the population site-frequency spectrum, significant allele frequency differences between populations under contrasting selection regimes, among others (Ellegren, 2014; Jensen, Foll, & Bernatchez, 2016; Vitti et al., 2013). Currently, genome scans are based on: (*1*) information on the physical distance (LD blocks) between loci for the detection of *selective sweeps* (Messer & Petrov, 2013; Vatsiou, Bazin, & Gaggiotti, 2016), or on (*2*) knowledge of allele frequency differences between unlinked loci using (*i*) outlier loci tests, or (*ii*) Genetic-Environment Association (GEA) analysis (Bernatchez, 2016; Gagnaire & Gaggiotti, 2016; Hoban et al., 2016). (*1*) *Selective sweeps* can be "hard", when the beneficial mutations of large effect are new and increase in frequency in the population in a short period of time, or they can be "soft", when they comprise numerous alleles of small effect that were already present in the population or resulted from recurrent independent mutational events (Messer & Petrov, 2013). The genetic signal for hard sweeps is generally easier to detect in genomic data as it includes elevated differentiation at particular loci, whereas the soft sweeps signal can be confounded with the genomic background because the genetic changes involved are more subtle. The detection of either type of selective sweeps and their distinction requires the use of specific statistical tools (reviewed by Messer & Petrov (2013) and Vatsiou *et al.* (2016)). (*2*)(*i*) Outlier loci tests rely on the detection of putatively selected loci showing elevated levels of differentiation with respect to expectations under a neutral model and usually involve a window-based approach where summary statistics (e.g., F_{ST}) are estimated and averaged for all the variants present in the window (Barrett & Hoekstra, 2011). Window size choice is thus important as, by modifying the number and the range of physical separation between variants, the outcome could change. A very large window can lead to an overestimation of outlier loci due to false positives, whereas an excessively narrow window can lead to an underestimation of outlier loci by excluding from the window sections of low differentiated genomic background useful for outlier detection. Window size selection should thus account for average genome-wide LD or, in its absence, for the relative genomic position and separation of variants, population polymorphism level (Hoban et al., 2016), or could also be statistically inferred as breakpoints in the genomic data (Beissinger, Rosa, Kaeppler, Gianola, & de Leon, 2015). (*ii*) GEA analysis uncovers putatively adaptive loci through the comparison of the genetic variation between populations adapted to contrasting environments (Barrett & Hoekstra, 2011; Martin & Jiggins, 2013; Pardo-Diaz et al., 2015). The choices of populations and environmental variables to compare are relevant as they define the power of such comparison. The population spatial resolution and the temporal resolution of environmental variables need to be considered as they will directly affect the correlations between outlier loci and environment (Hoban et al., 2016).

Other limitations of genome scans include the fact that: (*i*) some outlier loci may not themselves be under selection but may instead be located in the proximity of a causal mutation, implying that follow-up functional molecular studies testing the phenotypic effect of an outlier locus are needed to consider it adaptive (Barrett & Hoekstra, 2011); (*ii*) signals of selection can be confounded with footprints of demographic history (e.g. populations structure) (Tiffin & Ross-Ibarra, 2014); (*iii*) mutation and recombination rates, type and strength of selection, and the genetic architecture of adaptive traits all modulate the genomic heterogeneity of a species, restraining the capacity to detect the genetic basis of adaptation (Haasl & Payseur, 2016). Several solutions have been proposed to overcome these limitations (and others not commented here) and are described in more detail in the referenced papers.

The genetic architecture of an adaptive trait refers to the total number of genes contributing to a given character, their location, effect size and heritability, and the interactions among them (i.e. additivity, epistasis, dominance, pleiotropy) and with the environment (genetic-environment interactions) (Gagnaire & Gaggiotti, 2016; Hansen, 2006). This architecture determines the range of allele frequency changes in a population responding to selection (Gagnaire & Gaggiotti, 2016). For instance, in oligogenic traits (i.e. characters underlined by a few large effect genes) a large shift in allele frequencies is expected, whereas a small change is assumed in polygenic traits where characters result from the interaction of multiple small effect genes. The power of genome scans to identify the genetic basis of quantitative traits based on allele frequency methods therefore depends on how much of the total adaptive genetic variation of a trait is explained by the summed effect of the outlier loci (Berg & Coop, 2014; Gagnaire & Gaggiotti, 2016). Sampling design in a complex environmental landscape (Lotterhos & Whitlock, 2015), as well as sample size and number and density of markers thus play an important role in our capacity to reveal the genetic basis of adaptive traits, especially for traits of polygenic architecture (Gagnaire & Gaggiotti, 2016) which are common in nature (Bernatchez, 2016; Rockman, 2012). The choice of sequencing technique for the collection of genomic data is thus not trivial as it will define the proportion and type of genetic variation in a genome that has been sampled and thus the potential inclusion or exclusion of relevant loci (Hoban et al., 2016). The marker density required for a genome scan should thus ideally account for the average LD decay to ensure that most variants contributing to a trait are surveyed (Gagnaire & Gaggiotti, 2016).

The importance of a proper planning of sequencing approaches for the study of local adaptation has been brought into sharp focus by a recent debate on the power of RAD-seq for the detection of adaptive genetic variation in natural populations. Lowry *et al.* (2017) used computer simulations to argue that a large proportion of putatively adaptive loci are missed by RAD-seq studies because typically only a small fraction of the genome is surveyed with loci being too widely spaced. The problem is expected to be more important for species with large census and effective population sizes which tend to exhibit short LD blocks (high LD decay and high recombination rate). On the other hand, McKinney *et al.* (2017) claim that a properly designed RAD-seq study that takes advantage of the flexibility of restriction enzymes, can provide enough markers to achieve high genome coverage. Authors also argue that LD and recombination are not homogeneous across the genome and adaptation signatures can frequently result in extended LD blocks [or genomic islands of divergence, Nosil *et al.* (2009)] spanning several Kb, that can be easily screened with the low marker density provided by RAD-seq, regardless of effective population size (McKinney *et al.* 2017). In addition, Catchen *et al.* (2017) argue that even with short LD blocks, RAD-seq has been successful at detecting adaptive loci (e.g., the *Eda* locus in three-spine sticklebacks), and that endangered species usually exhibit small effective population sizes for which large LD blocks should be expected. They also pointed out that some studies may be focused on detecting adaptive differentiation only at the fraction of the genome sampled and not at all adaptive loci present. Finally, Lowry *et al.* (2017b) emphasize that the average LD block size and variation of recombination rate along the genome is usually unknown for non-model species, thus, it is difficult to estimate *a priori* the minimum marker density required for a RRS approach. They propose RADseq studies aiming to detect adaptive loci should follow these basic principles: 1) report the limitations of a given study, 2) in the absence of a reference genome or linkage map, efforts should be centered first on obtaining this information, 3) complement genome scans with alternative sources of evidence (i.e. field experiments or functional molecular tests) that demonstrate the phenotypic effect of outlier loci, and 4) conduct pilot tests to assess the viability of a sequencing experiment plan.

Although previous knowledge on the extend of LD decay or recombination rate is generally lacking for non-model species, LD block size can be estimated from a dense genetic map or from RAD-seq data with a reference genome (Catchen *et al.* 2017). Fig. 2.4 synthesizes the decision making process for sequencing approach as a function of the existence of a reference genome; expected LD block size; whether the interest is on neutral, adaptive or both types of genetic variation; relative cost; and type of genetic variation assessed.

Sample size is another practical consideration that has not received much attention in population genomics, in part because of the perception that large sample sizes are not required because of the very large number of markers genotyped per individual. This may be fine in some cases (studies in Table 2.2 with small sample sizes), but in general, the establishment of a minimum sample size depends on the research question and the genetic architecture of the focal species. For example, in human genetics studies for the detection of small effect variants associated with rare diseases, even the screening of thousands of individuals has not provided enough power to detect and track such variation (Agarwala, Flannick, Sunyaev, & Altshuler, 2013; Lee, Abecasis, Boehnke, & Lin, 2014; Moutsianas et al., 2015). Therefore, a presumably large sample size may be required in studies of non-model species aiming to identify adaptive variation associated to polygenic traits.

In conclusion, multiple considerations need to be taken into account when planning genome scans for the detection of signatures of selection and local adaptation from genomic data. The choice of minimum sample size and sequencing technique for the collection of genomic data should respond to the research question and should be informed by the expected (or ideally verified) LD block size (or physical LD), and the genetic architecture of a given phenotypic trait. These factors will determine the marker density needed for the successful detection of putatively adaptive variation in the genome.

2.11 Acknowledgements

We thank Leif Andersson, Sangeet Lamichhaney, Nima Rafati, and Alvaro Martínez-Barrio (Uppsala University) for training on Pool-seq. Thanks to Martin Fischer, Simone Fior, Erik Garrison, and Anders Albrechtsen for helpful discussions on Pool-seq and lcWGR during the 2017 Workshop on Bioinformatics for Adaptation Genomics (ETH-Zurich). Thanks to three anonymous reviewers that helped improve this manuscript. A.P.F.P. thanks to the Vanier Canada Graduate Scholarship, the President’s Award of Dalhousie University, the NSERC Strategic grant earned by D.E.R., and the Nova Scotia

Graduate Scholarship for funding. A.P.F.P. and D.E.R. thank the Killam Trust. The study was funded by NSERC Discovery and Strategic grants to D.E.R.

2.12 Author contribution

A.P.F.P. and D.E.R. conceived the study. A.P.F.P. compiled the papers, analyzed the data and wrote the first draft of the manuscript with input from D.E.R. Abbreviations of names as described in the statement of co-authorship (page 7).

2.13 Tables

Table 2.1 Comparison of requirements and different aspects of RAD-seq, Pool-seq and lcWGR approaches for population genomics studies.

Aspect	**RAD-seq (original protocol)**[1]	**Pool-seq**	**lcWGR**
Expected percentage of the genome covered	~1-5 %	> 70 % (depending on reference genome completeness)	> 70 % (depending on reference genome completeness)
DNA quality	High molecular weight	High molecular weight	High molecular weight
DNA quantity per sample	> 200 ng	> 1 µg per pool (for TruSeq PCR-free kit) > 200 ng per pool (for TruSeq Nano kit)	> 50 ng (for Nextera kit), although it could be less[2]
Need of a reference genome	Not indispensable but desirable	Required	Required
Type of library	Usually non-commercial	Commercial	Commercial
Library insert size	~350-550 bp (standard library)	~350-550 bp (standard library). For detection of large structural variants with short-read sequencing, ~2-20Kb (mate-pair library)	~350-550 bp (standard library)
Cost of library preparation	~USD$ 5-10/individual[3]	~$USD 46/pool[3]	~$USD 6/individual[2,3]
Minimum number of individuals per population	Usually ≥ 20[4]	≥ 50[5]	Number not established but usually ≥ 50
Popular sequencing platform	Illumina MiSeq and HiSeq, IonProton	Illumina HiSeq	Illumina HiSeq
Type of sequence reads	Single-end or paired-end reads, ≥ 100 bp per read	Paired-end reads, ≥ 100 bp per read	Single-end or paired-end reads, ≥ 100 bp per read
Minimum sequencing depth of coverage	High coverage: ≥ 20x per individual for diploids and higher depth for polyploids[7]	High coverage: ≥ 50-100x per pool[5]	Low coverage: ~1-4x (per individual) depending on ploidy, e.g. 2x recommended for diploid organisms[8]
Minimum computing resources	Desktop computer with multicore processor (≥ 24) and ≥ 64GB RAM	QC and trimming of raw reads, read mapping, duplicate marking, and read sorting of small to large size genomes (~1Gb) can be performed in a desktop computer with multi-core processor (≥ 32) and ≥ 128-256GB RAM. Greater computing resources (i.e. computer cluster or computing facilities in the cloud) are required for larger gen omes (≥ 1Gb), specially for SNP calling	
Computer data storage[9]	MBs per sample (depending on genome size and depth)	GBs per pool (depending on genome size and depth). For example, for one population of Atlantic herring (genome size ~900Mb, 50 individuals per pool, depth 50x per pool): raw data ~50GB, clean data ~40GB, BAM file ~35GB, gVCF file ~40GB	GBs per individual (depending on genome size and depth)
Programming skills	Basic-Intermediate	Intermediate	Intermediate

Aspect	RAD-seq (original protocol)[1]	Pool-seq	lcWGR
Expected number of SNP loci per sample	Thousands (without reference), hundreds of thousands (with reference genome)	Millions	Millions
Type of variant assessed	Mostly SNPs, inversions when a reference genome is available	SNPs, INDELs, large SVs, CNVs	SNPs, INDELs detected but not used in software, some SVs (depending on genome coverage per individual)
Type of genetic variation screened	Mostly neutral and sometimes functional (depending on marker density)	Neutral and functional	Neutral and functional
Output data obtained per sample	Individual SNP genotypes (based on coverage)	Population-level allele frequency per SNP (based on read counts per variant site)	Population-level genotype likelihood (based on reads of multiple individuals in a population)
Possibility to do individual-based analyses	Yes	No, individual information is missed during library preparation	No, reliable individual SNP calls cannot be obtained from low coverage data
Scalability (+): most positive feature (−): most negative feature	High (+) Cheaper method than Pool-seq and lcWGR enabling the analysis of numerous individuals and populations (−) Low marker density limits the capacity to detect adaptive variation	High (+) Many individuals per population can be mixed in one pool for the same library preparation cost (−) Sequencing depth should be increased accordingly. More expensive than RAD-seq but cheaper than lcWGR	High (+) Low depth per individual enables the analysis of numerous individuals per population (−) More expensive than Pool-seq, especially for >50 individuals analyzed per population

Abbreviations: GB = gigabytes, MB = megabytes, CNVs = copy number variations, SVs = structural variations, INDELs = insertions and deletions, SNPs = single nucleotide polymorphisms, QC = quality control.

[1] Baird *et al.* (2008), as cited in (Andrews et al., 2016).
[2] When following the protocol by Therkildsen & Palumbi (2017).
[3] Price for March 2017.
[4] Hohenlohe *et al.* (2010)
[5] Schlötterer *et al.* (2014)
[6] Fumagalli (2013)
[7] Andrews *et al.* (2016)
[8] Additional data storage space is required for temporary files generated during data analysis, which can exceed (2-3x) the size of the final data file (e.g. for Atlantic herring, SAM file ~150GB)

Table 2.2 Key methodological aspects of case studies using WGR in different conservation biology topics.

Organism	Main findings	WGR method	Sample size	Sequencing design and output	Number of markers	Software for mapping [parameters]	IR/ BQSR	SNP calling software	Ref.
Topic: Phylogenomics, hybridization and taxonomical species resolution									
Modern birds (Neoaves)	Genome-scale phylogenetic tree of 48 species representing all orders of Neoaves	huWGR	48 species, 1 ind. per species	Illumina HiSeq 2000, Roche 454, pair-end libraries of different insert sizes, 24x-160x	Whole-genome DNA sequences	SOAPdenovo[1]	-	-	Jarvis et al. (2014)
Snub-nosed monkey (*Rhinopithecus roxellana, R. bieti, R. brelichi, R. strykeri*)	Genome-scale phylogenetic relationships that indicate functional evolution and leaf-eating dietary adaptations	huWGR	4 species, 1 ind. per species	Illumina HiSeq, 30x huWGR, 146x *de novo* assembly, multiple paired-end and mate-pair libraries spanning size range of 180 bp to 20 kb	Whole-genome DNA sequences	SOAPdenovo[1]	-	-	Zhou et al. (2014)
North American wolves	Lack of unique ancestry in eastern and red wolves	huWGR	28 ind.	Illumina HiSeq, paired- and single-end sequencing libraries, average insert size of 300 to 500 bp, 4-29x	5,424,934 SNPs	Stampy[2]	NI	ANGSD[6]	vonHoldt et al. (2016)
Common milkweed *(Asclepias syriaca*)	Marker development including complete chloroplast genome, a nearly complete rDNA cistron and 5S rDNA sequence, a partial mitochondrial genome sequence, and some single copy ortholog genes	lcWGR	1 ind.	Illumina GAII, 1 lane, 40 cycles, 0.5x	-	-	-	-	Straub *et al.* (2011)
Beardtongue plant (*Penstemon. centranthifolius, P. grinnellii*)	Primer design of phylogenetic markers in the plant genus based on annotation and gene prediction of more than 10 000 contigs	lcWGR	1 ind. of each species	Roche 454 platform, ~0.005x–0.007x	-	-	-	-	Blischak *et al.* (2014)
Yellow baboons (*Papio cynocephalus*), Anubis baboons (*P. anubis*), and hybrids	Genetic differentiation between parent taxa, complex admixture history involving intermittent but multiple hybridization events that did not indicate fitness reduction in hybrids	lcWGR	46 ind. in total	Illumina HiSeq, for huWGR, 2.09x-19.6x, paired-end 100 bp reads	~2.1 million	BWA-MEM[4] [minimum seed length of 20]	Yes/Yes	GATK *UnifiedGenotyper*[7], (discarding indels)	Wall *et al.* (2016)
Topic: Population structure and admixture									
Giant pandas (*Ailuropoda melanoleuca*)	Multiple demographic events including population expansion, bottlenecks, and divergence, human activities most likely contributed to decline in the last 3000 years	lcWGR	34 ind.	llumina HiSeq2000 platform, and 100-bp paired-end reads, 4.7x	13,020,055 SNPs	BWA-ALN[3] [NI]	NI	SOAPsnp[13]	Zhao et al. (2012)

Organism	Main findings	WGR method	Sample size	Sequencing design and output	Number of markers	Software for mapping [parameters]	IR/ BQSR	SNP calling software	Ref.
Killer whale (*Orcinus orca*)	Differentiations between pairs of contemporary allopatric and sympatric ecotypes most likely are the consequence of ecological divergence and genetic drift resulting from bottlenecks experienced during past founder events	lcWGR	(lcWGR): 48 ind., (huWGR): 2 ind.	Illumina HiSeq 2000 platform using single-read 100-bp chemistry, 2x	603,519 variant sites	BWA[3] [NI]	NI	ANGSD[6]	Foote *et al.* (2016)
Western palearctic black-and-white flycatchers of the genus *Ficedula* (*F. speculigera*, *F. albicollis*, *F. hypoleuca*, *F. semitorquata*)	Most recent common ancestor of the four species dates back to 1–2 million years (Mya) ago and each species followed separated evolutionary paths involving population growth, decline (~100–200 thousand years ago) and expansion	huWGR	200 ind. in total	HiSeq 2000, paired-end 100 bp, insert size ~450 bp, ≥20x	NI	BWA[3] [NI]	Yes/Yes	SAMtools[8]	Nadachowska-Brzyska *et al.* (2016)
Cattle (*Bos taurus*)	Historical events such as domestication or modern breeding are related with population decline	huWGR	15 to 25 ind. from each of 4 breeds	Illumina	NI	BWA[3] [NI]	NI	SAMtools[8]	Boitard *et al.* (2016)
European dark honey bee (*Apis mellifera mellifera*) and two introduced honey bee subspecies (*A. m. carnica* and buckfast)	Genetic differentiation between subspecies that coincides with geography. Observed presence of admixed individuals in protected areas	huWGR	151 drones	Illumina HiSeq2500, pair-end 2x125 bp reads, 10x	3.375 million SNPs	BWA-MEM[4] [NI]	Yes/Yes	GATK-*UnifiedGenotyper*[7]	Parejo et al. (2016)
Arabidopsis halleri	Weak genetic differentiation among populations. SNPs more informative than microsatellites about genome-wide genetic diversity	Pool-seq	20 ind. in 9 populations	(Pool-seq): Illumina HiSeq2000, paired-end 2x100 bp reads, 250–300 bp insert size, 60.7x (range 52.7 to 69.3x per pool)	2,178,204 SNPs	BWA-ALN[3] and sampe [allowing 10% mismatch]	NI	SAMtools[8]	Fischer *et al.* (2017)
Almond (*Prunus dulcis*) and peach (*P. persica*) plants	Almond genome-wide nucleotide diversity was ~7-fold higher than in peach, excess of rare alleles likely consistent with a recent population expansion event, no evidence of population bottleneck related with domestication, and a strong genetic differentiation between almond and peach	lcWGR	13 ind. per species	Illumina HiSeq2000, 100 bp paired-end reads, depth averaged 15.8x (4.7x to 34.6x) in almond and 19.7x (11.2x to 35.4x in peach	NI	BWA-MEM[4] [minimum seed length of 10 and internal seed length of 2.85]	NI	ANGSD[6]	Velasco *et al.* (2016)

Organism	Main findings	WGR method	Sample size	Sequencing design and output	Number of markers	Software for mapping [parameters]	IR/ BQSR	SNP calling software	Ref.
Topic: Signatures of selection, genetic basis of phenotypic traits, and local adaptation									
Darwin finches (*Geospiza* spp.)	A 240 kilobase haplotype encompassing the ALX1 gene that encodes a transcription factor affecting craniofacial development is strongly associated with beak shape	huWGR	200 ind. distributed in 15 species	Illumina Hiseq2000, 2x100 bp paired-end reads, insert size ~400 bp, 10x coverage	44,753,624 SNPs	BWA[3] [default]	Yes/Yes	GATK-*UnifiedGenotyper*[7]	Lamichhaney *et al.* (2015a)
Atlantic salmon, (*Salmon salar*)	Locus that maintains variation in age at maturity	huWGR	54 wild populations, ~500 ind., (SNP array), and 32 ind. from 7 populations (huWGR)	(huWGR): HiSeq2500, paired-end reads 2x125 bp, average coverage 18x (8x to 32x)	208,704 SNPs	BWA-MEM[4] [default]	No/No	FreeBayes[9]	Barson *et al.* (2015)
Atlantic salmon, (*Salmon salar*)	Locus vgll3 controls age at maturity in wild and domesticated Atlantic salmon males	Pool-seq	20 ind. per river and phenotype	Illumina HiSeq2000 platform, 12.3x per pool	4,326,591 SNPs	Bowtie2[5] [no soft clipping, end-to-end mode, seed length 18, only 1 mismatch]	NI	SAMtools[8]	Ayllon *et al.* (2015)
Red siskins (*Spinus cucullata*), common canaries (*Serinus canaria*), and”red factor” canaries	Gene encoding a cytochrome P450 enzyme, CYP2J19, is the ketolase that mediates red coloration in birds	Pool-seq	12 to 39 ind. per pool	Illumina Hiseq2500, 2x 100 bp reads, 19.3x per pool	9,414,439 SNPs	BWA-MEM[4] [default]	Yes/NI	SAMtools[8] and VarScan2[10]	Lopes *et al.* (2016)
Marine midge (*Clunio marinus*)	Locus calcium/calmodulin-dependent kinase II.1 (CaMKII.1) splice variants strongly associated with circadian timing	Pool-seq	5 populations, 100-300 ind. each	Illumina HiSeq2000, paired-end 2x100bp reads, 0.2-0.4Kb insert size, 68x to 251x per pool	1,010,052 SNPs	BWA-ALN[3] and sampe [maximal insert size 1500bp]	Yes/No	Poopolation 2[11]	Kaiser *et al.* (2016)
Domestic and wild rabbits (*Oryctolagus cuniculus*)	Mapped genes affecting brain and neuronal development likely associated with domestication	Pool-seq	10-20 ind. in 7 domestic populations and 14 wild populations	Illumina Genome Analyzer II, paired-end 2x76bp reads, average coverage ~10x per pool	50,165,386 SNPs	BWA[3] [default except –q 5, base quality cut-off for soft-clipping reads]	NI	SAMtools[8]	Carneiro *et al.* (2014)

Organism	Main findings	WGR method	Sample size	Sequencing design and output	Number of markers	Software for mapping [parameters]	IR/ BQSR	SNP calling software	Ref.
Chicken (*Gallus gallus domesticus*)	Genes associated to breed-related traits of pathogen resistance and reproductive ability	Pool-seq	16 ind. for each of 2 inbred lines	Illumina Hiseq 2000, 22-24x per pool	~4 million SNPs	BWA[3] [default]	NI	GATK-*UnifiedGenotyper*[7], down sampling was turned off, ploidy option was used	Fleming *et al.* (2016)
Yellow monkeyflower plant (*Mimulus guttatus*)	Candidate genes potentially driving the morphological, life history and salt tolerance differences between the two ecotypes	Pool-seq	'Coastal perennial' pool with 101 ind., 'inland annual' pool with 92 individuals	Illumina HiSeq2500, 2x250 bp paired-end reads,	29,693,578 SNPs	BWA[3] [NI]	NI	SNAPE[12]	Gould *et al.* (2017)
Atlantic salmon (*Salmon salar*)	Mapped immune related genes	Pool-seq	30 ind. in each of 19 rivers	Illumina HiSeq2000, paired-end reads, average 26.7x per pool	~4.5 million SNPs	Bowtie2[5] [without soft clipping, end-to-end mode, seed length 18 and the interval between extracted seeds set to S,1,1.5, maximum number of mismatches per seed set to L,0,0.1]	NI	SAMtools[8]	Kjærner-Semb *et al.* (2016)
Topic: Signatures of selection, genetic basis of phenotypic traits, and local adaptation									
Mosquito *Anopheles gambiae* s.l. (GOUNDRY) and *A. coluzzii*	A genomic barrier (large inversion) to gene flow between a *A. gambiae* s.l. (GOUNDRY) and *A. coluzzii*	lcWGR	11-12 ind. each	Illumina HiSeq2000, paired-end 100-bp, insert size of 500 base pairs, 9.79x to 16.44x	162 -180 million SNPs per subgroup	BWA-MEM[4] [NI]	Yes/No	ANGSD[6]	Crawford *et al.* (2016)
Native sheep (*Ovis aries*)	Loci presumably involved in adaptation to high altitude and arid environments in native sheeps	lcWGR	77 ind.	Illumina HiSeq 2000, 75 ind. with average depth of ~5x and ~42x for 2 samples	~21.26 million SNPs	BWA[3] [NI]	Yes/No	SAMtools[8] and ANGSD[6]	Yang *et al.* (2016)

Organism	Main findings	WGR method	Sample size	Sequencing design and output	Number of markers	Software for mapping [parameters]	IR/ BQSR	SNP calling software	Ref.
Shetland ponies (*Equus caballus*)	Deletions at the *SHOX* locus associated with skeletal atavism	Pool-seq	6 affected and 21 control ind.	Illumina, Hiseq2000, paired-end reads 2x100bp, ~56x	9,844,628 SNPs and 1,111,009 INDELs	BWA[3] [NI]	Yes/NI	GATK *UnifiedGenotyper*[7]	Rafati *et al.* (2016)
Bird ruff (*Philomachus pugnax*)	Large chromosomal inversion underlines the variety of male mating morphs	huWGR	15 independent and 9 satellite males from a single location	Illumina HiSeq 2000, 2x125 bp paired-end reads, average fragment size ~500 bp, average ~8x	NI	BWA[3] [default]	Yes/Yes	GATK[7]	Lamichhaney *et al.* (2015b)
Bird ruff (*Philomachus pugnax*)	Large chromosomal inversion underlines the variety of male mating morphs	huWGR	300 ind. in total	(huWGR): 80x for 5 ind. and low- (SbfI) and high- (PstI) density RAD-seq data from a pedigree population	1,068,556 SNPs	BWA-MEM[4] [default]	Yes/NI	GATK-*HaplotypeCaller*[7] (huWGR and lo-density RAD-seq data), BCFtools (for high-density RAD-sed data)	Küpper *et al.* (2015)
Plant *Arabidopsis halleri*	175 genes highly associated with some of the five environmental factors tested (precipitation, slope, solar radiation, site water balance and temperature)	Pool-seq	5 populations, 20 ind. per pool	Illumina HiSeq2000, 250–300 bp insert size, 100 bp paired-end reads, average 99x per pool	~2 million SNPs	BWA-ALN[3] and sampe [allowing 10% mismatch with the *A. thaliana* reference genome]	NI	SAMtools[8]	Fischer *et al.* (2013)
Atlantic/Baltic herring (*Clupea harengus*) and Pacific herring (*Clupea pallasii*)	Genetic differences between populations spawning in different seasons and oceanic and brackish water in Europe	Pool-seq	20 populations of *C. harengus*, 1 of *C. pallasii*, 47-100 ind. per pool, 16 ind. for WGR	Illumina Hiseq2000, paired-end 2x100bp reads, insert size ~350 bp, ~30x per pool, ~10x per individual	8.83 million SNPs (with Pacific herring), 6.04 million among Atlantic and Baltic herring	BWA-MEM[4] [NI]	No/No	GATK-*HaplotypeCaller*[7]	Martinez Barrio *et al.* (2016)

Organism	**Main findings**	**WGR method**	**Sample size**	**Sequencing design and output**	**Number of markers**	**Software for mapping [parameters]**	**IR/ BQSR**	**SNP calling software**	**Ref.**
Atlantic/Baltic herring (*Clupea harengus*) and Pacific herring (*Clupea pallasii*)	6 333 SNPs showed significant allele frequency differences between spring and fall spawning populations in Canada. About 25% of these SNPs were previously observed in Baltic Sea/NE Atlantic populations	Pool-seq	(NE): Martinez Barrio *et al.* (2016) (NW): 6 populations of *C. harengus*, 41-50 ind. per pool, 12 ind. for WGR	(NE): Martinez Barrio *et al.* (2016) (NW): Hiseq2500, paired-end 2x125bp reads, insert size ~450-550 bp, ~40x per pool, ~10x per individual	~8,9 million SNPs	BWA-MEM[4] [default]	No/No	GATK-*HaplotypeCaller*[7]	Lamichhaney *et al.* (2017)
Topic: Inbreeding, conservation breeding and restoration									
Mountain gorilla (eastern species: *Gorilla beringei beringei*, *G. beringei graueri*, western species: *G. gorilla gorilla*, *G. gorilla diehli*)	Extensive inbreeding (34% homozygosis) observed, indicating a steady population decline over the past 100 000 years	huWGR	44 ind. in total	Illumina HiSeq 2000, average 26x	1,649,453,084 SNPs	BWA-MEM[4] [default]	NI	FreeBayes[9]	Xue *et al.* (2015)
Tree *Eucalyptus grandis*	Progeny retained high and different heterozygosity percentage (52% to 79%, average 66%), in disagreement with an expectation of 50% homozygosis (IBD) produced by selfing without selection	huWGR	1 outbred parent, 28 selfed offspring	Illumina HiSeq, paired-end 2x100bp reads, average 6.733x	308,784 heterozygous SNPs	BWA[3] [default, -q 15]	NI	SAMtools[8], BAQ scores disabled.	Myburg *et al.* (2014)
Tree *Eucalyptus grandis*	Pseudo-overdominance most likely explains observed inbreeding depression, and it could be underlined by 100 or more genes of large effect associated with viability	huWGR	1 outbred parent, 28 selfed offspring	Illumina HiSeq, paired-end 2x100bp reads, average 6.733x	308,784 heterozygous SNPs	BWA[3] [default, -q 15]	NI	SAMtools[8], BAQ scores disabled.	Hedrick et al. (2016)
Rice (*Oryza sativa*)	Identification of causal mutations of three phenotypic traits in inbred rice varieties	lcWGR	203 varieties	Illumina Hiseq2000, peired-end 90bp reads, insert size 400-500 bp, average coverage 1.53x	2,288,867 SNPs	BWA[3] [default, remapping using Stampy]	Yes/NI	ANGSD[6]	Wang *et al.* (2016b)

Abbreviations: IR = INDEL recalibration, BQSR = base quality score recalibration, NI = no information, SNPs = single nucleotide polymorphisms, ind. = individuals, x = depth of coverage, huWGR = high-coverage individual whole-genome resequencing, Pool-seq = whole-genome resequencing of pooled DNA, lcWGR = low-coverage individual whole-genome resequencing.

[1] Li *et al.* (2010)
[2] Lunter & Goodson (2011)
[3] Li & Durbin (2009), Li & Durbin (2010)

[4] Li (2013)
[5] Langmead & Salzberg (2012)
[6] Korneliussen *et al.* (2014)
[7] McKenna *et al.* (2010)
[8] Li *et al.* (2009a)
[9] Garrison & Marth (2012)
[10] Koboldt *et al.* (2012)
[11] Kofler *et al.* (2011b)
[12] Raineri *et al.* (2012)
[13] Li *et al.* (2009b)

2.14 Figures

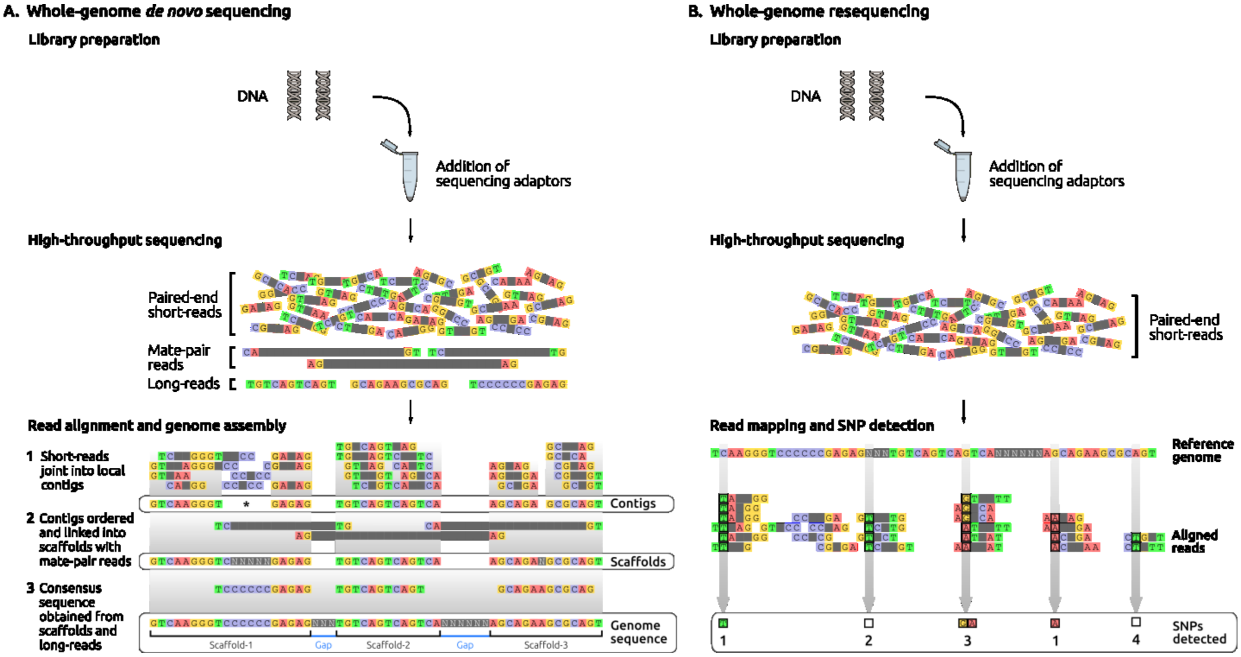

A. Whole-genome *de novo* sequencing
Library preparation
DNA
Addition of sequencing adaptors
High-throughput sequencing
Paired-end short-reads
Mate-pair reads
Long-reads
TGTCAGTCAGT
GCAGAAGCGCAG
TCCCCCCGAGAG
Read alignment and genome assembly
1 Short-reads joint into local contigs
GTCAAGGGT
*
GAGAG
TGTCAGTCAGTCA
AGCAGA
GCGCAGT
Contigs
2 Contigs ordered and linked into scaffolds with mate-pair reads
GTCAAGGGTCNNNNNGAGAG
TGTCAGTCAGTCA
AGCAGANGCGCAGT
Scaffolds
3 Consensus sequence obtained from scaffolds and long-reads
TCCCCCCGAGAG
TGTCAGTCAGT
GCAGAAGCGCAG
GTCAAGGGTCCCCCCGAGAGNNNTGTCAGTCAGTCANNNNNNAGCAGAAGCGCAGT
Genome sequence
Scaffold-1
Gap
Scaffold-2
Gap
Scaffold-3
B. Whole-genome resequencing
Library preparation
DNA
Addition of sequencing adaptors
High-throughput sequencing
Paired-end short-reads
Read mapping and SNP detection
Reference genome
Aligned reads
1
2
3
1
4
SNPs detected

Figure 2.1 Whole-genome sequencing. (**A**) *De novo* whole-genome assembly consists on sequencing and assembling a species complete genome for the first time. First, high quality genomic DNA is fragmented for library preparation that involves addition of sequencing adaptors to DNA fragments. Paired-end short-reads (~100 bp) are obtained using high-throughput sequencing from libraries with different insert sizes to maximize coverage of the genome [standard libraries: ~350-550 bp, mate-pair libraries: ~2-20 kilobases (Kb), fosmid-end libraries: ~40Kb, not shown]. Long-read sequences (~2-10Kb long) can also complement the sequence pool. (1) Read alignment starts with the building of local contigs (i.e. sequence formed by overlapping DNA fragments). (*) Repetitive regions are difficult to assemble with short-reads. (2) Mate-pair reads can help orient and link contigs, building larger sequence stretches called scaffolds (or supercontigs). Gaps in a scaffold are denoted with 'Ns'. (3) Long-reads can help in the assembly of repetitive regions. The final product is the genome assembly that results in a consensus sequence often corresponding to a series of contiguous scaffolds separated by gaps of unknown sequence (represented by runs of 'Ns'). (**B**) Whole-genome resequencing compares variable sites between the genomes of individuals or populations and requires the species genome sequence for read mapping. This image shows an example for one individual and using short-read sequencing. High quality genomic DNA of an individual is fragmented for library preparation that adds sequencing adaptors to the DNA fragments and have an insert size of ~350-500 bp. Paired-end short-reads (~100 bp) are obtained from the DNA library using high-throughput sequencing. Short-reads are mapped onto the species reference genome based on sequence similarity. A SNP is detected when the specific base observed in a position in the reference genome differs from the base observed in the

reads. Notice the uneven read coverage for some positions. (1) Variant sites correspond to a base change present only in the subject reads but not in the reference genome, (2) some SNPs may be lost because they are absent in the reference genome, (3) some SNPs may be heterozygous in the subject reads, (4) others may be lost because of low coverage. The final result is a file that contains the variable sites of the individual. In this image paired-end reads are represented by a rectangular shape with bases at both ends but not in the middle. The middle part is in grey and corresponds to unknown sequence in between paired reads. Figures 1-5 were created using the free software Inkscape (https://inkscape.org/en/).

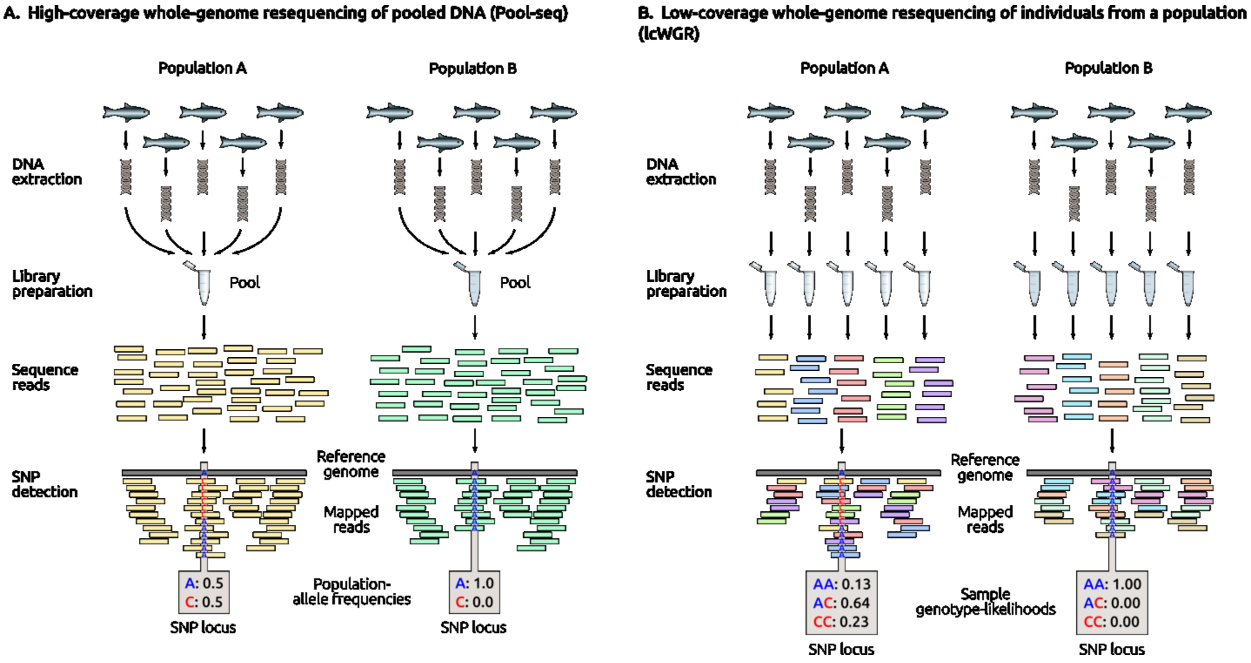

A. High-coverage whole-genome resequencing of pooled DNA (Pool-seq)
B. Low-coverage whole-genome resequencing of individuals from a population (lcWGR)
Population A
Population B
DNA extraction
Library preparation
Pool
Sequence reads
SNP detection
Reference genome
Mapped reads
Population-allele frequencies
A: 0.5
C: 0.5
A: 1.0
C: 0.0
SNP locus
Sample genotype-likelihoods
AA: 0.13
AC: 0.64
CC: 0.23
AA: 1.00
AC: 0.00
CC: 0.00

Figure 2.2 Data acquisition in current population-based WGR methods. (**A**) Pool-seq starts with mixing in a single tube equimolar amount of DNA of several individuals from a population. An aliquot of the DNA pool is used for sequencing library preparation and a single barcode is assigned to each population. Barcodes are represented by different colours in the sequence reads, yellow for population A and green for population B. The pooled-DNA library is sequenced to a high depth of coverage (>50x). SNP detection and population-allele frequency estimation require the mapping of reads to the reference genome and are based on sequence read coverage. Allele frequency differences between populations are then detected from allele read counts for a given polymorphic site. (**B**) lcWGR, starts with the preparation of a single sequencing library per individual, each with its own barcode (represented as ten different colours of short-reads). Individual DNA libraries are sequenced to a low depth of coverage (~1-4x). Read mapping to the reference genome is required for SNP detection and sample genotype likelihoods calculation, which are based on the alleles present in the individual reads supporting a variant site.

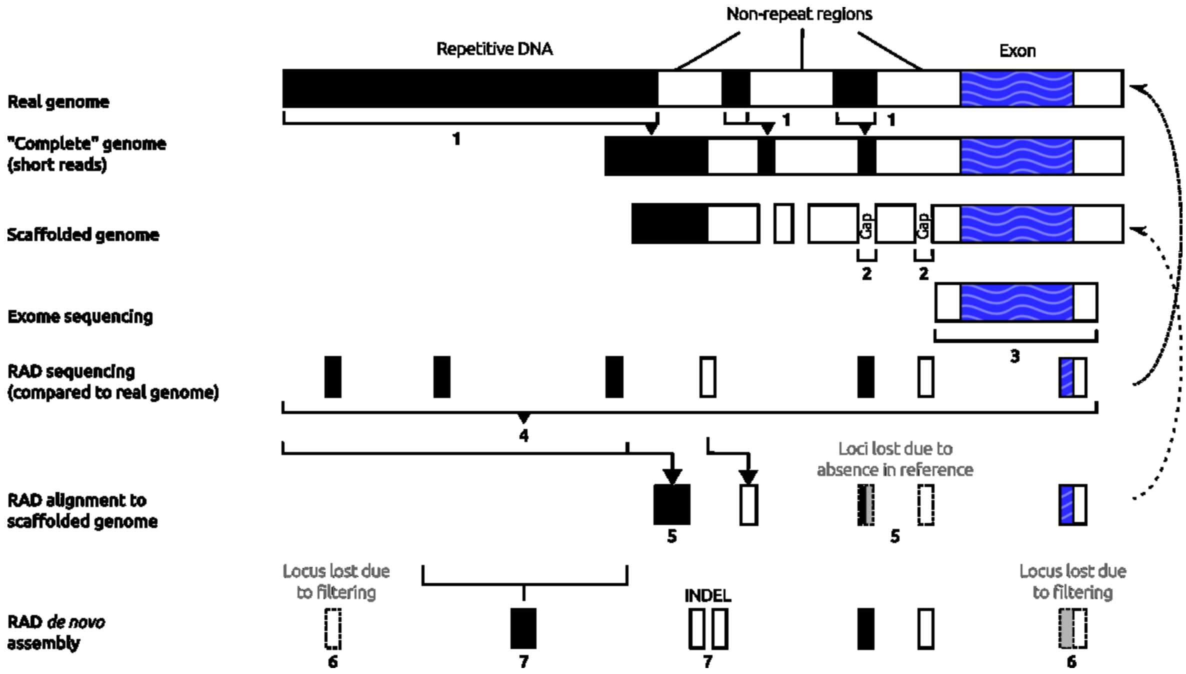
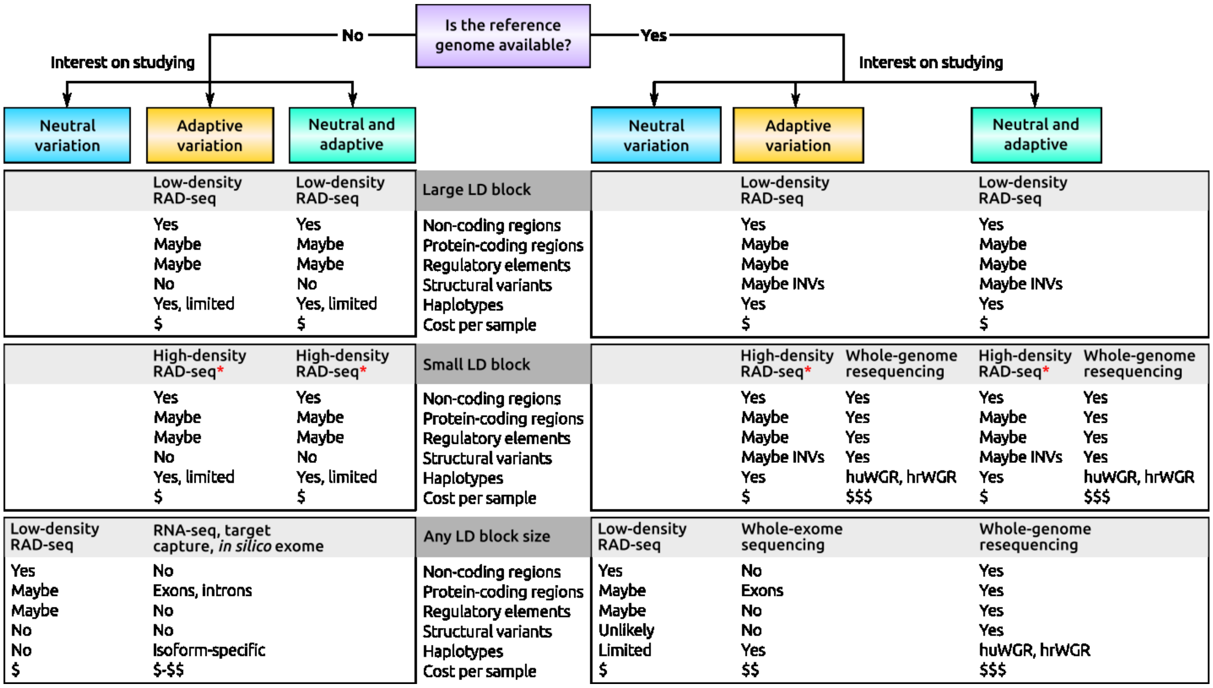


Figure 2.3 Proportion of the genome assessed by different approaches. A real genome comprises mostly repetitive and non-coding sequences with a small percentage corresponding to protein-coding regions. (1) Currently, a "complete genome" reference sequence usually misses a great proportion of repetitive regions as they are particularly challenging for base-calling and assembly algorithms based on short-read sequences; (2) A scaffold genome consists of large sequence blocks, resulting from the overlapping of short-read sequences (or contigs), that have gaps in between, corresponding to unknown parts of the sequence. A few repetitive regions are usually represented in it; (3) An exome sequence encompasses protein-coding regions (exons) and flanking sequences only, missing variation in regulatory and other noncoding regions; (4) RAD-seq randomly screens a small and dispersed amount of a real genome and includes protein-coding, non-coding, and repetitive regions; (5) When RAD-seq reads are aligned onto a scaffold genome, some RAD tags are lost because they are not present in the reference sequence. Repetitive regions would be collapsed to the fraction represented in the scaffold genome; (6) When there is no reference sequence available, RAD-seq reads are assembled to each other to form contigs (known as *de novo* assembly). Fewer unordered loci are usually recovered with this approach than when using a reference genome, mainly because of low read coverage. Also, (7) some contigs may be misaligned into different RAD loci when there are INDELs, and repetitive sequences may collapse into a single locus. Squares in gray and dashed lines indicate missing portions of the genome. The arrows point to the comparison being made and explained in the legend. Figure adapted from Hoban *et al.* (2016) with permission provided by The University of Chicago Press and Copyright Clearance Center.



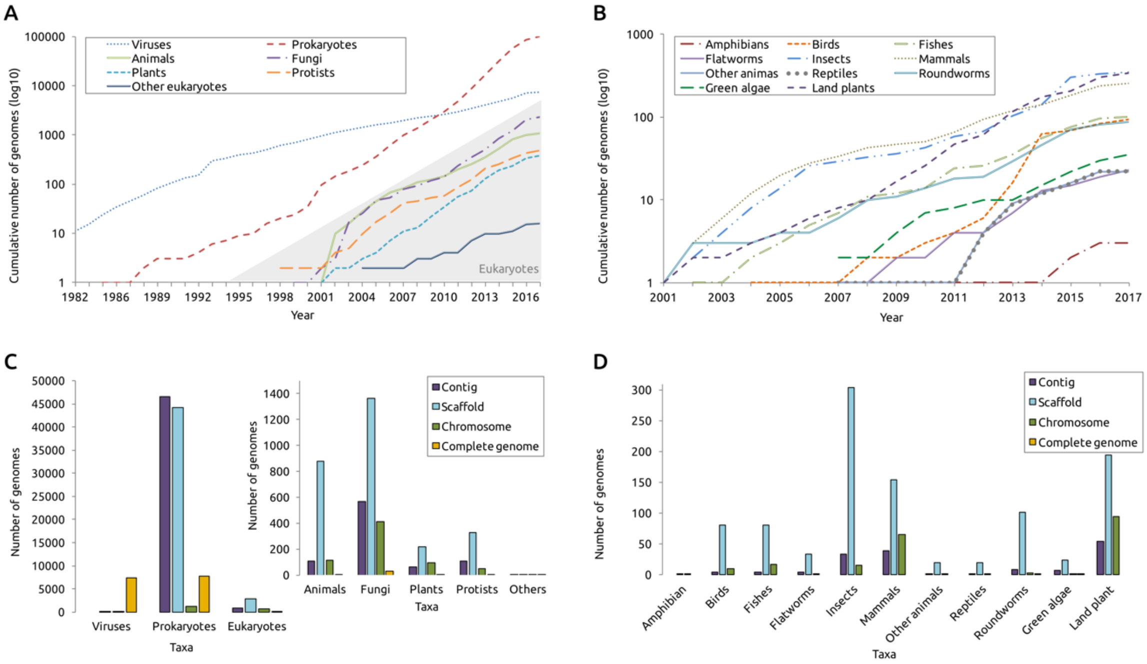
	Low-density RAD-seq	Low-density RAD-seq	Large LD block		Low-density RAD-seq	Low-density RAD-seq
	Yes	Yes	Non-coding regions		Yes	Yes
	Maybe	Maybe	Protein-coding regions		Maybe	Maybe
	Maybe	Maybe	Regulatory elements		Maybe	Maybe
	No	No	Structural variants		Maybe INVs	Maybe INVs
	Yes, limited	Yes, limited	Haplotypes		Yes	Yes
	$	$	Cost per sample		$	$

	High-density RAD-seq*	High-density RAD-seq*	Small LD block		High-density RAD-seq*	Whole-genome resequencing	High-density RAD-seq*	Whole-genome resequencing
	Yes	Yes	Non-coding regions		Yes	Yes	Yes	Yes
	Maybe	Maybe	Protein-coding regions		Maybe	Yes	Maybe	Yes
	Maybe	Maybe	Regulatory elements		Maybe	Yes	Maybe	Yes
	No	No	Structural variants		Maybe INVs	Yes	Maybe INVs	Yes
	Yes, limited	Yes, limited	Haplotypes		Yes	huWGR, hrWGR	Yes	huWGR, hrWGR
	$	$	Cost per sample		$	$$$	$	$$$

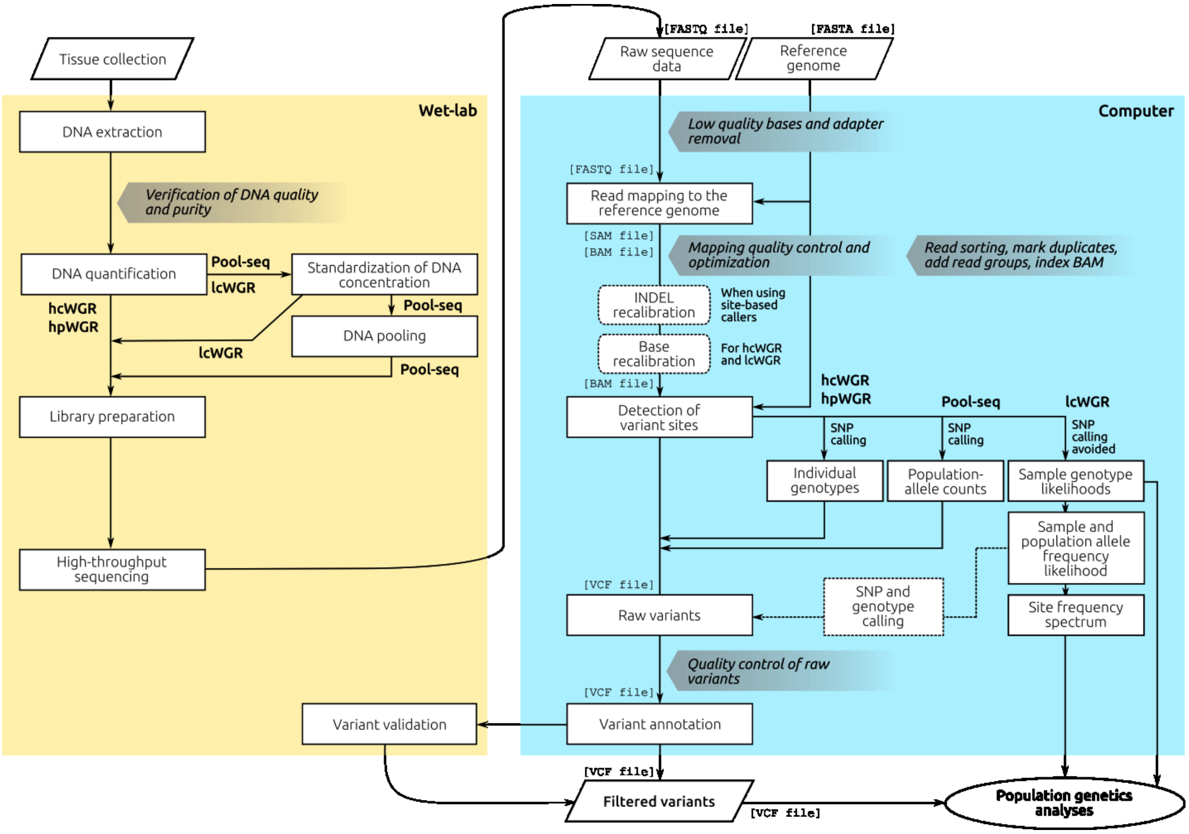
Low-density RAD-seq	RNA-seq, target capture, *in silico* exome	Any LD block size	Low-density RAD-seq	Whole-exome sequencing	Whole-genome resequencing
Yes	No	Non-coding regions	Yes	No	Yes
Maybe	Exons, introns	Protein-coding regions	Maybe	Exons	Yes
Maybe	No	Regulatory elements	Maybe	No	Yes
No	No	Structural variants	Unlikely	No	Yes
No	Isoform-specific	Haplotypes	Limited	Yes	huWGR, hrWGR
$	$-$$	Cost per sample	$	$$	$$$

Figure 2.4 Selection of sequencing approach for population genomics studies depending on reference genome availability, type of genetic variation of interest, and expected linkage disequilibrium block size. Abbreviations: RAD-seq = restriction associated DNA sequencing, INVs = inversions, huWGR = high-coverage unresolved-haplotype whole-genome-resequencing methods, hr = high-coverage resolved-haplotype whole-genome-resequencing methods, LD block = Linkage disequilibrium block size. The red asterisk indicates RAD-seq methods commonly assess a reduced fraction of the genome and experimental adjustments are required to obtain higher marker density. Relative cost per sample is represented by the number of '$' signs: '$': affordable technique, '$$': more expensive than '$', and '$$$': most expensive approach. For each block size category (large, small, or any size), the type of genetic variation surveyed is listed underneath. When the study of only neutral variation is sought, for any LD block size and in the presence or absence of a reference genome, the most cost-effective approach is low-density RAD-seq. This method screens single nucleotide polymorphisms (SNPs) mostly in non-coding regions, although some variants may fall in protein-coding regions and regulatory elements. Examination of large structural variants (SVs) (>50bp) is restricted by short-reads and low marker density. Haplotypes can be assigned although with some limitations (Arnold, Corbett-Detig, Hartl, & Bomblies, 2013). In the absence of a reference genome, when the interest is only analysing adaptive variation, there are two alternatives: RAD-seq and methods targeting protein-coding regions (i.e. RNA-seq, target capture, and *in silico* exome (Lamichhaney et al., 2012)). RAD-seq may require some fine-tuning to increase the chances of screening putatively adaptive loci (low-density markers needed for large LD block size and high-density markers for small LD block size). The methods targeting protein-coding regions only assess variation in exons and introns and may allow the reconstruction of haplotype specific isoforms. When the interest is both, neutral and adaptive variation, RAD-seq is the best approach and may

require some fine-tuning depending on LD block size. In the case a reference genome is available, when the focus is the study of adaptive variation only there are three options, first, RAD-seq as it, inferred surveys SNPs genome-wide to a fraction of the genome determined by marker density adjusted to the expected LD block size. Inversions (Küpper et al., 2015) and haplotypes (Miller et al., 2012) can be assessed with this method, the latter being inferred from population data; second, whole-exome sequencing, screens SNPs located only in exons across the entire genome, and may allow for the reconstruction of some haplotypes; and third, whole-genome resequencing that offers the greatest marker density of all current approaches, assessing SNPs across the genome including non-coding and protein-coding regions, regulatory elements, as well as structural variants. Haplotype assignment can be achieved indirectly with huWGR and directly with hrWGR methods. WGR is the most expensive approach of all. When the interest is to evaluate both, neutral and adaptive variation, the same applies as in 'adaptive variation only' and WGR is the best option for the multiple benefits it offers. Notice that the detection of adaptive variation not only depend on marker density, sampling size also plays an important role and depends on the effect size of a locus (see Box 5).



Box 1 Figure 2.1 State of the art of genomes publicly available in GenBank to date (data retrieved in June 2017). (**A**) Cumulative number of genomes per year for some major taxonomic groups. The gray shadow indicates eukaryotic groups. (**B**) Cumulative number of genomes per year for some taxonomic groups within animals and plants. (**C**) Genome completeness of some major taxonomic groups and within eukaryotes (inset). (**D**) Genome completeness of some taxonomic groups within animals and plants. Charts C and D were made based on the "assembly level" annotation associated to each genome listed in the Genome browser of Genbank (https://www.ncbi.nlm.nih.gov/genome/browse/#) and were used as proxy of genome completeness. Four levels of assembly were used (from lowest to highest): contigs, scaffolds, chromosomes, and complete genome (https://www.ncbi.nlm.nih.gov/assembly/help/#definition).



Box 2 Figure 2.1 Schematic illustration of the general workflow of the four WGR approaches (i.e. huWGR, hrWGR, Pool-seq, and lcWGR). Explanation in Box 2.

2.15 References

Adey, A., Burton, J. N., Kitzman, J. O., Hiatt, J. B., Lewis, A. P., Martin, B. K., … Shendure, J. (2013). The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature*, *500*(7461), 207–211. doi:10.1038/nature12064

Agarwala, V., Flannick, J., Sunyaev, S., & Altshuler, D. (2013). Evaluating empirical bounds on complex disease genetic architecture. *Nature Genetics*, *45*(12), 1418–27. doi:10.1038/ng.2804

Aird, D., Ross, M. G., Chen, W., Danielsson, M., Fennell, T., Russ, C., … Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, *12*(2), R18. doi:10.1186/gb-2011-12-2-r18

Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews. Genetics*, *12*(5), 363–76. doi:10.1038/nrg2958

Allendorf, F. W. (2016). Genetics and the conservation of natural populations: Allozymes to genomes. *Molecular Ecology*, *38*(1), 42–49. doi:10.1111/mec.13948

Allendorf, F. W., Hohenlohe, P. A., & Luikart, G. (2010). Genomics and the future of conservation genetics. *Nature Reviews Genetics*, *11*(10), 697–709. doi:10.1038/nrg2844

Allendorf, F. W., Luikart, G., & Aitken, S. N. (2013). *Conservation and the Genetics of Populations* (2nd ed.). Wiley-Blackwell.

Anderson, E. C., Skaug, H. J., & Barshis, D. J. (2014). Next-generation sequencing for molecular ecology: a caveat regarding pooled samples. *Molecular Ecology*, *23*(3), 502–12. doi:10.1111/mec.12609

Andersson, L. S., Larhammar, M., Memic, F., Wootz, H., Schwochow, D., Rubin, C.-J., … Kullander, K. (2012). Mutations in DMRT3 affect locomotion in horses and spinal circuit function in mice. *Nature*, *488*(7413), 642–646. doi:10.1038/nature11399

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, *17*(2), 81–92. doi:10.1038/nrg.2015.28

Angeloni, F., Wagemaker, N., Vergeer, P., & Ouborg, J. (2012). Genomic toolboxes for conservation biologists. *Evolutionary Applications*, *5*(2), 130–143. doi:10.1111/j.1752-4571.2011.00217.x

Arnold, B., Corbett-Detig, R. B., Hartl, D., & Bomblies, K. (2013). RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*, *22*(11), 3179–3190. doi:10.1111/mec.12276

Auer, P. L., & Lettre, G. (2015). Rare variant association studies: considerations, challenges and opportunities. *Genome Medicine*, *7*(1), 16. doi:10.1186/s13073-015-0138-2

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. doi:10.1038/nature15393

Avise, J. C. (2010). Perspective: Conservation genetics enters the genomics era. *Conservation Genetics*, *11*(2), 665–669. doi:10.1007/s10592-009-0006-y

Aykanat, T., Lindqvist, M., Pritchard, V. L., & Primmer, C. R. (2016). From population genomics to conservation and management: a workflow for targeted analysis of markers identified using genome-wide approaches in Atlantic salmon Salmo salar. *Journal of Fish Biology*, *89*(6), 2658–2679. doi:10.1111/jfb.13149

Ayllon, F., Kjærner-Semb, E., Furmanek, T., Wennevik, V., Solberg, M. F., Dahle, G., … Wargelius, A. (2015). The vgll3 Locus Controls Age at Maturity in Wild and Domesticated Atlantic Salmon (Salmo salar L.) Males. *PLoS Genetics*, *11*(11), 1–15. doi:10.1371/journal.pgen.1005628

Baird, N., Etter, P., Atwood, T., Currey, M., Shiver, A., Lewis, Z., … Johnson, E. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS One*, *3*(10), e3376. doi:10.1371/journal.pone.0003376

Baker, M. (2012). De novo genome assembly: what every biologist should know. *Nature Methods*, *9*(4), 333–337. doi:10.1038/nmeth.1935

Barrett, R. D. H., & Hoekstra, H. E. (2011). Molecular spandrels: tests of adaptation at the genetic level. *Nature Review Genetics*, *12*(11), 767–780. doi:10.1038/nrg3015

Barson, N. J., Aykanat, T., Hindar, K., Baranski, M., Bolstad, G. H., Fiske, P., … Primmer, C. R. (2015). Sex-dependent dominance at a single locus maintains variation in age at maturity in salmon. *Nature*, *528*(7582), 405–408. doi:10.1038/nature16062

Baym, M., Kryazhimskiy, S., Lieberman, T. D., Chung, H., Desai, M. M., & Kishony, R. K. (2015). Inexpensive multiplexed library preparation for megabase-sized genomes. *PLoS ONE*, *10*(5), 1–15. doi:10.1371/journal.pone.0128036

Beissinger, T. M., Rosa, G. J., Kaeppler, S. M., Gianola, D., & de Leon, N. (2015). Defining window-boundaries for genomic analyses using smoothing spline techniques. *Genetics Selection Evolution*, *47*(1), 30. doi:10.1186/s12711-015-0105-9

Bekkevold, D., Helyar, S. J., Limborg, M. T., Nielsen, E. E., Hemmer-Hansen, J., Clausen, L. A. W., & Carvalho, G. R. (2015). Gene-associated markers can assign origin in a weakly structured fish, Atlantic herring. *ICES Journal of Marine Science: Journal Du Conseil*, *72*(6), 1790–1801. doi:10.1093/icesjms/fsu247

Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q. B., Antipenko, A., … Abel, L. (2015). Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proceedings of the National Academy of Sciences*, *112*(17), 5473–5478. doi:10.1073/pnas.1418631112

Benestan, L. M., Ferchaud, A.-L., Hohenlohe, P. A., Garner, B. A., Naylor, G. J. P., Baums, I. B., … Luikart, G. (2016). Conservation genomics of natural and managed populations: building a conceptual and practical framework. *Molecular Ecology*, *25*(13), 2967–2977. doi:10.1111/mec.13647

Berg, J. J., & Coop, G. (2014). A Population Genetic Signal of Polygenic Adaptation. *PLoS Genetics*, *10*(8), e1004412. doi:10.1371/journal.pgen.1004412

Berlin, K., Koren, S., Chin, C.-S., Drake, J. P., Landolin, J. M., & Phillippy, A. M. (2015). Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature Biotechnology*, *33*(6), 623–630. doi:10.1038/nbt.3238

Bernatchez, L. (2016). On the maintenance of genetic variation and adaptation to environmental change: considerations from population genomics in fishes. *Journal of Fish Biology*, 1–38. doi:10.1111/jfb.13145

Bi, K., Vanderpool, D., Singhal, S., Linderoth, T., Moritz, C., & Good, J. M. (2012). Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics*, *13*(1), 403. doi:10.1186/1471-2164-13-403

Bickhart, D. M., Rosen, B. D., Koren, S., Sayre, B. L., Hastie, A. R., Chan, S., … Smith, T. P. L. (2017). Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nature Genetics*, *53*(9), 1689–1699. doi:10.1038/ng.3802

Bleidorn, C. (2016). Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. *Systematics and Biodiversity*, *14*(1), 1–8. doi:10.1080/14772000.2015.1099575

Blischak, P. D., Wenzel, A. J., & Wolfe, A. D. (2014). Gene Prediction and Annotation in Penstemon (Plantaginaceae): A Workflow for Marker Development from Extremely Low-Coverage Genome Sequencing. *Applications in Plant Sciences*, *2*(12), 1400044. doi:10.3732/apps.1400044

Boitard, S., Rodríguez, W., Jay, F., Mona, S., & Austerlitz, F. (2016). Inferring Population Size History from Large Samples of Genome-Wide Molecular Data - An Approximate Bayesian Computation Approach. *PLOS Genetics*, *12*(3), e1005877. doi:10.1371/journal.pgen.1005877

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. doi:10.1093/bioinformatics/btu170

Bossdorf, O., Richards, C. L., & Pigliucci, M. (2008). Epigenetics for ecologists. *Ecology Letters*, *11*(2), 106–115. doi:10.1111/j.1461-0248.2007.01130.x

Bradbury, I. R., Hamilton, L. C., Dempson, B., Robertson, M. J., Bourret, V., Bernatchez, L., & Verspoor, E. (2015). Transatlantic secondary contact in Atlantic Salmon, comparing microsatellites, a single nucleotide polymorphism array and restriction-site associated DNA sequencing for the resolution of complex spatial structure. *Molecular Ecology*, *24*(20), 5130–5144. doi:10.1111/mec.13395

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., … Korf, I. F. (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, *2*(1), 10. doi:10.1186/2047-217X-2-10

Browning, S. R., & Browning, B. L. (2011). Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, *12*(10), 703–714. doi:10.1038/nrg3054

Buerkle, A. C., & Gompert, Z. (2013). Population genomics based on low coverage sequencing: how low should we go? *Molecular Ecology*, *22*(11), 3028–3035. doi:10.1111/mec.12105

Bustamante, C. D., Wakeley, J., Sawyer, S., & Hartl, D. L. (2001). Directional selection and the site-frequency spectrum. *Genetics*, *159*(4), 1779–1788.

Cariou, M., Duret, L., & Charlat, S. (2016). How and how much does RAD-seq bias genetic diversity estimates? *BMC Evolutionary Biology*, *16*(1), 240. doi:10.1186/s12862-016-0791-0

Carneiro, M., Rubin, C.-J., Di Palma, F., Albert, F. W., Alfoldi, J., Barrio, A. M., … Andersson, L. (2014). Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. *Science*, *345*(6200), 1074–1079. doi:10.1126/science.1253714

Catchen, J. M., Hohenlohe, P. A., Bernatchez, L., Funk, W. C., Andrews, K. R., & Allendorf, F. W. (2017). Unbroken: RADseq remains a powerful tool for understanding the genetics of adaptation in natural populations. *Molecular Ecology Resources*, *38*(1), 42–49. doi:10.1111/1755-0998.12669

Chakraborty, M., Baldwin-Brown, J. G., Long, A. D., & Emerson, J. J. (2016). Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Research*, *44*(19), 1–12. doi:10.1093/nar/gkw654

Chan, C. X., & Ragan, M. A. (2013). Next-generation phylogenomics. *Biology Direct*, *8*(1), 3. doi:10.1186/1745-6150-8-3

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., … Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, *6*(2), 80–92. doi:10.4161/fly.19695

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2009). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, *38*(6), 1767–1771. doi:10.1093/nar/gkp1137

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., … Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*(1), 13. doi:10.1186/s13059-016-0881-8

Crawford, J. E., Riehle, M. M., Markianos, K., Bischoff, E., Guelbeogo, W. M., Gneme, A., … Lazzaro, B. P. (2016). Evolution of GOUNDRY, a cryptic subgroup of Anopheles gambiae s.l., and its impact on susceptibility to Plasmodium infection. *Molecular Ecology*, *25*(7), 1494–1510. doi:10.1111/mec.13572

Cutler, D. J., & Jensen, J. D. (2010). To pool, or not to pool? *Genetics*, *186*(1), 41–3. doi:10.1534/genetics.110.121012

da Fonseca, R. R., Albrechtsen, A., Themudo, G. E., Ramos-Madrigal, J., Sibbesen, J. A., Maretty, L., … Pereira, R. J. (2016). Next-generation biology: Sequencing and data analysis approaches for non-model organisms. *Marine Genomics*, *30*, 1–11. doi:10.1016/j.margen.2016.04.012

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. doi:10.1093/bioinformatics/btr330

De Wit, P., Pespeni, M. H., Ladner, J. T., Barshis, D. J., & Palumbi, S. R. (2012). The simple fool' s guide to population genomics via RNA-Seq : an introduction to high-throughput sequencing data analysis. *Molecular Ecology Resources*, *12*, 1058–1067. doi:10.1111/1755-0998.12003

Delsuc, F., Brinkmann, H., & Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nature Reviews. Genetics*, *6*(5), 361–375. doi:10.1038/nrg1603

Dennenmoser, S., Vamosi, S. M., Nolte, A. W., & Rogers, S. M. (2017). Adaptive genomic divergence under high gene flow between freshwater and brackish-water ecotypes of prickly sculpin (Cottus asper) revealed by Pool-Seq. *Molecular Ecology*, *26*(1), 25–42. doi:10.1111/mec.13805

DePristo, M. a, Banks, E., Poplin, R., Garimella, K. V, Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–8. doi:10.1038/ng.806

Durbin, R. M., Altshuler, D. L., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., … Africa, W. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–73. doi:10.1038/nature09534

Durrett, R. (2008). *Probability Models for DNA Sequence Evolution* (2nd ed.). New York, NY: Springer New York. doi:10.1007/978-0-387-78168-6

Earl, D., Bradnam, K., St. John, J., Darling, A., Lin, D., Fass, J., … Paten, B. (2011). Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Research*, *21*(12), 2224–2241. doi:10.1101/gr.126599.111

Ekblom, R., & Wolf, J. B. W. (2014). A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications*, *7*(9), 1026–1042. doi:10.1111/eva.12178

Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution*, *29*(1), 51–63. doi:10.1016/j.tree.2013.09.008

Evans, J. D., Brown, S. J., Hackett, K. J. J., Robinson, G., Richards, S., Lawson, D., … Zhou, X. (2013). The i5K initiative: Advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *Journal of Heredity*, *104*(5), 595–600. doi:10.1093/jhered/est050

Ferretti, L., Ramos-Onsins, S. E., & Pérez-Enciso, M. (2013). Population genomics from pool sequencing. *Molecular Ecology*, *22*(22), 5561–5576. doi:10.1111/mec.12522

Field, Y., Boyle, E. A., Telis, N., Gao, Z., Gaulton, K. J., Golan, D., … Pritchard, J. K.

(2016). Detection of human adaptation during the past 2000 years. *Science*, *354*(6313), 760–764. doi:10.1126/science.aag0776

Fischer, M. C., Rellstab, C., Leuzinger, M., Roumet, M., Gugerli, F., Shimizu, K. K., … Widmer, A. (2017). Estimating genomic diversity and population differentiation – an empirical comparison of microsatellite and SNP variation in Arabidopsis halleri. *BMC Genomics*, *18*(1), 69. doi:10.1186/s12864-016-3459-7

Fischer, M. C., Rellstab, C., Tedder, A., Zoller, S., Gugerli, F., Shimizu, K. K., … Widmer, A. (2013). Population genomic footprints of selection and associations with climate in natural populations of Arabidopsis halleri from the Alps. *Molecular Ecology*, *22*(22), 5594–5607. doi:10.1111/mec.12521

Fleming, D. S., Koltes, J. E., Fritz-Waters, E. R., Rothschild, M. F., Schmidt, C. J., Ashwell, C. M., … Lamont, S. J. (2016). Single nucleotide variant discovery of highly inbred Leghorn and Fayoumi chicken breeds using pooled whole genome resequencing data reveals insights into phenotype differences. *BMC Genomics*, *17*(1), 812. doi:10.1186/s12864-016-3147-7

Fonseca, N. A., Rung, J., Brazma, A., & Marioni, J. C. (2012). Tools for mapping high-throughput sequencing data. *Bioinformatics*, *28*(24), 3169–3177. doi:10.1093/bioinformatics/bts605

Fontanesi, L., Di Palma, F., Flicek, P., Smith, A. T., Thulin, C.-G., & Alves, P. C. (2016). LaGomiCs—Lagomorph Genomics Consortium: An International Collaborative Effort for Sequencing the Genomes of an Entire Mammalian Order. *Journal of Heredity*, *107*(4), 295–308. doi:10.1093/jhered/esw010

Foote, A. D., Vijay, N., Ávila-Arcos, M. C., Baird, R. W., Durban, J. W., Fumagalli, M., … Wolf, J. B. W. (2016). Genome-culture coevolution promotes rapid divergence of killer whale ecotypes. *Nature Communications*, *7*(May), 11693. doi:10.1038/ncomms11693

Fracassetti, M., Griffin, P. C., & Willi, Y. (2015). Validation of pooled whole-genome re-sequencing in Arabidopsis lyrata. *PLoS ONE*, *10*(10), 1–15. doi:10.1371/journal.pone.0140462

Frankham, R. (2010). Challenges and opportunities of genetic approaches to biological conservation. *Biological Conservation*, *143*(9), 1919–1927. doi:10.1016/j.biocon.2010.05.011

Fumagalli, M. (2013). Assessing the effect of sequencing depth and sample size in population genetics inferences. *PloS One*, *8*(11), e79667. doi:10.1371/journal.pone.0079667

Fumagalli, M., Vieira, F. G., Korneliussen, T. S., Linderoth, T., Huerta-Sánchez, E., Albrechtsen, A., & Nielsen, R. (2013). Quantifying population genetic differentiation from next-generation sequencing data. *Genetics*, *195*(3), 979–92. doi:10.1534/genetics.113.154740

Fumagalli, M., Vieira, F. G., Linderoth, T., & Nielsen, R. (2014). NgsTools: Methods for population genetics analyses from next-generation sequencing data. *Bioinformatics*, *30*(10), 1486–1487. doi:10.1093/bioinformatics/btu041

Funk, W. C., McKay, J. K., Hohenlohe, P. a, & Allendorf, F. W. (2012). Harnessing genomics for delineating conservation units. *Trends in Ecology & Evolution*, *27*(9), 489–496. doi:10.1016/j.tree.2012.05.012

Fussi, B., Westergren, M., Aravanopoulos, F., Baier, R., Kavaliauskas, D., Finzgar, D., … Kraigher, H. (2016). Forest genetic monitoring: an overview of concepts and definitions. *Environmental Monitoring and Assessment*, *188*(8), 493. doi:10.1007/s10661-016-5489-7

Futschik, A., & Schlötterer, C. (2010). The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics*, *186*(1), 207–18. doi:10.1534/genetics.110.114397

Gagnaire, P.-A., & Gaggiotti, O. E. (2016). Detecting polygenic selection in marine populations by combining population genomics and quantitative genetics approaches. *Current Zoology*, *62*(August), 1–14. doi:10.1093/cz/zow088

Garner, B. A., Hand, B. K., Amish, S. J., Bernatchez, L., Foster, J. T., Miller, K. M., … Luikart, G. (2016). Genomics in Conservation: Case Studies and Bridging the Gap between Data and Application. *Trends in Ecology & Evolution*, *31*(2), 81–83. doi:10.1016/j.tree.2015.10.009

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *ArXiv Preprint ArXiv:1207.3907*, 9. doi:arXiv:1207.3907

Gautier, M., Foucaud, J., Gharbi, K., Cézard, T., Galan, M., Loiseau, A., … Estoup, A. (2013). Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. *Molecular Ecology*, *22*(14), 3766–79. doi:10.1111/mec.12360

GIGA. (2014). The Global Invertebrate Genomics Alliance (GIGA): Developing Community Resources to Study Diverse Invertebrate Genomes. *Journal of Heredity*, *105*(1), 1–18. doi:10.1093/jhered/est084

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333–351. doi:10.1038/nrg.2016.49

Gould, B. A., Chen, Y., & Lowry, D. B. (2017). Pooled ecotype sequencing reveals candidate genetic mechanisms for adaptive differentiation and reproductive isolation. *Molecular Ecology*, *26*(1), 163–177. doi:10.1111/mec.13881

Grigoriev, I. V., Nikitin, R., Haridas, S., Kuo, A., Ohm, R., Otillar, R., … Shabalov, I. (2014). MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Research*, *42*(D1), D699–D704. doi:10.1093/nar/gkt1183

Grossen, C., Biebach, I., Angelone-Alasaad, S., Keller, L. F., & Croll, D. (2017). Population genomics analyses of European ibex species show lower diversity and higher inbreeding in reintroduced populations. *Evolutionary Applications*. doi:10.1111/eva.12490

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

doi:10.1093/bioinformatics/btt086

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genetics*, *5*(10), e1000695. doi:10.1371/journal.pgen.1000695

Haasl, R. J., & Payseur, B. A. (2016). Fifteen years of genomewide scans for selection: Trends, lessons and unaddressed genetic sources of complication. *Molecular Ecology*, *25*(1), 5–23. doi:10.1111/mec.13339

Habicht, C., Munro, A., Dann, T., Eggers, D., Templin, W., Witteveen, M., … Volk, E. (2012). *Harvest and Harvest Rates of Sockeye Salmon Stocks in Fisheries of the Western Alaska Salmon Stock Identification Program (WASSIP), 2006– 2008*. Alaska, US.

Haig, S. M., Miller, M. P., Bellinger, R., Draheim, H. M., Mercer, D. M., & Mullins, T. D. (2016). The conservation genetics juggling act: integrating genetics and ecology, science and policy. *Evolutionary Applications*, *9*(1), 181–195. doi:10.1111/eva.12337

Han, E., Sinsheimer, J. S., & Novembre, J. (2015). Fast and accurate site frequency spectrum estimation from low coverage sequence data. *Bioinformatics*, *31*(5), 720–727. doi:10.1093/bioinformatics/btu725

Hansen, T. F. (2006). The Evolution of Genetic Architecture. *Annual Review of Ecology, Evolution, and Systematics*, *37*(1), 123–157. doi:10.1146/annurev.ecolsys.37.091305.110224

Hatem, A., Bozdağ, D., & Çatalyürek, Ü. V. (2013). Benchmarking short sequence mapping tools. *BMC Bioinformatics*, *14*(184), 1–25. Retrieved from http://www.biomedcentral.com/1471-2105/14/184%0ARESEARCH

Head, S. R., Komori, H. K., LaMere, S. A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D. R., & Ordoukhanian, P. (2014). Library construction for next-generation sequencing: Overviews and challenges. *BioTechniques*, *56*(2), 167–203. doi:10.2144/000114133

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), 1–8. doi:10.1016/j.ygeno.2015.11.003

Hedrick, P. W., Hellsten, U., & Grattapaglia, D. (2016). Examining the cause of high inbreeding depression: Analysis of whole-genome sequence data in 28 selfed progeny of Eucalyptus grandis. *New Phytologist*, *209*(2), 600–611. doi:10.1111/nph.13639

Hedrick, P. W., & Miller, P. S. (1992). Conservation Genetics: Techniques and Fundamentals. *Ecological Applications*, *2*(1), 30–46.

Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., … Whitlock, M. C. (2016). Finding the Genomic Basis of Local Adaptation: Pitfalls, Practical Solutions, and Future Directions. *The American Naturalist*, *188*(4), 379–397. doi:10.1086/688018

Hoffmann, A. A., & Rieseberg, L. H. (2008). Revisiting the Impact of Inversions in Evolution: From Population Genetic Markers to Drivers of Adaptive Shifts and Speciation? *Annual Review of Ecology, Evolution, and Systematics*, *39*(2008), 21–42. doi:10.1146/annurev.ecolsys.39.110707.173532

Hohenlohe, P. a, Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. a, & Cresko, W. a. (2010). Population Genomics of Parallel Adaptation in Threespine Stickleback using Sequenced RAD Tags. *PLoS Genetics*, *6*(2), e1000862. doi:10.1371/journal.pgen.1000862

Holsinger, K. E., & Weir, B. S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nature Reviews. Genetics*, *10*(September), 639–650. doi:10.1038/nrg2611

Human Genome Sequencing Consortium, I. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*(7011), 931–945. doi:10.1038/nature03001

Ivy, J. A., Putnam, A. S., Navarro, A. Y., Gurr, J., & Ryder, O. A. (2016). Applying SNP-derived molecular coancestry estimates to captive breeding programs. *Journal of Heredity*, *107*(5), 403–412. doi:10.1093/jhered/esw029

Jarvis, E., Zhang, G., Li, C., Li, Q., Li, B., Larkin, D. M., … Froman, D. P. (2014). Comparative genomics reveals insights into avian genome evolution and adaptation. *Science*, *346*(6215), 1311–1320. doi:10.1126/science.1251385

Jensen, J. D., Foll, M., & Bernatchez, L. (2016). The past, present and future of genomic scans for selection. *Molecular Ecology*, *25*(1), 1–4. doi:10.1111/mec.13493

Jonas, A., Taus, T., Kosiol, C., Schlotterer, C., & Futschik, A. (2016). Estimating the Effective Population Size from Temporal Allele Frequency Changes in Experimental Evolution. *Genetics*, *204*(2), 723–735. doi:10.1534/genetics.116.191197

Jones, M. R., & Good, J. M. (2016). Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, *25*(1), 185–202. doi:10.1111/mec.13304

Joron, M., Frezal, L., Jones, R. T., Chamberlain, N. L., Lee, S. F., Haag, C. R., … ffrench-Constant, R. H. (2011). Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature*, *477*(7363), 203–206. doi:10.1038/nature10341

Kaiser, T. S., Poehn, B., Szkiba, D., Preussner, M., Sedlazeck, F. J., Zrim, A., … Tessmar-Raible, K. (2016). The genomic basis of circadian and circalunar timing adaptations in a midge. *Nature*, *540*(7631), 69–73. doi:10.1038/nature20151

Kardos, M., Taylor, H. R., Ellegren, H., Luikart, G., & Allendorf, F. W. (2016). Genomics advances the study of inbreeding depression in the wild. *Evolutionary Applications*, *9*(10), 1205–1218. doi:10.1111/eva.12414

Kim, S., Lohmueller, K., Albrechtsen, A., Li, Y., Korneliussen, T., Tian, G., … Nielsen, R. (2011). Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC Bioinformatics*, *12*(1), 231. doi:10.1186/1471-2105-12-231

Kjærner-Semb, E., Ayllon, F., Furmanek, T., Wennevik, V., Dahle, G., Niemelä, E., … Edvardsen, R. B. (2016). Atlantic salmon populations reveal adaptive divergence of immune related genes - a duplicated genome under selection. *BMC Genomics*, *17*(1), 610. doi:10.1186/s12864-016-2867-z

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., … Wilson, R. K. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, *22*(3), 568–576. doi:10.1101/gr.129684.111

Koepfli, K., Paten, B., Genome 10K Community of Scientists, & O'Brien, S. J. (2015). The Genome 10K Project: a way forward. *Annual Review of Animal Biosciences*, *3*, 57–111. doi:10.1146/annurev-animal-090414-014900

Kofler, R., Langmuller, A. M., Nouhaud, P., Otte, K. A., & Schlotterer, C. (2016). Suitability of Different Mapping Algorithms for Genome-wide Polymorphism Scans with Pool-Seq Data. *Genes|Genomes|Genetics*, *6*(November), 1–20. doi:10.1534/g3.116.034488

Kofler, R., Orozco-terWengel, P., De Maio, N., Pandey, R. V., Nolte, V., Futschik, A., … Schlötterer, C. (2011). PoPoolation: A Toolbox for Population Genetic Analysis of Next Generation Sequencing Data from Pooled Individuals. *PLoS ONE*, *6*(1), e15925. doi:10.1371/journal.pone.0015925

Kofler, R., Pandey, R. V., & Schlötterer, C. (2011). PoPoolation2: Identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics*, *27*(24), 3435–3436. doi:10.1093/bioinformatics/btr589

Korneliussen, T. S. T., Albrechtsen, A., Nielsen, R., Nielsen, R., Paul, J., Albrechtsen, A., … Ballinger, Dge. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, *15*(356), 1–13. doi:10.1186/s12859-014-0356-4

Küpper, C., Stocks, M., Risse, J. E., Remedios, N., Farrell, L. L., Mcrae, B., … Burke, T. (2015). A supergene determines highly divergent male reproductive morphs in the ruff. *Nature Publishing Group*, *48*(1), 79–83. doi:10.1038/ng.3443

Laehnemann, D., Borkhardt, A., & McHardy, A. C. (2016). Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction. *Briefings in Bioinformatics*, *17*(1), 154–179. doi:10.1093/bib/bbv029

Lamichhaney, S., Barrio, A. M., Rafati, N., Sundstrom, G., Rubin, C.-J., Gilbert, E. R., … Andersson, L. (2012). Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. *Proceedings of the National Academy of Sciences*, *109*(47), 19345–19350. doi:10.1073/pnas.1216128109

Lamichhaney, S., Berglund, J., Almén, M. S., Maqbool, K., Grabherr, M., Martinez-Barrio, A., … Andersson, L. (2015). Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature*, *518*(7539), 371–375. doi:10.1038/nature14181

Lamichhaney, S., Fan, G., Widemo, F., Gunnarsson, U., Thalmann, D. S., Hoeppner, M. P., … Andersson, L. (2015). Structural genomic changes underlie alternative reproductive strategies in the ruff (Philomachus pugnax). *Nature Genetics*, *48*(1), 84–88. doi:10.1038/ng.3430

Lamichhaney, S., Fuentes-Pardo, A. P., Rafati, N., Ryman, N., McCracken, G. R., Bourne, C., … Andersson, L. (2017). Parallel adaptive evolution of geographically distant herring populations on both sides of the North Atlantic Ocean. *Proceedings of the National Academy of Sciences*, *114*(17), E3452–E3461. doi:10.1073/pnas.1617728114

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods*, *9*(4), 357–359. doi:10.1038/nmeth.1923

Larsson, L. C., Laikre, L., André, C., Dahlgren, T. G., & Ryman, N. (2010). Temporally stable genetic structure of heavily exploited Atlantic herring (Clupea harengus) in Swedish waters. *Heredity*, *104*(1), 40–51. doi:10.1038/hdy.2009.98

Lee, H., Gurtowski, J., Yoo, S., Nattestad, M., Marcus, S., Goodwin, S., … Schatz, M. (2016). Third-generation sequencing and the future of genomics. *BioRxiv*, (Table 2.1), 048603. doi:doi.org/10.1101/048603

Lee, S., Abecasis, G. R., Boehnke, M., & Lin, X. (2014). Rare-variant association analysis: Study designs and statistical tests. *American Journal of Human Genetics*, *95*(1), 5–23. doi:10.1016/j.ajhg.2014.06.009

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv Preprint ArXiv*, *00*(00), 3. doi:arXiv:1303.3997 [q-bio.GN]

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. doi:10.1093/bioinformatics/btp324

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, *26*(5), 589–595. doi:10.1093/bioinformatics/btp698

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. doi:10.1093/bioinformatics/btp352

Li, H., Ruan, J., & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, *18*(11), 1851–1858. doi:10.1101/gr.078212.108

Li, H., & Wren, J. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, *30*(20), 2843–2851. doi:10.1093/bioinformatics/btu356

Li, R., Li, Y., Fang, X., Yang, H., Wang, J., Kristiansen, K., & Wang, J. (2009). SNP detection for massively parallel whole-genome resequencing. *Genome Research*, *19*(6), 1124–1132. doi:10.1101/gr.088013.108

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., … Wang, J. (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, *20*(2), 265–272. doi:10.1101/gr.097261.109

Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Matthew, P., Leong, J. S., … Vik, J. O. (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature*, *533*(6020), 200–205. doi:10.1038/nature17164

Limborg, M. T., Helyar, S., De Bruyn, M., Taylor, M. I., Nielsen, E. E., Ogden, R., … Bekkevold, D. (2012). Environmental selection on transcriptome-derived SNPs in a high gene flow marine fish, the Atlantic herring (Clupea harengus). *Molecular Ecology*, *21*(15), 3686–3703. doi:10.1111/j.1365-294X.2012.05639.x

Lopes, R. J., Johnson, J. D., Toomey, M. B., Ferreira, M. S., Araujo, P. M., Melo-Ferreira, J., … Carneiro, M. (2016). Genetic Basis for Red Coloration in Birds. *Current Biology*, *26*(11), 1427–1434. doi:10.1016/j.cub.2016.03.076

Lotterhos, K. E., & Whitlock, M. C. (2015). The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology*, *24*(5), 1031–1046. doi:10.1111/mec.13100

Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., & Storfer, A. (2017a). Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Molecular Ecology Resources*, *17*(2), 142–152. doi:10.1111/1755-0998.12635

Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., & Storfer, A. (2017b). Responsible RAD: Striving for best practices in population genomic studies of adaptation. *Molecular Ecology Resources*, *38*(1), 42–49. doi:10.1111/1755-0998.12677

Lunter, G., & Goodson, M. (2011). Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, *21*(6), 936–939. doi:10.1101/gr.111120.110

Mace, G. M. (2004). The role of taxonomy in species conservation. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *359*(1444), 711–9. doi:10.1098/rstb.2003.1454

Malmstrøm, M., Matschiner, M., Tørresen, O. K., Jakobsen, K. S., & Jentoft, S. (2017). Whole genome sequencing data and de novo draft assemblies for 66 teleost species. *Scientific Data*, *4*, 160132. doi:10.1038/sdata.2016.132

Manthey, J. D., Campillo, L. C., Burns, K. J., & Moyle, R. G. (2016). Comparison of Target-Capture and Restriction-Site Associated DNA Sequencing for Phylogenomics: A Test in Cardinalid Tanagers (Aves, Genus: Piranga). *Systematic Biology*, *65*(4), 640–650. doi:10.1093/sysbio/syw005

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10–12. doi:http://dx.doi.org/10.14806/ej.17.1.200

Martin, S. H., & Jiggins, C. D. (2013). Genomic Studies of Adaptation in Natural Populations. In *eLS*. Chichester, UK: John Wiley & Sons, Ltd. doi:10.1002/9780470015902.a0024613

Martinez Barrio, A., Lamichhaney, S., Fan, G., Rafati, N., Pettersson, M., Zhang, H., …

Andersson, L. (2016). The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. *ELife*, *5*(MAY2016), 1–32. doi:10.7554/eLife.12081

Martinsohn, J. T., & Ogden, R. (2009). FishPopTrace—Developing SNP-based population genetic assignment methods to investigate illegal fishing. *Forensic Science International: Genetics Supplement Series*, *2*(1), 294–296. doi:10.1016/j.fsigss.2009.08.108

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. doi:10.1101/gr.107524.110

McKinney, G. J., Larson, W. A., Seeb, L. W., & Seeb, J. E. (2017). RADseq provides unprecedented insights into molecular ecology and evolutionary genetics: comment on Breaking RAD by Lowry et al . (2016). *Molecular Ecology Resources*, *17*(3), 356–361. doi:10.1111/1755-0998.12649

McMahon, B. J., Teeling, E. C., & Höglund, J. (2014). How and why should we implement genomics into conservation? *Evolutionary Applications*, *7*(9), 999–1007. doi:10.1111/eva.12193

Melton, C., Reuter, J. A., Spacek, D. V, & Snyder, M. (2015). Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nature Genetics*, *47*(7), 710–716. doi:10.1038/ng.3332

Messer, P. W., & Petrov, D. A. (2013). Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology and Evolution*, *28*(11), 659–669. doi:10.1016/j.tree.2013.08.003

Miller, M. R., Brunelli, J. P., Wheeler, P. A., Liu, S., Rexroad, C. E., Palti, Y., … Thorgaard, G. H. (2012). A conserved haplotype controls parallel adaptation in geographically distant salmonid populations. *Molecular Ecology*, *21*(2), 237–249. doi:10.1111/j.1365-294X.2011.05305.x

Moutsianas, L., Agarwala, V., Fuchsberger, C., Flannick, J., Rivas, M. A., Gaulton, K. J., … McCarthy, M. I. (2015). The Power of Gene-Based Rare Variant Methods to Detect Disease-Associated Variation and Test Hypotheses About Complex Disease. *PLoS Genetics*, *11*(4), 1–24. doi:10.1371/journal.pgen.1005165

Mukherjee, S., Stamatis, D., Bertsch, J., Ovchinnikova, G., Verezemska, O., Isbandi, M., … Reddy, T. B. K. (2017). Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. *Nucleic Acids Research*, *45*(D1), D446–D456. doi:10.1093/nar/gkw992

Muñoz, I., Henriques, D., Johnston, J. S., Ch?vez-Galarza, J., Kryger, P., & Pinto, M. A. (2015). Reduced SNP Panels for Genetic Identification and Introgression Analysis in the Dark Honey Bee (Apis mellifera mellifera). *PLOS ONE*, *10*(4), e0124365. doi:10.1371/journal.pone.0124365

Myburg, A. A., Grattapaglia, D., Tuskan, G. A., Hellsten, U., Hayes, R. D., Grimwood, J., … Schmutz, J. (2014). The genome of Eucalyptus grandis. *Nature*, *510*(7505), 356–362. doi:10.1038/nature13308

Nadachowska-Brzyska, K., Burri, R., Smeds, L., & Ellegren, H. (2016). PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white Ficedula flycatchers. *Molecular Ecology*, *25*(5), 1058–1072. doi:10.1111/mec.13540

Nagasaki, M., Yasuda, J., Katsuoka, F., Nariai, N., Kojima, K., Kawai, Y., … Yamamoto, M. (2015). Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nature Communications*, *6*, 8018. doi:10.1038/ncomms9018

Neale, D. B., Wegrzyn, J. L., Stevens, K. a, Zimin, A. V, Puiu, D., Crepeau, M. W., … Langley, C. H. (2014). Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biology*, *15*(3), R59. doi:10.1186/gb-2014-15-3-r59

Nevado, B., Ramos-Onsins, S. E., & Perez-Enciso, M. (2014). Resequencing studies of nonmodel organisms using closely related reference genomes: Optimal experimental designs and bioinformatics approaches for population genomics. *Molecular Ecology*, *23*(7), 1764–1779. doi:10.1111/mec.12693

Nielsen, R. (2009). Adaptionism - 30 years after gould and lewontin. *Evolution*, *63*(10), 2487–2490. doi:10.1111/j.1558-5646.2009.00799.x

Nielsen, R., Akey, J. M., Jakobsson, M., Pritchard, J. K., Tishkoff, S., & Willerslev, E. (2017). Tracing the peopling of the world through genomics. *Nature*, *541*(7637), 302–310. doi:10.1038/nature21347

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., & Wang, J. (2012). SNP Calling, Genotype Calling, and Sample Allele Frequency Estimation from New-Generation Sequencing Data. *PLoS ONE*, *7*(7), e37558. doi:10.1371/journal.pone.0037558

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews. Genetics*, *12*(6), 443–51. doi:10.1038/nrg2986

Norman, A. J., Street, N. R., & Spong, G. (2013). De Novo SNP Discovery in the Scandinavian Brown Bear (Ursus arctos). *PLoS ONE*, *8*(11), e81012. doi:10.1371/journal.pone.0081012

Nosil, P., Funk, D. J., & Ortiz-Barrientos, D. (2009). Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, *18*(3), 375–402. doi:10.1111/j.1365-294X.2008.03946.x

O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., … Lyon, G. J. (2013). Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Medicine*, *5*(3), 28. doi:10.1186/gm432

Okonechnikov, K., Conesa, A., & García-Alcalde, F. (2015). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, *32*(2), btv566. doi:10.1093/bioinformatics/btv566

Ouborg, N. J., Pertoldi, C., Loeschcke, V., Bijlsma, R. K., & Hedrick, P. W. (2010).

Conservation genetics in transition to conservation genomics. *Trends in Genetics*, *26*(4), 177–187. doi:10.1016/j.tig.2010.01.001

Ozsolak, F., & Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, *12*(2), 87–98. doi:10.1038/nrg2934

Pardo-Diaz, C., Salazar, C., & Jiggins, C. D. (2015). Towards the identification of the loci of adaptive evolution. *Methods in Ecology and Evolution*, *6*(4), 445–464. doi:10.1111/2041-210X.12324

Parejo, M., Wragg, D., Gauthier, L., Vignal, A., Neumann, P., & Neuditschko, M. (2016). Using Whole-Genome Sequence Information to Foster Conservation Efforts for the European Dark Honey Bee, Apis mellifera mellifera. *Frontiers in Ecology and Evolution*, *4*(December), 1–15. doi:10.3389/fevo.2016.00140

Pasaniuc, B., Rohland, N., McLaren, P. J., Garimella, K., Zaitlen, N., Li, H., … Price, A. L. (2012). Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nature Genetics*, *44*(6), 631–635. doi:10.1038/ng.2283

Paten, B., Novak, A. M., Eizenga, J. M., & Garrison, E. (2017). Genome graphs and the evolution of genome inference. *Genome Research*, *27*(5), 665–676. doi:10.1101/gr.214155.116

Pedersen, B. S., Layer, R. M., Quinlan, A. R., Li, H., Wang, K., Li, M., … Kang, H. (2016). Vcfanno: fast, flexible annotation of genetic variants. *Genome Biology*, *17*(1), 118. doi:10.1186/s13059-016-0973-5

Pettersson, E., Lundeberg, J., & Ahmadian, A. (2009). Generations of sequencing technologies. *Genomics*, *93*(2), 105–111. doi:10.1016/j.ygeno.2008.10.003

Pfeifer, S. P. (2017). From next-generation resequencing reads to a high-quality variant data set. *Heredity*, *118*(2), 111–124. doi:10.1038/hdy.2016.102

Phan, V., Gao, S., Tran, Q., & Vo, N. S. (2014). How genome complexity can explain the hardness of aligning reads to genomes. *2014 IEEE 4th International Conference on Computational Advances in Bio and Medical Sciences, ICCABS 2014*, *16*(Suppl 17), 1–15. doi:10.1109/ICCABS.2014.6863916

Phillippy, A. M. (2017). New advances in sequence assembly. *Genome Research*, *27*(5), xi–xiii. doi:10.1101/gr.223057.117

Primmer, C. R. (2009). From conservation genetics to conservation genomics. *Annals of the New York Academy of Sciences*, *1162*, 357–368. doi:10.1111/j.1749-6632.2009.04444.x

Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., … Carroll, M. W. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature*, *530*(7589), 228–232. doi:10.1038/nature16996

Rabbani, B., Tekin, M., & Mahdieh, N. (2014). The promise of whole-exome sequencing in medical genetics. *Journal of Human Genetics*, *59*(1), 5–15. doi:10.1038/jhg.2013.114

Rafati, N., Andersson, L. S., Mikko, S., Feng, C., Pettersson, J., Janecka, J., … Evan, E. (2016). Large Deletions at the SHOX Locus in the Pseudoautosomal Region are associated with Skeletal Atavism in Shetland ponies. *Genes|Genomes|Genetics*, *6*(July), 2213–2223. doi:10.1534/g3.116.029645

Raineri, E., Ferretti, L., Esteve-Codina, A., Nevado, B., Heath, S., & Pérez-Enciso, M. (2012). SNP calling by sequencing pooled samples. *BMC Bioinformatics*, *13*(1), 239. doi:10.1186/1471-2105-13-239

Reinert, K., Langmead, B., Weese, D., & Evers, D. J. (2015). Alignment of Next-Generation Sequencing Reads. *Annual Review of Genomics and Human Genetics*, *16*, 133–51. doi:10.1146/annurev-genom-090413-025358

Rellstab, C., Fischer, M. C., Zoller, S., Graf, R., Tedder, A., Shimizu, K. K., … Gugerli, F. (2016). Local adaptation (mostly) remains local: reassessing environmental associations of climate-related candidate SNPs in Arabidopsis halleri. *Heredity*, *118*(July), 1–9. doi:10.1038/hdy.2016.82

Richards, C. L., Bossdorf, O., & Pigliucci, M. (2010). What Role Does Heritable Epigenetic Variation Play in Phenotypic Evolution? *BioScience*, *60*(3), 232–237. doi:10.1525/bio.2010.60.3.9

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E., Getz, G., & Mesirov, J. P. (2011). Integrative Genomics Viewer. *Nature Biotechnology*, *29*(1), 24–26. doi:10.1038/nbt0111-24

Rockman, M. V. (2012). The QTN program and the alleles that matter for evolution: All that's gold does not glitter. *Evolution*, *66*(1), 1–17. doi:10.1111/j.1558-5646.2011.01486.x

Ronen, R., Udpa, N., Halperin, E., & Bafna, V. (2013). Learning natural selection from the site frequency spectrum. *Genetics*, *195*(1), 181–193. doi:10.1534/genetics.113.152587

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., … Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data. *Genome Biology*, *14*(5), R51. doi:10.1186/gb-2013-14-5-r51

Ruffalo, M., Koyutürk, M., Ray, S., & LaFramboise, T. (2012). Accurate estimation of short read mapping quality for next-generation genome sequencing. *Bioinformatics*, *28*(18), 349–355. doi:10.1093/bioinformatics/bts408

Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., … Yorke, J. a. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research*, *22*(3), 557–567. doi:10.1101/gr.131383.111

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, *74*(12), 5463–5467. doi:10.1073/pnas.74.12.5463

Schiffels, S., & Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, *46*(8), 919–925. doi:10.1038/ng.3015

Schlötterer, C., Tobler, R., Kofler, R., & Nolte, V. (2014). Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. *Nature Reviews Genetics*, *15*(11), 749–763. doi:10.1038/nrg3803

Sedlackova, T., Repiska, G., Celec, P., Szemes, T., & Minarik, G. (2013). Fragmentation of DNA affects the accuracy of the DNA quantitation by the commonly used methods. *Biological Procedures Online*, *15*(1), 5. doi:10.1186/1480-9222-15-5

Shafer, A. B. A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., & Wolf, J. B. W. (2016). Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods in Ecology and Evolution*, 1–11. doi:10.1111/2041-210X.12700

Shafer, A. B. A., Wolf, J. B. W., Alves, P. C., Bergström, L., Bruford, M. W., Brännström, I., … Zieliński, P. (2015). Genomics and the challenging translation into conservation practice. *Trends in Ecology and Evolution*, *30*(2), 78–87. doi:10.1016/j.tree.2014.11.009

Shapiro, M. D., Marks, M. E., Peichel, C. L., Blackman, B. K., Nereng, K. S., Jónsson, B., … Kingsley, D. M. (2006). Corrigendum: Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature*, *439*(7079), 1014–1014. doi:10.1038/nature04500

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, *26*(10), 1135–1145. doi:10.1038/nbt1486

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews. Genetics*, *15*(2), 121–32. doi:10.1038/nrg3642

Sinclair-Waters, M. (2017). *Genomic perspectives for conservation and management of Atlantic cod in costal Labrador. (Unpublished master's thesis).* Dalhousie University, Halifax, Canada.

Skotte, L., Korneliussen, T. S., & Albrechtsen, A. (2013). Estimating individual admixture proportions from next generation sequencing data. *Genetics*, *195*(3), 693–702. doi:10.1534/genetics.113.154138

Snyder-Mackler, N., Majoros, W. H., Yuan, M. L., Shaver, A. O., Gordon, J. B., Kopp, G. H., … Tung, J. (2016). Efficient genome-wide sequencing and low-coverage pedigree analysis from noninvasively collected samples. *Genetics*, *203*(2), 699–714. doi:10.1534/genetics.116.187492

Snyder, M. W., Adey, A., Kitzman, J. O., & Shendure, J. (2015). Haplotype-resolved genome sequencing: experimental methods and applications. *Nature Reviews Genetics*, *16*(6), 344–358. doi:10.1038/nrg3903

Steiner, C. C., Putnam, A. S., Hoeck, P. E. A., & Ryder, O. A. (2013). Conservation genomics of threatened animal species. *Annual Review of Animal Biosciences*, *1*, 261–281. doi:10.1146/annurev-animal-031412-103636

Stetz, J. B., smith, S., Sawaya, M. A., Ramsey, A. B., Amish, S. J., Schwartz, M. K., & Luikart, G. (2016). Discovery of 20,000 RAD–SNPs and development of a 52-SNP array for monitoring river otters. *Conservation Genetics Resources*, *8*(3), 299–302. doi:10.1007/s12686-016-0558-3

Straub, S. C. K., Fishbein, M., Livshultz, T., Foster, Z., Parks, M., Weitemier, K., … Liston, A. (2011). Building a model: developing genomic resources for common milkweed (Asclepias syriaca) with low coverage genome sequencing. *BMC Genomics*, *12*(1), 211. doi:10.1186/1471-2164-12-211

Teacher, A. G., André, C., Jonsson, P. R., & Merilä, J. (2013). Oceanographic connectivity and environmental correlates of genetic structuring in Atlantic herring in the Baltic Sea. *Evolutionary Applications*, *6*(3), 549–567. doi:10.1111/eva.12042

Teacher, A. G., André, C., Merilä, J., & Wheat, C. W. (2012). Whole mitochondrial genome scan for population structure and selection in the Atlantic herring. *BMC Evolutionary Biology*, *12*, 248. doi:10.1186/1471-2148-12-248

The Computational Pan-genomics Consortium. (2016). Computational pan-genomics: status, promises and challenges. *Briefings in Bioinformatics*, (August), 1–18. doi:10.1093/bib/bbw089

The FAASG Consortium. (2016). Functional Analysis of All Salmonid Genomes (FAASG): an international initiative supporting future salmonid research, conservation and aquaculture, 1–18. doi:http://dx.doi.org/10.1101/095737

Therkildsen, N. O., & Palumbi, S. R. (2017). Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. *Molecular Ecology Resources*, *17*(2), 194–208. doi:10.1111/1755-0998.12593

Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, *14*(2), 178–192. doi:10.1093/bib/bbs017

Tiffin, P., & Ross-Ibarra, J. (2014). Advances and limits of using population genetics to understand local adaptation. *Trends in Ecology and Evolution*, *29*(12), 673–680. doi:10.1016/j.tree.2014.10.004

Treangen, T. J., & Salzberg, S. L. (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, *13*(1), 36–46. doi:10.1038/nrg3117

Trowsdale, J., & Knight, J. C. (2013). Major Histocompatibility Complex Genomics and Human Disease. *Annual Review of Genomics and Human Genetics*, *14*(1), 301–323. doi:10.1146/annurev-genom-091212-153455

Tung, J., Zhou, X., Alberts, S. C., Stephens, M., & Gilad, Y. (2015). The genetic architecture of gene expression levels in wild baboons. *ELife*, *4*, 1–22. doi:10.7554/eLife.04729

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., … DePristo, M. A. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. In *Current Protocols in Bioinformatics* (Vol. 11, p. 11.10.1-11.10.33). Hoboken, NJ, USA:

John Wiley & Sons, Inc. doi:10.1002/0471250953.bi1110s43

Vandergast, A. (2017). *Incorporating genetic sampling in long-term monitoring and adaptive management in the San Diego County Management Strategic Plan Area, Southern California*. Virginia, US. doi:/10.3133/ofr20171061

VanderMeer, J. E., & Ahituv, N. (2011). cis-regulatory mutations are a genetic cause of human limb malformations. *Developmental Dynamics*, *240*(5), 920–930. doi:10.1002/dvdy.22535

Varshney, G. K., Pei, W., LaFave, M. C., Idol, J., Xu, L., Gallardo, V., … Burgess, S. M. (2015). High-throughput gene targeting and phenotyping in zebrafish using CRISPR/Cas9. *Genome Research*, *25*(7), 1030–1042. doi:10.1101/gr.186379.114

Vatsiou, A. I., Bazin, E., & Gaggiotti, O. E. (2016). Detection of selective sweeps in structured populations: A comparison of recent methods. *Molecular Ecology*, *25*(1), 89–103. doi:10.1111/mec.13360

Veeckman, E., Ruttink, T., & Vandepoele, K. (2016). Are We There Yet? Reliably Estimating the Completeness of Plant Genome Sequences. *The Plant Cell*, *28*(8), 1759–1768. doi:10.1105/tpc.16.00349

Velasco, D., Hough, J., Aradhya, M., & Ross-Ibarra, J. (2016). Evolutionary Genomics of Peach and Almond Domestication. *Genes|Genomes|Genetics*, *6*(December), 3985–3993. doi:10.1534/g3.116.032672

Vieira, F. G., Albrechtsen, A., & Nielsen, R. (2016). Estimating IBD tracts from low coverage NGS data. *Bioinformatics*, *32*(14), 2096–2102. doi:10.1093/bioinformatics/btw212

Vieira, F. G., Fumagalli, M., Albrechtsen, A., & Nielsen, R. (2013). Estimating inbreeding coefficients from NGS data: Impact on genotype calling and allele frequency estimation. *Genome Research*, *23*(11), 1852–1861. doi:10.1101/gr.157388.113

Vitti, J. J., Grossman, S. R., & Sabeti, P. C. (2013). Detecting Natural Selection in Genomic Data. *Annual Review of Genetics*, *47*(1), 97–120. doi:10.1146/annurev-genet-111212-133526

von der Heyden, S., Beger, M., Toonen, R. J., van Herwerden, L., Juinio-Meñez, M. A., Ravago-Gotanco, R., … Bernardi, G. (2014). The application of genetics to marine management and conservation: examples from the Indo-Pacific. *Bulletin of Marine Science*, *90*(1), 123–158. doi:10.5343/bms.2012.1079

vonHoldt, B. M., Cahill, J. A., Fan, Z., Gronau, I., Robinson, J., Pollinger, J. P., … Wayne, R. K. (2016). Whole-genome sequence analysis shows that two endemic species of North American wolf are admixtures of the coyote and gray wolf. *Science Advances*, *2*(7), e1501714–e1501714. doi:10.1126/sciadv.1501714

Wall, J. D., Schlebusch, S. A., Alberts, S. C., Cox, L. A., Snyder-Mackler, N., Nevonen, K. A., … Tung, J. (2016). Genomewide ancestry and divergence patterns from low-coverage sequencing data reveal a complex history of admixture in wild baboons. *Molecular Ecology*, *25*(14), 3469–3483. doi:10.1111/mec.13684

Wang, H., Xu, X., Vieira, F. G., Xiao, Y., Li, Z., Wang, J., … Chu, C. (2016). The Power of Inbreeding: NGS-Based GWAS of Rice Reveals Convergent Evolution during Rice Domestication. *Molecular Plant*, *9*(7), 975–985. doi:10.1016/j.molp.2016.04.018

Wang, J., Skoog, T., Einarsdottir, E., Kaartokallio, T., Laivuori, H., Grauers, A., … Jiao, H. (2016). Investigation of rare and low-frequency variants using high-throughput sequencing with pooled DNA samples. *Scientific Reports*, *6*(August), 33256. doi:10.1038/srep33256

Waples, R. K., Larson, W. A., & Waples, R. S. (2016). Estimating contemporary effective population size in non-model species using linkage disequilibrium across thousands of loci. *Heredity*, *117*(4), 233–240. doi:10.1038/hdy.2016.60

Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N., & Watson, M. (2015). Exome Sequencing: Current and Future Perspectives. *Genes|Genomes|Genetics*, *5*(8), 1543–1550. doi:10.1534/g3.115.018564

Wittkopp, P. J., & Kalay, G. (2012). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics*, *13*(1), 59–69. doi:10.1038/nrg3095

Wong, P. B., Wiley, E. O., Johnson, W. E., Ryder, O. A., O'Brien, S. J., Haussler, D., … Murphy, R. W. (2012). Tissue sampling methods and standards for vertebrate genomics. *GigaScience*, *1*(1), 8. doi:10.1186/2047-217X-1-8

Wray, G. A. (2007). The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics*, *8*(3), 206–216. doi:10.1038/nrg2063

Xue, Y., Prado-Martinez, J., Sudmant, P. H., Narasimhan, V., Ayub, Q., Szpak, M., … Scally, A. (2015). Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science*, *348*(6231), 242–245. doi:10.1126/science.aaa3952

Yang, H., & Wang, K. (2015). Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nature Protocols*, *10*(10), 1556–1566. doi:10.1038/nprot.2015.105

Yang, J., Li, W. R., Lv, F. H., He, S. G., Tian, S. L., Peng, W. F., … Liu, M. J. (2016). Whole-Genome Sequencing of Native Sheep Provides Insights into Rapid Adaptations to Extreme Environments. *Molecular Biology and Evolution*, *33*(10), 2576–2592. doi:10.1093/molbev/msw129

Ye, H., Meehan, J., Tong, W., & Hong, H. (2015). Alignment of short reads: A crucial step for application of next-generation sequencing data in precision medicine. *Pharmaceutics*, *7*(4), 523–541. doi:10.3390/pharmaceutics7040523

Zhang, G. (2015). Genomics: Bird sequencing project takes off. *Nature*, *522*(7554), 34–34. doi:10.1038/522034d

Zhao, S., Zheng, P., Dong, S., Zhan, X., Wu, Q., Guo, X., … Wei, F. (2012). Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation. *Nature Genetics*, *45*(1), 67–71. doi:10.1038/ng.2494

Zhou, X., Wang, B., Pan, Q., Zhang, J., Kumar, S., Sun, X., … Li, M. (2014). Whole-genome sequencing of the snub-nosed monkey provides insights into folivory and evolutionary history. *Nature Genetics*, *46*(12), 1303–1310. doi:10.1038/ng.3137

CHAPTER 3. PARALLEL ADAPTIVE EVOLUTION OF GEOGRAPHICALLY DISTANT HERRING POPULATIONS ON BOTH SIDES OF THE NORTH ATLANTIC OCEAN

3.1 Abstract

Atlantic herring is an excellent species for studying the genetic basis of adaptation in geographically distant populations, because of its characteristically large population sizes and low genetic drift. In this study, we compared whole genome resequencing data of Atlantic herring populations from both sides of the Atlantic Ocean. An important finding was the very low degree of genetic differentiation among geographically distant populations ($F_{ST} = 0.026$) suggesting lack of reproductive isolation across the ocean. This feature of the Atlantic herring facilitates the detection of genetic factors affecting ecological adaptation because of the sharp contrast between loci showing strong genetic differentiation and the background noise due to genetic drift. We show that genetic factors associated with timing of reproduction are shared between genetically distinct and geographically distant populations. The genes for thyroid-stimulating hormone receptor (*TSHR*), the SOX11 transcription factor (*SOX11*) and calmodulin (*CALM*), all of which have established roles in photoperiodic regulation of reproduction, and estrogen receptor 2 (*ESR2A*), with a significant role in reproductive biology in birds and mammals, were among the loci that showed the most consistent association with spawning time throughout the species range. In fact, the same two SNPs located at the 5'end of *TSHR* showed the most significant association to spawning time both in the East and West Atlantic. We also identified unexpected haplotype sharing between spring-spawning oceanic herring and autumn-spawning populations across the Atlantic Ocean and the Baltic Sea. The genomic regions showing this pattern are unlikely to control spawning time but may be involved in adaptation to ecological factor(s) shared among these populations.

3.2 Significance Statement

Identification of genetic changes that allow a species to adapt to different environmental conditions is an important topic in evolutionary biology. In this study, we analyzed whole genome resequencing data of Atlantic herring populations from both sides of the Atlantic Ocean and identified a number of loci that show consistent associations to spawning time (spring or autumn). Several of these loci have a well-established role in reproductive biology, such as *TSHR*, whereas others have never been implicated in controlling reproduction. Genetic variants associated with adaptation to spring- or autumn-spawning are to a large extent shared among populations across the Atlantic Ocean and the Baltic Sea providing evidence for parallel adaptive evolution.

3.3 Introduction

Widely dispersed and abundant species generally exhibit populations inhabiting divergent habitats. Such populations need to adapt to local biotic and abiotic factors, a process that results in higher fitness in the local environment and leads to genetic and phenotypic differentiation among subpopulations (Miller et al., 2012; Savolainen et al., 2013). In addition to such local adaptation, if a trait responds to similar forces of natural selection independently across multiple populations or species, parallel evolution will lead to similar adaptive changes among geographically distant populations (Pearse & Pogson, 2000). Such parallel adaptation may be caused by convergent evolution or the sharing of similar (or identical) genetic changes across populations (Hoekstra & Nachman, 2003; Jones et al., 2012; Stern, 2013). Identification of the genetic basis for ecological adaptation is a fundamental goal in evolutionary biology (Pearse, Miller, Abadia-Cardoso, & Garza, 2014) and current technological advances in population-scale high-throughput sequencing provide powerful tools to explore these processes at a genomic scale (Hohenlohe et al., 2010). However, many adaptive traits are expected to have a highly polygenic background (Pritchard & Di Rienzo, 2010) where genes of small effect are hard to detect with traditional genome scans, and adaptive changes can often be confounded with demographic history effects such as population structure and genetic drift (Luo, Widmer, & Karrenberg, 2015) making the dissection of such genetic differentiations challenging.

Our recent population-scale genomic study on Atlantic herring has demonstrated that this species is an ideal model for the detection of signatures of selection (Lamichhaney et al., 2012; Martinez Barrio et al., 2016). The large effective population size (one of the most abundant fish species on earth) probably combined with gene flow between populations results in extremely low levels of genetic differentiation at selectively neutral loci across populations exposed to different ecological conditions. This allowed us to identify about 500 independent loci associated with local adaptation as regards the colonization of the brackish Baltic Sea and timing of reproduction in Northeast (NE) Atlantic herring populations (Martinez Barrio et al., 2016).

Atlantic herring is a schooling pelagic fish distributed throughout the North Atlantic Ocean and adjacent waters including the Baltic Sea (Fig. 3.1). The population structure of Atlantic herring is considered to be one of the most complex of any marine fish and there is a long history of attempts to describe it (14). Traditionally, herring stocks have been described based on morphology and life history traits such as spawning time and location (Gröger, Kruse, & Rohlf, 2009; Iles & Sinclair, 1982; McQuinn, 1997). Populations are known to spawn in different seasons with some spawning in the spring, others in the autumn and others still in between. In each case, spawning generally takes place over a protracted period of a few weeks. The optimal spawning time is generally linked to environmental conditions associated with plankton blooms (Sinclair & Tremblay, 1984b). Our recent study revealed highly significant genetic differentiation between spring- and autumn-spawning herring in the NE Atlantic with some of the loci involved in this differentiation likely controlling the timing of reproduction (Martinez Barrio et al., 2016).

Our previous genome-scale studies of Atlantic herring were restricted to population samples from the NE Atlantic, but this species is ecologically important throughout the North Atlantic. In fact, herring support a commercially important fishery in the Northwest (NW) Atlantic (Department of Fisheries and Oceans Canada, 2011) where the species is recognized for the complexity and plasticity of its stocks (McQuinn, 1997; Stephenson et al., 2009). Similar to the NE Atlantic populations, herring in the NW Atlantic also undergo north-south and inshore-offshore migrations for feeding and reproduction (Reid et al., 1999), with spawning taking place mostly during spring and autumn (Department of Fisheries & Oceans Canada, 2012) from Cape Cod to northern Newfoundland (Iles & Sinclair, 1982). Previous genetic studies based on a small number of microsatellite markers reported weak but significant genetic structuring between NW and NE Atlantic populations as well as among spawning aggregations within the NW Atlantic (McPherson, O'Reilly, & Taggart, 2004; McPherson, Stephenson, O'Reilly, Jones, & Taggart, 2001).

The presence of spring- and autumn-spawning herring populations on both sides of the North Atlantic Ocean provides an exceptional opportunity to explore whether the same or similar genetic factors associated with spawning time are shared between geographically distant populations. In this study, we present the results of whole genome resequencing of six NW Atlantic populations (three spring- and three autumn-spawners) and compare their genetic architecture with that of NE Atlantic populations. We demonstrate that genetic factors associated with the timing of reproduction are to a large extent shared between herring populations from the NW and NE Atlantic Ocean.

3.4 Materials and Methods

3.4.1 Sample collection and DNA extraction

Total genomic DNA was extracted from muscle tissue of 287 maturing individuals (or in spawning condition) collected in 2014 during the spawning season at six localities on the East coast of Canada (NW Atlantic), from Newfoundland to the Scotian Shelf (Fig. 3.1, Table 3.1). Tissue samples were stored in 95% ethanol at -20°C until DNA isolation was performed following a standard Phenol:Chloroform:Isoamyl alcohol protocol. The geographic location of sampling sites is shown in Fig. 3.1A. Maps were created using ArcGIS® © Esri. DNA concentrations were measured as ng/µl using the Quant-iT™ PicoGreen® dsDNA Assay (Thermo Fisher Scientific, U.S.) in a Roche LightCycler® 480 Instrument. The integrity of DNA samples was verified with agarose gel electrophoresis and a molecular ladder, where non degraded genomic DNA corresponded to a high molecular weight band around 23 Kb.

3.4.2 Whole genome resequencing and variant calling

Genomic DNA of 41-50 individuals per population were pooled in equimolar concentrations, resulting in 6 DNA pools, one per population. An aliquot of each pool was used for independent library preparations using the TruSeq Nano Illumina kit (Illumina, CA, U.S.). Fragment size selection was performed following the instructions of the manufacturer using AMPURE beads, for an insert size between 450-550 bp. Each population library was paired-end sequenced at 126 cycles in one lane of a HiSeq-2500 sequencer to obtain a depth of coverage of >30x per pool. The quality of the raw reads was evaluated using FastQC v.10.1 (Andrews, 2010) and low quality reads were trimmed using Trimmomatic v.0.33 (Bolger et al., 2014). The trimming of reads followed a sliding window approach where a read was cut at the 3'-end when the average PHRED33 quality score fell below 20 within a 5 bp window. Remaining Illumina adapter sequences were removed using the function ILLUMINACLIP (settings 2:30:10) implemented in Trimmomatic. Only pairs where both reads were recovered after the quality-trimming step were used for downstream analysis.

The high quality trimmed reads were aligned to the reference herring genome assembly (Martinez Barrio et al., 2016) using default parameters of the algorithm BWA-MEM (v.0.6.2) (Li & Durbin, 2009b). The sequences of the six NW Atlantic populations, together with the data of 19 Baltic Sea/NE Atlantic populations and one Pacific herring population from our previous study (Martinez Barrio et al., 2016) were used to call SNP variants across all 26 populations. We followed the recommended workflows of the GATK tool (McKenna et al., 2010; Van der Auwera et al., 2013) for variant discovery. The raw variant calls were filtered using stringent in-house filtering pipeline set up in our previous study (Martinez Barrio et al., 2016). Various standard quality scores generated by GATK such as SNP quality, mapping quality, base quality, mapping quality rank sum, read positions rank sum, allele frequency, and minimum and maximum read depth were used to set up filtering parameters according to GATK best practices recommendations (DePristo et al., 2011). The cut-offs of these quality scores were chosen based on their genome-wide distributions. In addition, as the sequence data of NE and NW Atlantic pools were generated at two different places (NE Atlantic pool in Uppsala, Sweden and NW Atlantic pools in Halifax, Canada), there was a possibility of inflated genetic differentiation between NE and NW Atlantic due to sequencing platform specific technical bias. Hence, we applied more stringent hard filtering of these SNP quality scores, particularly while analyzing data for phylogenetic and NE vs. NW comparisons.

As individual sequencing for all NE and NW samples were done in Uppsala, we also utilized these individual data to evaluate the SNP calls from pooled DNA to exclude false calls that were specific to a sequencing platform.

To explore the haplotype structure at highly differentiated genomic regions, we further performed whole genome resequencing of 37 individuals that included 12 samples from NW Atlantic (two from each population used for pooled sequencing) and nine autumn-spawning samples from NE Atlantic (Table 2). In addition, we also sequenced six individual Pacific herring samples to be used as outgroup (Table 2). Sequencing libraries were constructed with average fragment size of 400 bp and 2x150 bp paired-end reads were generated using Illumina HiSeq2500 sequencing platform. Each library was sequenced to approximate 10x depth of coverage per individual. We combined these data with the whole genome resequencing data of 16 spring spawning NE populations (Table 2) from our previous study (Martinez Barrio et al., 2016). The quality trimming, sequence alignments and variant calling was done using similar pipeline as for pooled sequencing described above.

3.4.3 Genome-wide screens for genetic differentiation

We separately combined the read depth count per SNP from the Pool-seq data of three spring-spawning and three autumn-spawning NW Atlantic populations, obtaining two separate super-pools, one for spring and one for autumn. Then we compared the allele frequency differences SNP by SNP between the super-pools using a 2 by 2 contingency χ^2 test. We compared these results against the similar comparisons of autumn- vs. spring-spawners from NE Atlantic populations of our previous study (Martinez Barrio et al., 2016). In addition to the contingency χ^2 tests, we also screened for genetic differentiation using commonly used methods like F_{ST} and pooled heterozygosity (Rubin et al., 2010). The SNPs showing highly significant differentiation were clustered into independent genomic regions as described (Martinez Barrio et al., 2016).

3.4.4 Simulations of genetic drift

The simulations aimed at assessing the expected distribution of F_{ST} under selective neutrality were conducted as in our previous paper on divergence among herring populations in the NE Atlantic (Lamichhaney et al., 2012). In brief, we used a slightly modified version of the Powsim software (Ryman & Palm, 2006) that mimics sampling from populations at a predefined level of expected divergence through random number simulations under a classical Wright–Fisher model without migration or mutation. An infinitely large base population segregating for a specified number of independent, selectively neutral loci is divided into *s* subpopulations of equal effective size (N_e) through random sampling of $2N_e$ genes. Each of the subpopulations is allowed to drift for *t* generations, and the expected degree of divergence in generation *t* is then $F_{ST} = 1\text{-}(1\text{-}1/2N_e)^t$ (Nei, 1987). The populations are sampled when the expected degree of differentiation has been reached, F_{ST} (Weir & Cockerham, 1984) is then calculated for each locus, and the distribution of simulated F_{ST}-values is compared to the observed one.

To reduce statistical noise in the observed distribution we restricted the analysis to SNPs that had 40-45 reads in all populations, and among these SNPs we calculated F_{ST} for a random sample of 50 000 ones. The corresponding simulation was run with effective sizes of $N_e = 5\ 000$, the number of loci (SNPs) was 50 000, and the time of divergence (*t*) was set to result in an expected F_{ST} identical to the average of that found in the observed distribution.

3.4.5 Phylogenetic analysis

Genetic divergence between populations was calculated using PLINK (Purcell et al., 2007) and phylogenetic trees based on allele frequencies estimated from pooled sequencing data were generated using PHYLIP (Felsenstein, 1989). Genetic distances were calculated using an identity by state (IBS) similarity matrix (Table S3.4). The bootstrapping of the phylogenetic tree was done using the Phylip Seqboot package (Felsenstein, 1989). The phased haplotypes were generated from the genomic regions showing high differentiation among populations using BEAGLE (Browning & Browning, 2016) and maximum likelihood haplotype trees were built generated using FastTree (Price, Dehal, & Arkin, 2010).

3.4.6 CNV analysis of the *HCE* locus

We used GATK:DepthOfCoverage to extract read depth coverage of all populations. All reads with mapping quality below 20 were filtered out. We then generated 1Kb non-overlapping windows and normalized them against AB1 pool that had the highest sequence coverage among all samples in our previous study (Martinez Barrio et al., 2016). We scanned the genome to identify copy number variation (CNV) between autumn- and spring-spawners from both side of Atlantic by selecting the following populations:

i) Autumn spawners: WBoB, WGeB, WNsF, BF, BAH, NS.

ii) Spring spawners: WFB, WInB, WNsS, BAV, BH, AB1.

After filtering windows with low depth we compared both groups in 715 093 windows by analysis of variance ANOVA.

3.4.7 Genotyping individual fish in a subset of SNPs

To validate the candidate loci, we genotyped 384 individuals (192 from NW and 192 from NE Atlantic populations, using a Sequenom SNP panel. To do so, we chose 105 SNPs with the following three criteria: (i) 35 SNPs showing highly significant differences in allele frequencies between autumn- and spring-spawning populations that are shared in herring from NW and NE Atlantic; (ii) 35 SNPs that are unique to herring from the NE Atlantic and show highly significant differences in allele frequency between autumn- and spring-spawning populations; (iii) 35 SNPs that are unique to herring from NW Atlantic and show highly significant differences in allele frequency between autumn- and spring-spawning populations.

The genotype data were analyzed for standard quality control using PLINK (Purcell et al., 2007). SNPs with missing genotypes in >10% of the individuals were excluded. Similarly, individuals that had missing genotypes in >10% of SNPs were also excluded. 103 SNPs genotyped in 377 individuals passed the above-mentioned thresholds and were used for the subsequent downstream analysis. The allele frequency estimates from individual genotyping were compared with the estimates from pooled whole genome sequencing. The genotyping results were consistent with the results based on pooled sequencing. The correlation coefficients of allele frequency estimates from individual and pooled-sequencing were in the range r = 0.95 – 0.97 in respective populations. The haplotype structures of the candidate loci based on Sequenom genotyping of individual fish resembled the ones generated by pooled sequencing (Figure 4 and Figure S2).

3.5 Results

3.5.1 Whole genome resequencing

Whole genome resequencing of pooled DNA of 40-50 individuals per location was conducted for six population samples from the NW Atlantic: two populations from Newfoundland, three from the Gulf of St Lawrence, and one from the Scotian Shelf (Fig. 3.1, Table 3.1). Each library had a ~30-50x depth of coverage. These data were compared with pooled sequence data from 19 populations of Baltic Sea/NE Atlantic herring (Fig. 3.1) and one population of Pacific herring from our previous study (Martinez Barrio et al., 2016); the Pacific herring data was used as an outgroup in the phylogenetic analysis and allowed us to reveal the ancestral state for candidate causal mutations. In addition, 37 individual samples of spring- and autumn-spawning herring from both sides of North Atlantic and six individual samples of Pacific herring were sequenced to ~10x depth of coverage to characterize haplotypes showing genetic differentiation. These sequences were aligned to the reference herring genome (Martinez Barrio et al., 2016) and SNP calling was conducted using an in-house rigorous quality-filtering pipeline (see Methods) to identify 8.9 million SNPs that were polymorphic in at least one population (including Pacific herring).

3.5.2 Phylogeny and population genetics

In agreement with the results from our previous study (Martinez Barrio et al., 2016), the neighbor-joining tree generated using 1.2 million high-quality SNPs (see Methods) revealed a large genetic distance between the Pacific and all 25 Atlantic herring populations (Fig. 3.2, left panel). The Atlantic herring populations in general clustered according to their geographic origin (Fig. 3.2, right panel). The populations formed three major groupings: (i) Atlantic herring from NW Atlantic and NE Atlantic (AB1 and AI), (ii) Atlantic herring from the North Sea, Skagerrak and Kattegat and (iii) spring-spawning herring from the Baltic Sea. Two populations of autumn-spawning herring from the Baltic Sea (BÄH and BF) deviated from this pattern and did not cluster with spring-spawning herring from the Baltic Sea. Two populations stood out with relatively long branch lengths (NS and WFB); a careful examination of the data from these populations did not indicate that these long branch lengths were due to technical issues. Among the six NW Atlantic populations, spring- and autumn-spawners formed their own clusters, indicating that populations spawning in different seasons are genetically distinguishable. NW Atlantic autumn-spawning herring were more closely related to NW Atlantic spring-spawning herring than to autumn-spawning herring from the NE Atlantic.

The average F_{ST} among the 25 populations of Atlantic and Baltic herring was as low as 0.026, a value that drops to 0.019 if we exclude the WFB population. This result is consistent with the tight clustering of populations in the phylogenetic tree (Fig. 3.2). This minute level of genetic differentiation is remarkable given that our samples now include herring populations from both sides of the Atlantic, from North Sea, Skagerrak, Kattegat, and the brackish Baltic Sea (Fig. 3.1). We performed a computer simulation study to investigate if the genetic differentiation among these 25 populations seemed to be primarily driven by selection or drift. We used 50 000 randomly sampled SNPs with 40-45 reads in each population and estimated F_{ST} (Weir & Cockerham, 1984) among all 25 populations for each locus. We then used simulation to generate a dataset reflecting the expected distribution of F_{ST} values for 50 000 selectively neutral loci based on 25 populations each with an effective population size of $N_e = 5\ 000$ that were separated for $t = 263$ generations and have an expected F_{ST} identical to the observed one ($F_{\mathrm{ST}} = 0.026$; when excluding the WFB outlier population the corresponding values were $F_{\mathrm{ST}} = 0.019$ and $t = 192$ generations). The observed data deviated significantly from the simulated data due to a long tail of F_{ST} values > 0.10 (Fig. 3.3). We conclude that the great majority of SNPs in this dataset with $F_{\mathrm{ST}} > 0.10$ and with significant allele frequency differences between populations are located in the vicinity of sequence polymorphisms showing genetic differentiation due to natural selection.

3.5.3 Genetic differentiation between spring- and autumn-spawning populations

We explored the genomic regions showing differentiation between spring- and autumn-spawners from NW Atlantic by comparing allele frequencies between the two groups SNP by SNP and identified 6 333 SNPs with significant allele frequency differences ($P < 1x10^{-10}$; Fig. 3.4A, upper panel). These SNPs are conservatively estimated to represent at least 182 independent genomic loci (Martinez Barrio et al., 2016). We compared these results with the loci that were associated with spawning time in the Baltic Sea/NE Atlantic populations (Fig. 3.4A, lower panel) (Martinez Barrio et al., 2016). About 28% (1 747 out of 6 333) of the associated SNPs in the NW Atlantic also reached statistical significance in the Baltic Sea/NE Atlantic comparison. The genetic signals associated with seasonal reproduction in the NE and NW Atlantic populations showed a considerable overlap (Fig. 3.4B). Six of eight previously identified genomic regions exhibiting the most consistent association with spawning time in NE Atlantic and Baltic populations replicated in the data from NW Atlantic (Table S3.1). At these six loci the same haplotype group is associated with autumn- or spring-spawning in all populations included in this study.

The results suggest that genetic factors affecting the timing of reproduction in herring are to a large extent shared among herring populations from both sides of the Atlantic. In contrast, genome-wide data indicate a closer genetic relationship between spring- and autumn-spawning herring from the same geographic region than between either spring- or autumn-spawning populations from different regions (Fig. 3.2, right panel). Although many loci associated with the onset of reproduction were shared between the NE and NW Atlantic herring populations, there were certain genomic regions unique to each geographic region (Fig. 3.4A, B). Apart from the timing of reproduction, spring- and autumn-spawning populations need to adapt to a variety of other ecological factors not considered in this study. The differentiated loci between spring- and autumn-spawners not shared between NE and NW Atlantic populations most likely reflect such local adaptations. The list of loci showing differentiation between spring- and autumn-spawners, the genes in their vicinity and additional annotations of these regions are in Table S3.2.

3.5.4 Evidence of parallel evolution at TSHR, a major locus associated with timing of reproduction

TSHR, encoding thyroid-stimulating hormone receptor, has a key role in photoperiodic regulation of reproduction in birds and mammals (Hanon et al., 2015; Nakao et al., 2008; Ono et al., 2008). The *TSHR* region showed the most significant allele frequency difference between spring- and autumn-spawners in both NE and NW Atlantic populations (Fig. 3.4A). The signatures of genetic differentiation within the ~120 Kb block around *TSHR* were strikingly similar in NE and NW Atlantic populations (Fig. 3.4C). All autumn- and spring-spawning populations clustered separately regardless of their geographic origin in a neighbor-joining tree based on allele frequency data on 940 SNPs from this ~120 Kb region (Fig. 3.4D).

To reveal the haplotype structure at the *TSHR* locus, we used our whole genome resequencing data of individual fish (16 spring-/9 autumn-spawning herring from NE Atlantic populations, 6 spring-/6 autumn-spawning herring from NW Atlantic populations and 6 Pacific herring individuals as an outgroup). These data were used to generate a maximum likelihood tree for haplotypes at the 120 Kb *TSHR* region (Fig. 3.4E). Consistent with the results from pooled DNA sequencing (Fig. 3.4D), haplotypes from spring- and autumn-spawning individuals clustered as two distinct groups and there was no clear sub-structuring related to geographic origin. The short branch lengths for *TSHR* haplotypes in spring-spawners most likely reflect a recent selective sweep.

Two SNPs (upstream of *TSHR*) were found to be the most differentiated in the spring- vs. autumn-spawning contrast in both NE and NW Atlantic herring populations (Fig. 3.4B). We generated phylogenetic conservation scores around these sites by comparing nine fish genomes, including Atlantic herring from our previous study (Martinez Barrio et al., 2016). One of these SNPs (Scaffold1420:133,030 bp) overlapped a conserved region. These highly differentiated *TSHR* SNPs are candidate mutations that may regulate *TSHR* expression in cells that are critical for the initiation of reproduction.

3.5.5 Haplotype sharing between spring-spawning oceanic herring and autumn-spawning populations

We compared allele frequencies across all populations of herring at the loci showing highly significant allele frequency differences between spring- and autumn-spawners, partially to explore whether the very same alleles were associated with spawning time across the North Atlantic Ocean. The heat maps summarizing these data show two distinct patterns that we designate A and B (Fig. 3.5A). For loci belonging to pattern A, including the *TSHR* locus, allele frequency differences between spring- and autumn-spawning herring were remarkably consistent, independent of geographic origin. In contrast, in pattern B involving SNPs located on eight different genomic scaffolds, the two oceanic spring-spawning herring populations sampled along the coasts of Norway (AB1) and Iceland (AI) were fixed for the same alleles as the autumn-spawners from the NW and NE Atlantic, whereas the alternative alleles dominated in the majority of spring-spawning populations from both sides of the Atlantic including the Baltic Sea (Fig. 3.5A). Six of these eight genomic scaffolds include members of *Myosin heavy chain (MYHC)* gene family. Intermediate allele frequencies at the majority of these loci were observed in the two samples from Skagerrak (SB and SH), the transition area between the North Sea and the Baltic Sea (Figs 1, 5A).

A comparison with Pacific herring indicated that autumn spawners carried the ancestral allele at the great majority of SNPs (70.1%) belonging to pattern A, whereas they were associated with the derived allele at the majority of loci (69.9%) belonging to pattern B (Fig. 3.5A). The difference in the proportion of derived alleles associated with autumn-spawning among pattern A and B is highly significant ($P = 2.0 \times 10^{-18}$, Fisher's exact test) and is most likely explained by linkage drag when alleles under selection have increased in frequency.

We further investigated individual haplotypes at the eight scaffolds associated with pattern B (Fig. 3.5B) in 43 fish including spring- and autumn-spawners from both sides of the Atlantic and six Pacific herring (Table 3.1). These eight loci showed very strong linkage disequilibrium across populations despite the fact that they are spread in several genomic regions. All autumn-spawning herring were essentially fixed for the same allele associated with spawning season whereas more heterogeneity was detected in spring-spawning populations (Fig. 3.5B). For example, in a spring-spawning population from the Baltic Sea (BH) considerable heterogeneity was observed. If these were sequence polymorphisms in a randomly mating population one would expect that heterozygosity at these 189 SNPs from eight different scaffolds would be distributed more or less randomly among individuals, but in the BH population two individuals are heterozygous at most positions whereas five individuals are essentially homozygous at all positions. Thus, a possible explanation is that the heterozygous fish represent hybrids between populations fixed for different alleles at these genomic regions. Similarly, in a spring-spawning population from NW Atlantic (WFB) (Table 3.1), one fish was mostly homozygous for the autumn-spawning alleles while the other was heterozygous at most diagnostic sites. The latter fish was also heterozygous at the *TSHR* locus (Fig. 3.4E). The four individuals from the other two spring-spawning populations from NW Atlantic (WInB and WNsS) were all homozygous for spring-spawning alleles (Fig. 3.5B). The eight fish sampled in spring 2013 close to Bergen (Norway) (AB2) showed considerable heterogeneity at these eight loci in contrast to the homogeneity observed in the sample AB1 collected in the same geographic area in February 1980 (Fig. 3.5A). Some fish were homozygous for either (i) alleles associated with spring spawning or (ii) alleles abundant in autumn-spawning and oceanic herring whereas others had mixed haplotypes. It is possible that the AB2 sample represents a mix of individuals coming from different spring-spawning populations in Norway and their hybrids. Interestingly, data on somatic growth patterns and morphometric measurements indicated that interbreeding has occurred in the last 50 years between a resident, coastal spring-spawning population and the migratory, oceanic Norwegian spring spawning herring (Johannessen et al., 2014); the sample AB1 in the current study is expected to represent this latter stock (Fig. 3.5A). This interbreeding likely took place after the collapse of the oceanic Norwegian spring spawning population due to overfishing in the late 1960s.

3.5.6 Validation of loci associated with spawning season

The genome-wide scans used for the detection of genetic differentiation among herring populations were based on the comparison of allele frequencies estimated from pooled DNA sequencing data. Such data can be prone to biases due to different factors, such as sequencing and mapping errors or differences in library or sequencing protocols, that could lead to uneven read depth distribution among loci and populations which could affect allele frequency estimates (Schlötterer et al., 2014). Hence, to validate the results from pooled DNA sequencing, we genotyped a subset of highly differentiated SNPs between spring- and autumn-spawning populations in 377 individual fish (about 30 individuals per each of 13 populations). A total of 103 SNPs were selected using the following criteria, (i) those showing a consistent association with spawning time across the Atlantic and (ii) those unique to NE or NW Atlantic. The individual genotyping results were consistent with the genetic signatures detected by pooled sequencing (Fig. S3.1).

For SNPs shared between NE and NW Atlantic, the spring-spawners tend to carry one haplotype whereas the autumn-spawners carry the alternative haplotype regardless of geographical location (Fig. S3.1A). For SNPs uniquely associated with spawning time in the NE Atlantic, there were two clear patterns (Fig. S3.1B). In the first pattern, the autumn-spawners from the North Sea carried unique haplotypes, not seen in any other population. In the second pattern, spring- and summer-spawners from the locations Gävle (BÄV and BÄS) and Kalix (BK) in the Baltic Sea (Table 3.1) shared a unique haplotype. For SNPs unique to the NW Atlantic, there was also a set of SNPs showing consistent differences between spring- and autumn-spawning populations only within the NW Atlantic (Fig. S3.1C). Some individuals from the autumn-spawning North Sea (NS) population were segregating for some of these alleles present in the autumn-spawning populations in the NW Atlantic whereas autumn-spawning herring from the Baltic Sea as well as all spring-spawning herring from the NE and NW Atlantic tended to be fixed for the alternative alleles.

3.5.7 Validation of loci associated with adaptation to low salinity in the brackish Baltic Sea

In our previous study (Martinez Barrio et al., 2016), we identified about 3 000 SNPs representing about 100 independent genomic regions showing the most consistent correlation of allele frequencies with salinity by comparing populations adapted to the brackish Baltic Sea and the marine NE Atlantic (Table S3.3). The loci that are directly related to adaptations to low salinity are also expected to show strong genetic differentiation between the populations from the Baltic Sea and NW Atlantic, but not between NE and NW Atlantic where salinity is the same (35‰). To test this, we calculated delta allele frequencies (dAF) between pairwise comparisons of Baltic, NW Atlantic and NE Atlantic populations. The results showed that as many as 94.4% of the loci that showed a dAF>0.2 between Baltic and NE Atlantic also showed a dAF>0.2 in the comparison of Baltic vs. NW Atlantic (Fig. S3.2A). In contrast, few of these loci (12.7%) showed a dAF>0.2 in the contrast between NE and NW Atlantic populations (Fig. S3.2B). Thus, the great majority of loci showing a consistent association to differences in salinity reported in our previous study were supported using the new data from NW Atlantic populations. The statistical support for allele frequency differences in different pairwise comparisons among these groups are given in Table S3.3.

We previously identified a copy number variation (CNV) overlapping the gene for high choriolytic enzyme *(HCE)* that correlated with salinity. Populations from the brackish Baltic Sea (3-12‰) had a high copy number whereas populations from marine waters in the NE Atlantic (20-35‰) had a relatively low copy number. As expected, all six NW Atlantic populations (35‰) showed a low copy number at this locus (Fig. S3.2C).

3.6 Discussion

Independent populations that are exposed to similar environmental conditions often evolve similar phenotypic traits. There are widespread examples of such parallel evolution in nature and in some cases the genetic basis is known to some extent e.g. (Bradbury et al., 2010; Hohenlohe et al., 2010; Jones et al., 2012). The Atlantic herring provides an opportunity to study the genetic basis of such repeated parallel evolution in geographically distant populations. In this study, we have demonstrated that genetic variants associated with adaptation to different spawning times (spring and autumn) are to a large extent shared among geographically distant populations of herring. This finding resembles the reuse of standing genetic variation for adaptation to marine and freshwater environments in the three-spine stickleback (5). Thus, even though autumn-spawning herring from the Baltic Sea, the North Sea and the NW Atlantic show genetic differentiation related to other traits (e.g. salinity) they share very similar haplotypes at loci that are strong candidates for underlying the timing of reproduction. This haplotype sharing implies that these variants were present in a common ancestor of these subpopulations or that they have been spread among populations due to gene flow. The different populations are too closely related to efficiently distinguish which of these scenarios has been most important but the very low levels of genetic differentiation among the 25 population samples included in this study ($F_{\mathrm{ST}} = 0.026$) suggests a lack of reproductive isolation throughout the species range. This result is in sharp contrast to the situation for Atlantic salmon which shows strong genetic differentiation between populations from different continents (King, Kalinowski, Schill, Spidle, & Lubinski, 2001; Stahl, 1983).

We identified a total of six independent loci that show a consistent association with spawning time across populations from the NW and NE Atlantic Ocean as well as in the brackish Baltic Sea (Fig. 3.4A; Table S3.1). It is likely that these loci contribute to how the timing of reproduction is determined because one of the main environmental cues for this, the change in day length, should be the same on both sides of the Atlantic. In contrast, those loci that are only associated with spawning time in one geographic region may reflect local adaptation. However, at present, the allele substitution effects at these six loci and the extent to which spawning time is genetically determined are unknown.

The *TSHR* locus shows the most convincing association with spawning time as the two SNPs located upstream of this gene exhibit the most significant allele frequency difference on both sides of the Atlantic (Fig. 3.4B). We propose that one of these, if not both, are causative changes. Another interesting finding is the large haplotype block around *TSHR*, about 120 Kb in size (Fig. 3.4C), associated with spawning time in contrast to the very rapid decay of linkage disequilibrium in parts of the herring genome not associated with ecological adaptation (Martinez Barrio et al., 2016). The large haplotype block may be maintained by an inversion or more likely by the presence of multiple causative mutations across the associated region. In our previous study (Martinez Barrio et al., 2016) we found no indications for the presence of an inversion at the *TSHR* locus but the ability to detect inversions using short insert, paired reads is limited. The branch lengths in the phylogenetic tree for *TSHR* haplotypes associated with spring spawning are much shorter than the branch lengths for the haplotypes associated with autumn spawning, implying a more recent coalescence time for the former ones (Fig. 3.4E). This indicates that a relatively recent selective sweep has occurred in this region and that spring-spawning may be a derived trait in the herring. There is in fact a clear trend that spring-spawning haplotypes also carry the derived allele at many of the SNPs (70.1%) showing the most consistent association with spawning time (Fig. 3.5A, pattern A); the derived state was deduced using the Pacific herring as an outgroup. Furthermore, phylogenetic trees generated from individual haplotypes for other loci showing strong association to spawning time also showed a general trend for shorter coalescence time for spring-spawning haplotypes (Fig. S3.3). Thus, we propose that autumn-spawning is the ancestral state in Atlantic herring.

Three of the loci showing the most consistent association with spawning time in herring contain genes (*TSHR*, *SOX11* and *CALM*) with an established role in photoperiodic regulation of reproduction in birds and mammals (Hanon et al., 2015; Kim et al., 2011; Melamed et al., 2012; Nakao et al., 2008; Ono et al., 2008). Functional studies are now required to confirm that these candidate loci contribute to photoperiodic regulation of reproduction also in herring. The robust associations reported here provide a unique opportunity to dissect the underlying molecular mechanisms for how increasing (spring) or decreasing (autumn) day length is translated to the initiation of spawning in different populations. The estrogen receptor beta 2 locus (*ESR2A*) has never been implicated in photoperiodic regulation of reproduction but has an established role in reproductive biology (Bondesson, Hao, Lin, Williams, & Gustafsson, 2015). Other loci also with very convincing associations to spawning season such as *HERPUD2* (Homocysteine-responsive endoplasmic reticulum-resident ubiquitin-like domain member 2) and *SYNE2* (Spectrin Repeat Containing Nuclear Envelope Protein 2) have no known role in reproduction. Further studies on these associations provide an opportunity to establish new functional roles for these genes. *SYNE2* is in fact the neighboring gene to *ESR2* in all vertebrates sequenced to date and the peaks of association in the two genes are only about 100 Kb apart, but clearly separated by a region of weaker association (Fig. S3.4). It is an open question whether the *SYNE2* association is related to *SYNE2* function or if the region harbors long-range regulatory elements controlling *ESR2* expression (or vice versa). An association study like this reveals the location of causal mutations but if these are regulatory the target gene(s) showing altered expression may be located outside the associated region.

One of the most interesting observations in this study was the unexpected haplotype sharing between spring-spawning oceanic and autumn-spawning populations. There was an over-representation of members of the myosin heavy chain (*MYHC*) gene family in these regions. There are 23 annotated *MYHC* genes in the current herring genome assembly and as many as six (26%) of these are located in these regions that only constitute 0.04% of the assembly. *MYHC* genes play a critical role for myogenesis (Watabe, 1999). Previous studies have indicated that herring populations spawning at various times of the year have a variable degree of developmental plasticity as regards myogenesis (Johnston, Vieira, & Temple, 2001) and that differences in water temperature between spawning seasons are considered responsible for differential myogenesis in herring (Temple, Cole, & Johnston, 2001). Further research is required to reveal the ecological adaptation underlying the observed association.

This study has important practical implications for herring fishery in the NW Atlantic since it provides genetic markers that can distinguish spring- and autumn-spawning herring outside the breeding season. Such diagnostic method can be applied to develop a more sustainable fishery by optimizing the intensity of fishing among stocks according to their abundance.

3.7 Acknowledgements

This work was funded by the European Research Council project BATESON (L.A.), the GENSINC project funded by the Norwegian Research Council (L.A.), and Discovery Grant RGPIN-2014-04696 (to D.E.R.). A.P.F.-P. was funded by a Vanier Canada Graduate Scholarship, the Killam Trust, and a Nova Scotia Graduate Scholarship. Simulation work by N. Ryman was supported by the Swedish Research Council and the Swedish Research Council Formas grants to Linda Laikre. Pooled DNA sequencing of NW Atlantic populations was performed at The Centre for Applied Genomics of the Hospital for Sick Children, Toronto. Sequencing of individual samples was performed by the SNP&SEQ Technology Platform, which is supported by Uppsala University and Hospital, the Science for Life Laboratory, and Swedish Research Council Grants

0576801 and 70374401. Computer resources were provided by the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX), Uppsala University.

3.8 Author contributions

D.E.R. and L.A. conceived the study; S.L., A.P.F.-P., N. Rafati, and N. Ryman performed research; G.R.M., C.B., and R.S. contributed critical samples for the study; and S.L., A.P.F.P., D.E.R., and L.A. wrote the paper. Abbreviations of names as described in the statement of co-authorship (page 7).

3.9 Tables

Table 3.1 Population samples of herring used in the study. Abbreviations: n=number of fish; n2=number of fish used for individual sequencing.

Locality[a]	Sample	n (n2)	Position		Salinity	Date	Spawning
					(‰)	(yy/mm/dd)	season
Baltic Sea							
Gulf of Bothnia (Kalix)[b]	BK	47	N 65°52’	E 22°43’	3	800629	summer
Bothnian Sea (Hudiksvall)[b]	BU	100	N 61°45’	E 17°30’	6	120419	spring
Bothnian Sea (Gävle)[b]	BÄV	100	N 60°43’	E 17°18’	6	120507	spring
Bothnian Sea (Gävle)[b]	BÄS	100	N 60°43’	E 17°18’	6	120718	summer
Bothnian Sea (Gävle)[b]	BÄH	100 (3)	N 60°44’	E 17°35’	6	120904	autumn
Bothnian Sea (Hästskär)[b]	BH	50 (8)	N 60°35’	E 17°48’	6	130522	spring
Central Baltic Sea (Vaxholm)[b]	BV	50	N 59°26’	E 18°18’	6	790827	spring
Central Baltic Sea (Gamleby)[b]	BG	49	N 57°50’	E 16°27’	7	790820	spring
Central Baltic Sea (Kalmar)[b]	BR	100	N 57°39’	E 17°07’	7	120509	spring
Central Baltic Sea (Karlskrona)[b]	BA	100	N 56°10’	E 15°33’	7	120530	spring
Central Baltic Sea[b]	BC	100	N 55°24’	E 15°51’	8	111018	unknown
Southern Baltic Sea (Fehmarn)[b]	BF	50	N 54°50’	E 11°30’	12	790923	autumn
Kattegat, Skagerrak, North Sea, East Atlantic Ocean							
Kattegat (Träslövsläge)[b]	KT	50	N 57°03’	E 12°11’	20	781023	unknown
Kattegat (Björköfjorden)[b]	KB	100	N 57°43’	E 11°42’	23	120312	spring
Skagerrak (Brofjorden)[b]	SB	100	N 58°19’	E 11°21’	25	120320	spring
Skagerrak (Hamburgsund)[b]	SH	49	N 58°30’	E 11°13’	25	790319	spring
North Sea[b]	NS	49 (3)	N 58°06’	E 06°10’	35	790805	autumn
Atlantic Ocean (Bergen)[b]	AB1	49	N 64°52’	E 10°15’	35	800207	spring
Atlantic Ocean (Bergen)[b]	AB2	8	N 60°35’	E 05°00’	33	130522	spring
Atlantic Ocean (Höfn)[b]	AI	100	N 65°49’	W 12°58’	35	110915	spring
Pacific Ocean							
Strait of Georgia (Vancouver)[b]	PH	50 (6)	-	-	35	121124	-
West Atlantic Ocean							
Bonavista Bay[c]	WBoB	49 (2)	N 48º 49'	W 53º 20'	35	140625	autumn
Fortune Bay[c]	WFB	50 (2)	N 47º 17'	W 55º 38'	35	140526	spring

Locality[a]	Sample	n (n2)	Position		Salinity (‰)	Date (yy/mm/dd)	Spawning season
Inner Baie Des Chaleurs[c]	WInB	41 (2)	N 48º 00'	W 65º 51'	35	140508	spring
Northumberland Strait[c]	WNsS	49 (2)	N 46º 19'	W 64º 09'	35	140506	spring
Northumberland Strait[c]	WNsF	50 (2)	N 45º 44'	W 62º 36'	35	140916	autumn
German Banks[c]	WGeB	48 (2)	N 43º 16'	W 66º 18'	35	140828	autumn

[a] **Places where the sample was landed (if known) are given in parenthesis**

[b] **Samples from our previous studies** (Lamichhaney et al., 2012; Martinez Barrio et al., 2016)

[c] **New samples sequenced in this study.**

3.10 Figures

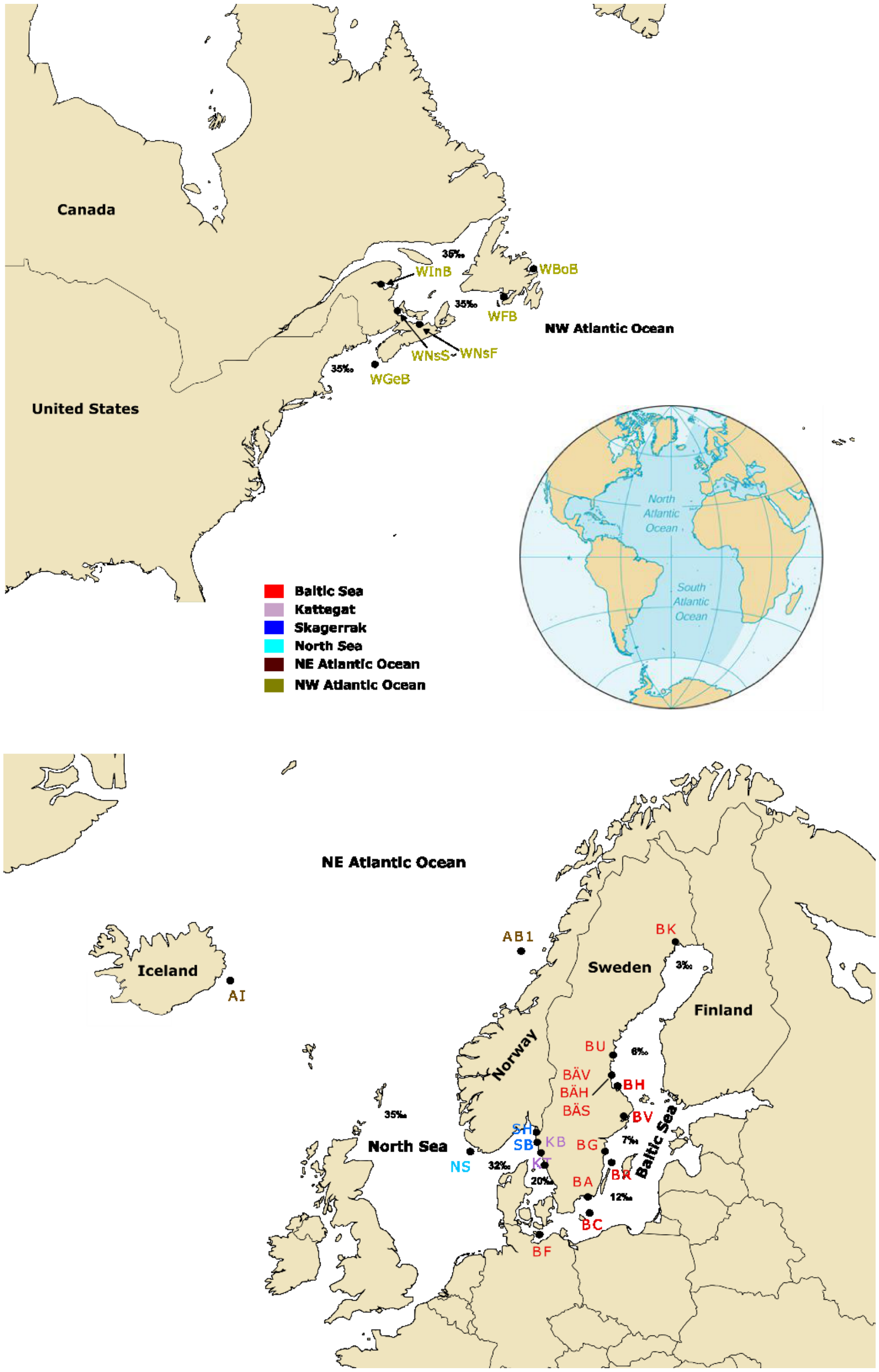


Figure 3.1 Sampling sites. Geographic location of all population samples. Abbreviations for localities are given in Table 3.1. The relative locations of populations sampled from NW and NE Atlantic Ocean are indicated in the inserted globe.

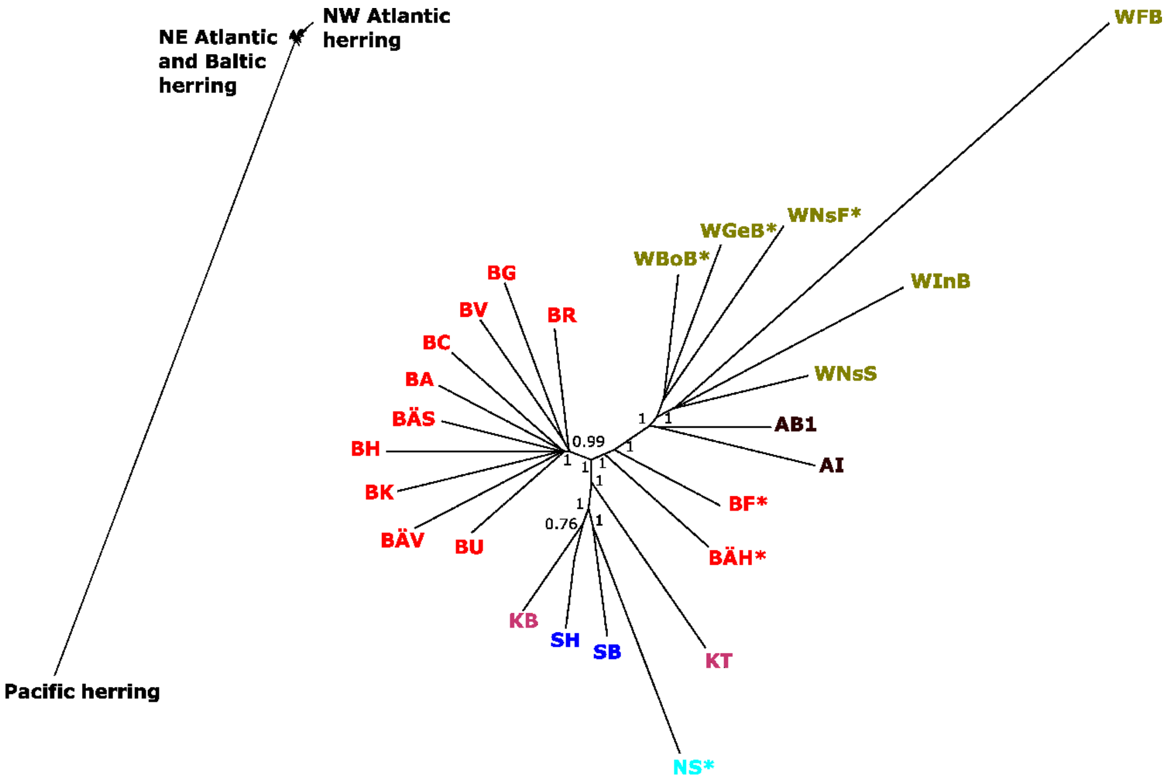


Figure 3.2 Neighbor-joining phylogenetic tree. Right panel, zoom-in on the cluster of NE and NW Atlantic herring populations. Color codes for sampling locations are the same as in Fig. 3.1. Autumn-spawning populations are marked with an asterisk. To restrict the sampling variance and sequencing bias across NE and NW, the phylogenetic tree was based on ~1.2 million SNPs each represented by 40-45 reads per population (see Methods).

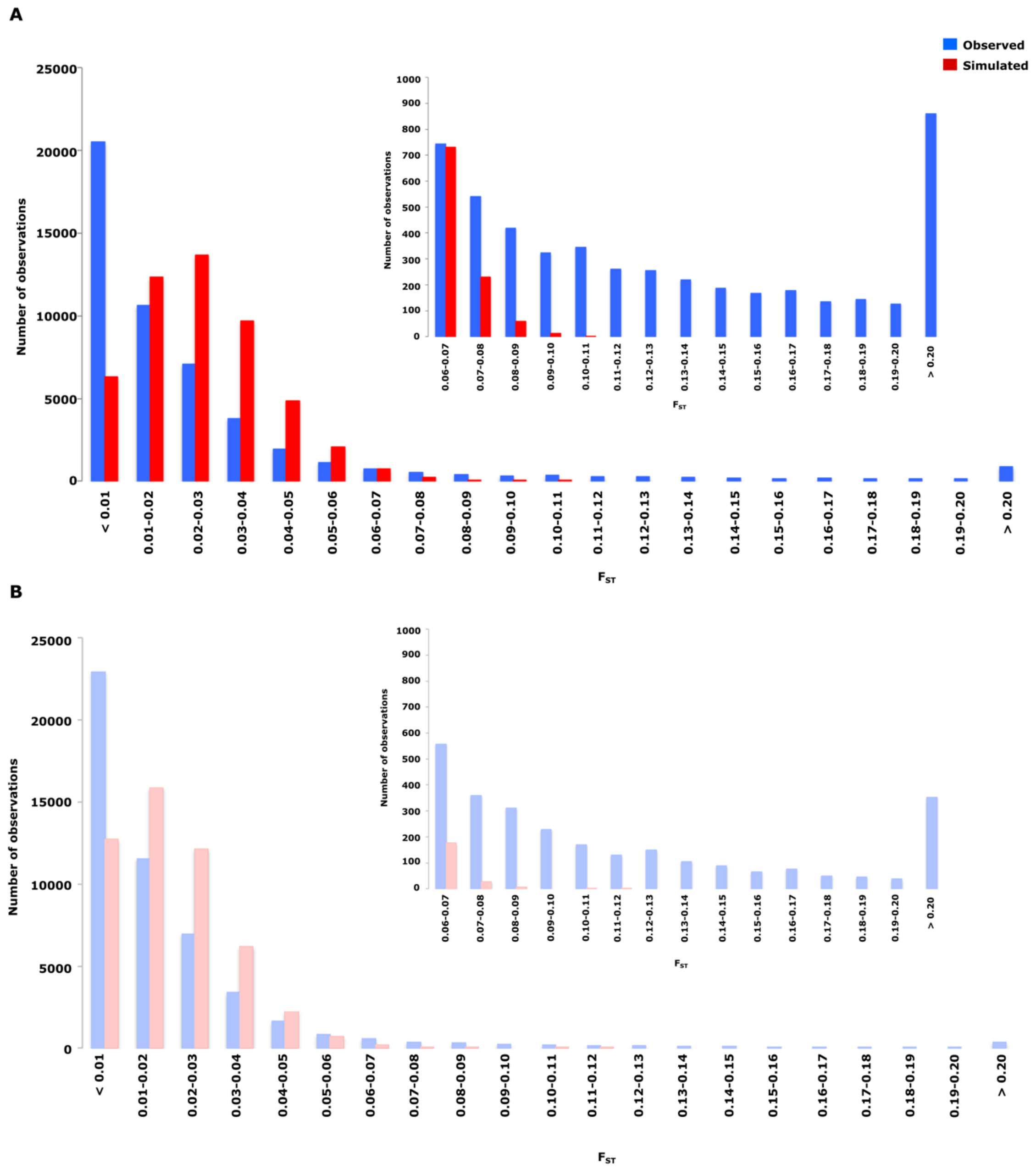


Figure 3.3 Comparison of genetic differentiation among herring populations against the expectation under genetic drift. The observed distribution is based on 50 000 randomly selected SNPs and the simulated data is based on the same number of selectively neutral loci. Histogram of F_{ST} values in the simulated and observed datasets among all 25 Atlantic herring populations from both sides of the Atlantic **(A)** and in 24 populations excluding the WFB population **(B)**. The right tail of the distribution is highlighted in the insets.

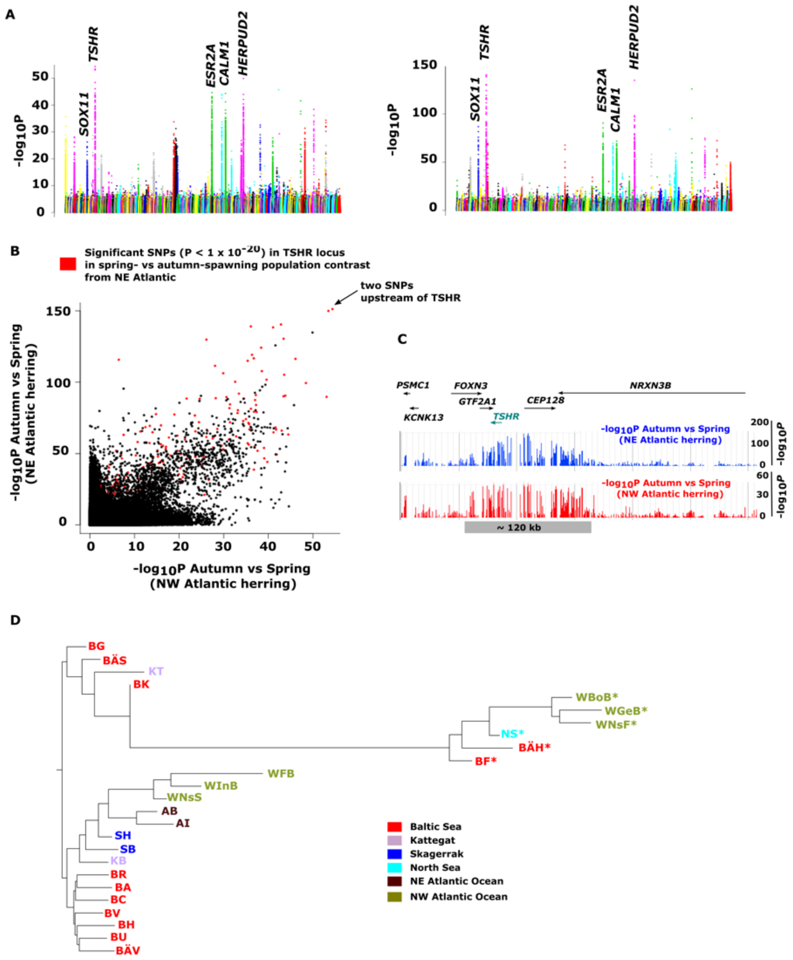


Figure 3.4 Genetic differentiation between autumn- and spring-spawning herring. (A) Manhattan plot of *P*-values for allele frequency differences between autumn- and spring-spawners in the NW Atlantic (upper panel) and in NE Atlantic/Baltic herring (lower panel). **(B)** Correlation plot of *P*-values for allele frequency differences between autumn- and spring-spawners in NW Atlantic and in NE Atlantic/Baltic herring. **(C)** Distribution of *P*-values around the 120 kb region harboring the *TSHR* locus **(D)** Neighbor-joining tree based on allele frequencies, determined by pooled sequencing, for all SNPs (n=1 313) in the *TSHR* region. Color codes for sampling locations are the same as in Fig. 3.1. **(E)** Phylogenetic trees for haplotypes at the *TSHR* locus determined using whole genome individual sequencing; A and B refer to the two haplotypes from the same individual.

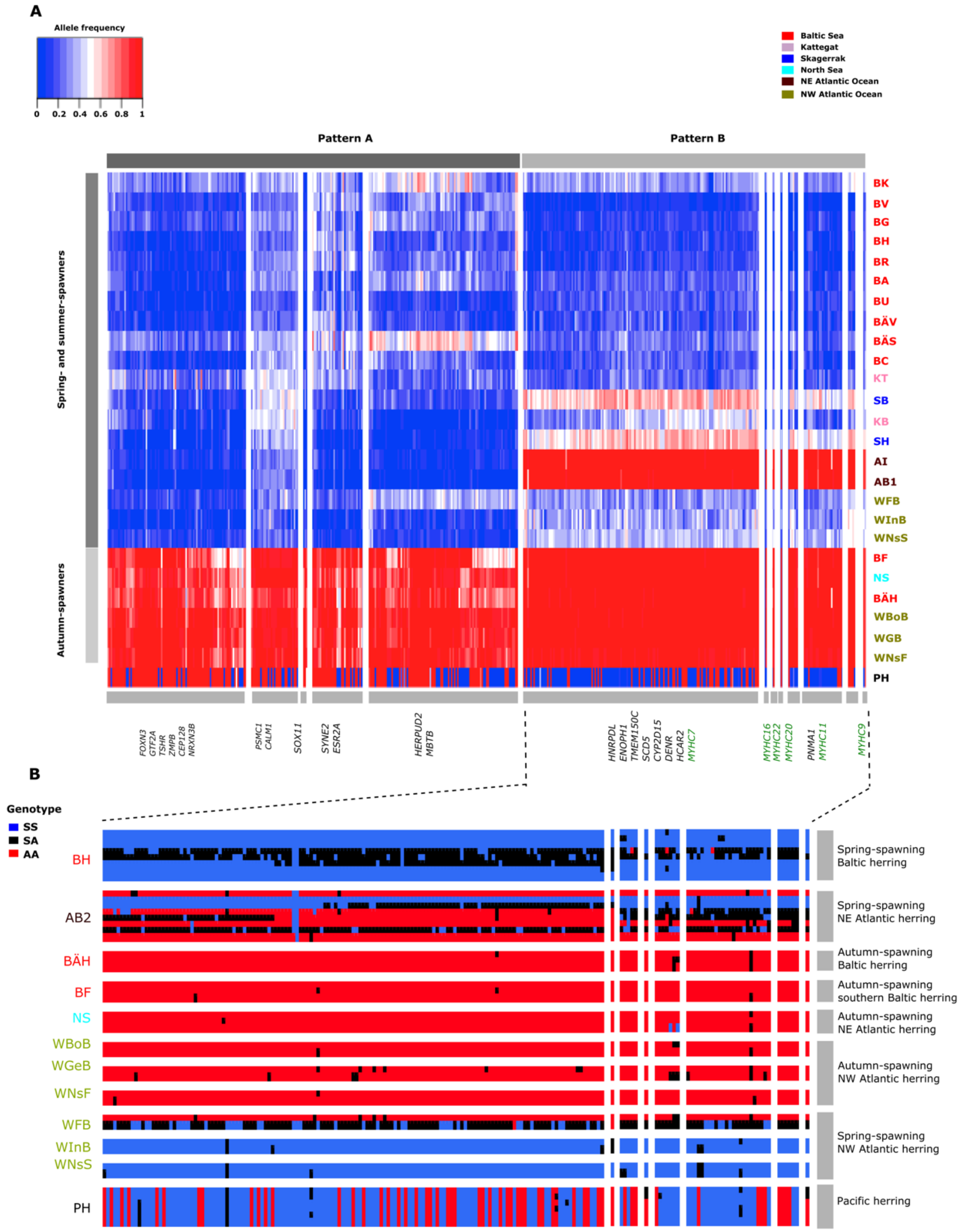


Figure 3.5 Loci showing highly significant allele frequency differences between spring- and autumn-spawning herring. (A) Heat map showing allele frequencies (estimated by pooled sequencing) in each of the 26 populations; color codes for each population are the same as in Fig. 3.1; genes overlapping these loci are listed at the bottom. **(B)** Haplotypes from individual herring samples at eight scaffolds showing pattern B (haplotype sharing between spring-spawning oceanic herring and autumn-spawning populations). The total number of SNPs used in these analyses is 189. Blocks of SNPs from different scaffolds are separated by blank lines.

3.12 Supporting Information

Table S3.1 Genomic regions that showed most consistent association with timing of spawning identified in our previous study (Martinez-Barrio et al, 2016). Loci that were not replicated in this study have their scaffold name, start and end base pair denoted in italic font. SNPs showing the strongest association with spawning time and that were common between NE and NW Atlantic populations are indicated with an asterisk.

Scaffold	**Start (bp)**	**End (bp)**	**Strongest SNP**[a]	**Strongest SNP $(-\log_{10}P)$**[a]	**Strongest SNP**[b]	**Strongest SNP $(-\log_{10}P)$**[b]	**Associated candidate genes**
scaffold139	*1,492,608*	*1,492,609*	1,492,609	14.89	1,492,609	1.70	*CPNE7*
scaffold1420	3,100	288,800	137,485*	151.19	137,485*	54.38	*PSMC1, KCNK13, FOXN3, GTF2A, TSHR, CEP128, NRXN3B*
scaffold1440	895,800	1,255,255	1,128,662*	26.98	1,128,662*	11.93	*SOX11, DCDC2, FEZ2, SMC6*
scaffold190	21,615	34,163	26,630	69.62	22,322	44.00	*CALM1*
scaffold312	2,556,980	2,701,506	2,642,301	66.54	2,637,207	42.87	*SYNE2*
scaffold312	2,704,099	2,885,098	2,744,359	90.42	2,761,114	44.61	*ESR2A*
scaffold46	*240,984*	*288,556*	240,985	22.37	240,985	0.84	*NELL1*
scaffold481	2,737,036	2,966,514	2,809,585*	134.74	2,809,585*	49.91	*HERPUD2, BAL1*

[a] Spring- vs. autumn-spawning populations in NE Atlantic, $-\log_{10}(P)$ values from standard chi-square tests to estimate the significance of allele frequency differences.

[b] Spring- vs. autumn-spawning populations in NW Atlantic, $-\log_{10}(P)$ values from standard chi-square tests to estimate the significance of allele frequency differences.

Table S3.2 List of loci showing strong genetic differentiation between spring- and autumn-spawning herring. An associated gene name is indicated if the SNP occurs within 5 Kb upstream or 5 Kb downstream of annotated genes. Loci significant in both east and west population are highlighted in green, the ones significant only in west Atlantic populations are highlighted in pink. **(electronic supplementary material)**

Table S3.3 Previously identified loci (Martinez Barrio et al., 2016) showing highly significant association to salinity. Loci with strong differentiation in Baltic vs NW Atlantic and non-significant differentiation in NE vs. NW Atlantic are highlighted in pink. *(-log10P, based on Chi-Square tests). **(electronic supplementary material)**

Table S3.4 Genetic distance matrix used for building the phylogenetic tree among 26 herring populations used for Figure 3.2. The details about sample ID are in Table 3.1. **(electronic supplementary material)**

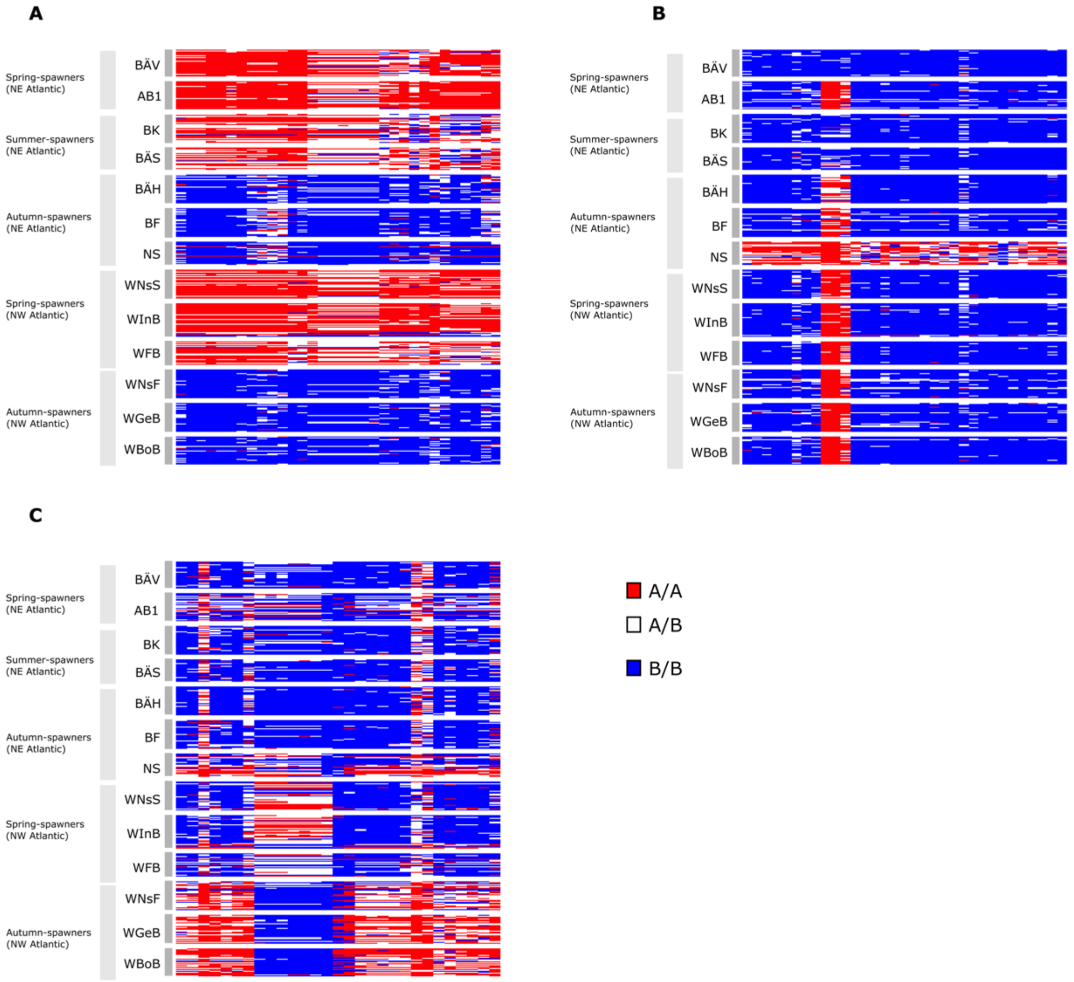


Figure S3.1 Individual genotype data for a subset of highly differentiated SNPs. (A) Highly differentiated SNPs between spring- and autumn-spawning population shared between NE and NW Atlantic. **(B)** Highly differentiated SNPs between spring- and autumn-spawning populations unique to the NE Atlantic. **(C)** Highly differentiated SNPs between spring- and autumn-spawning populations unique to the NW Atlantic.

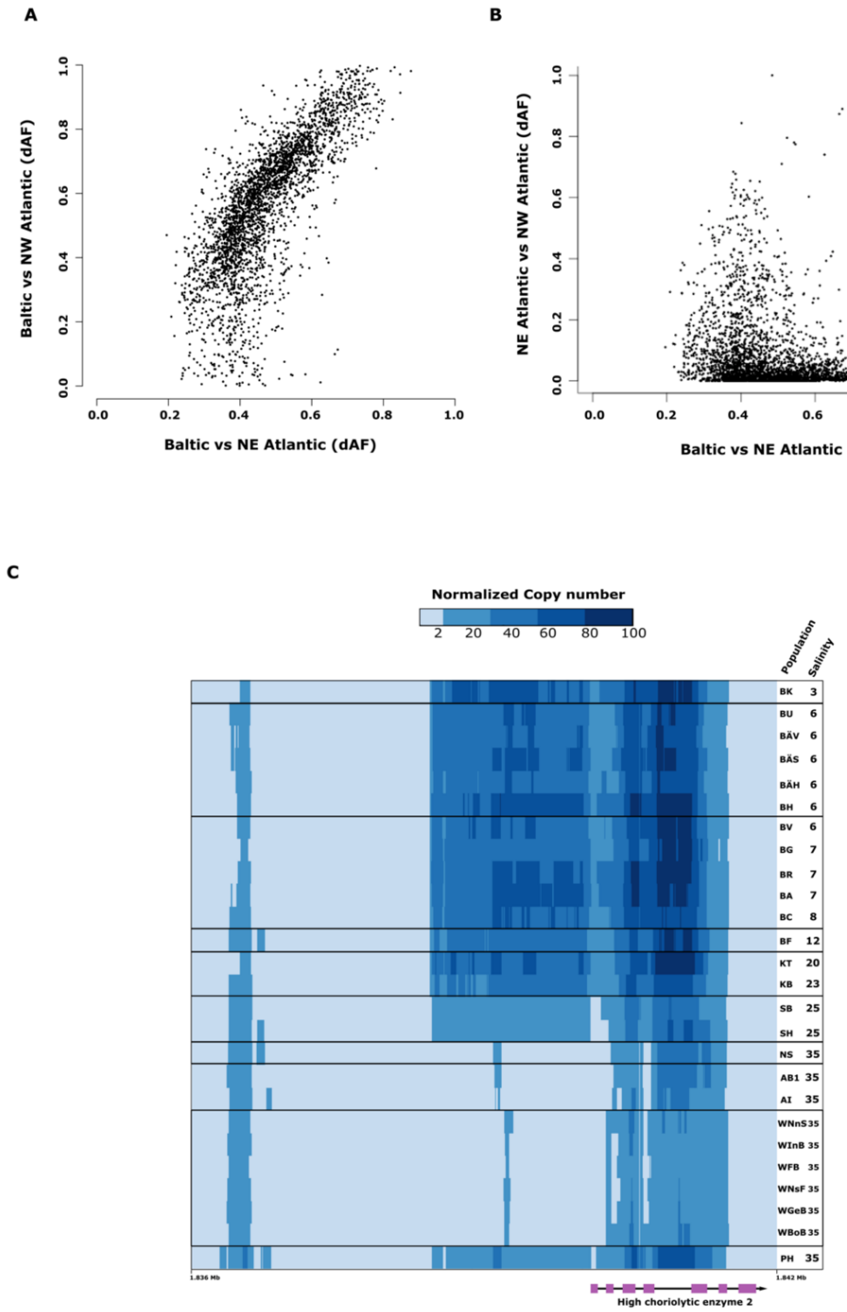
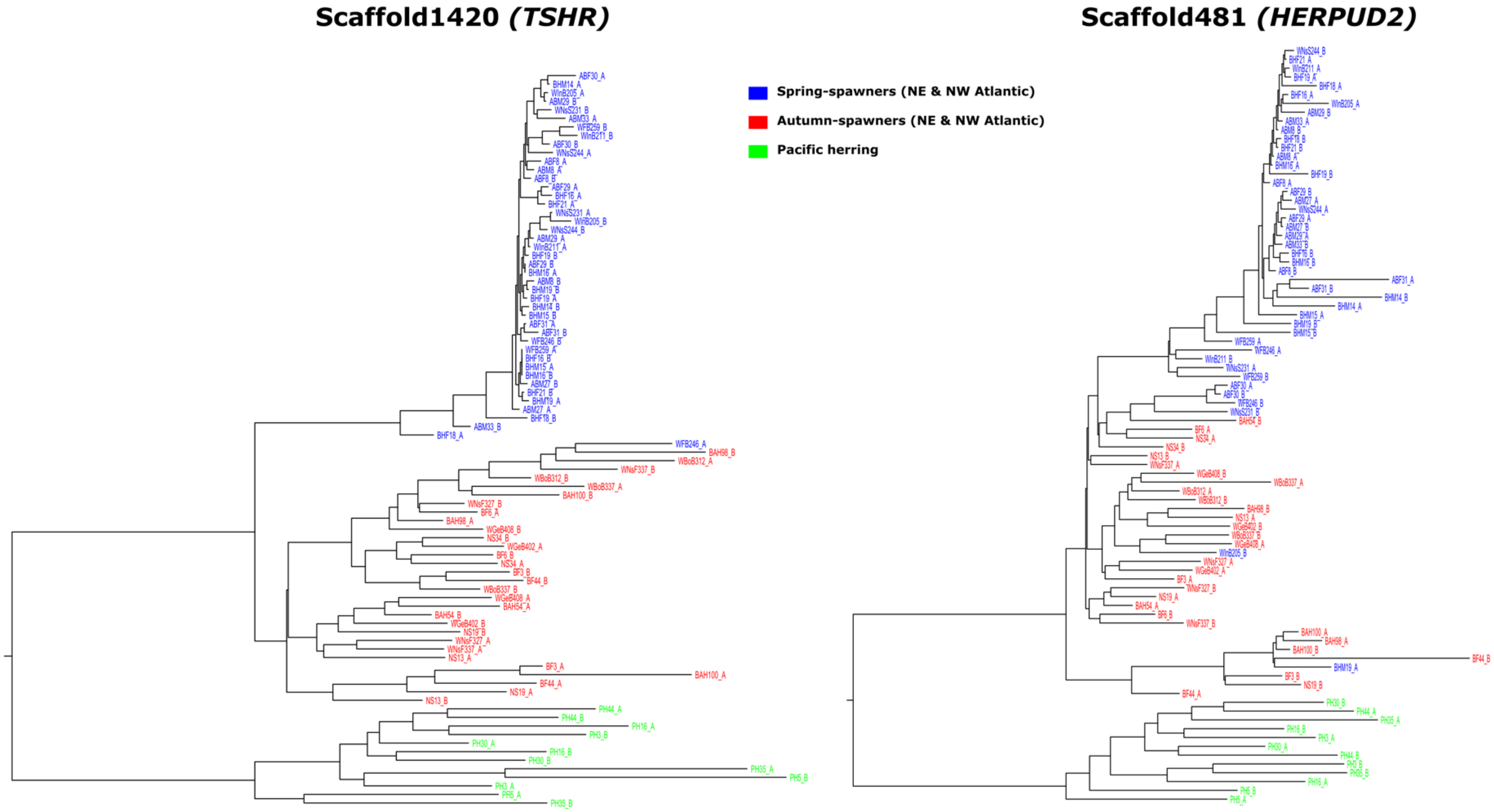
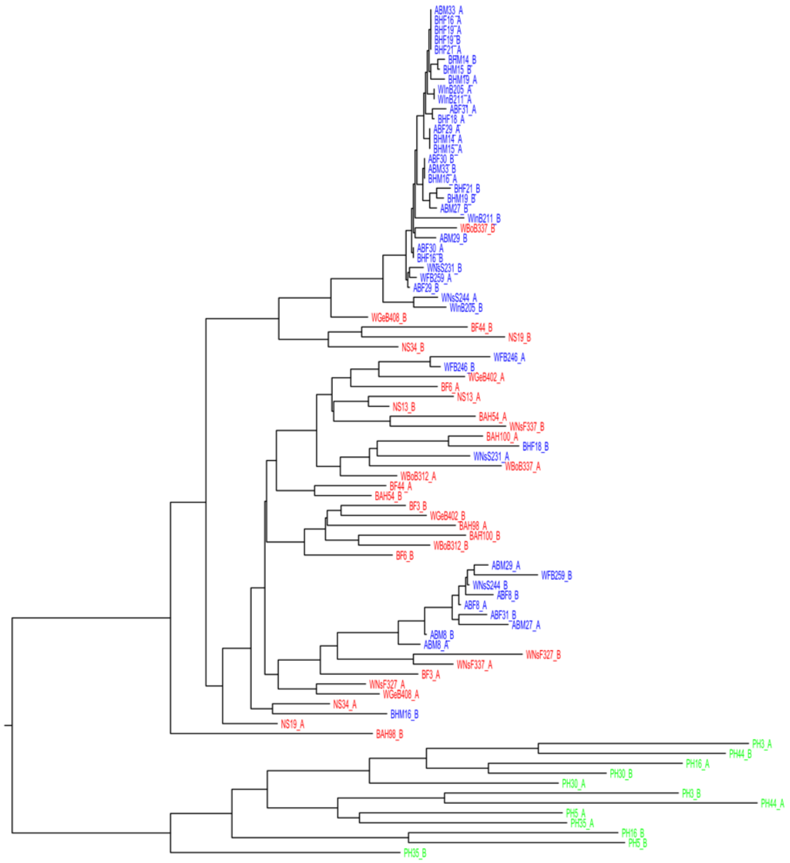


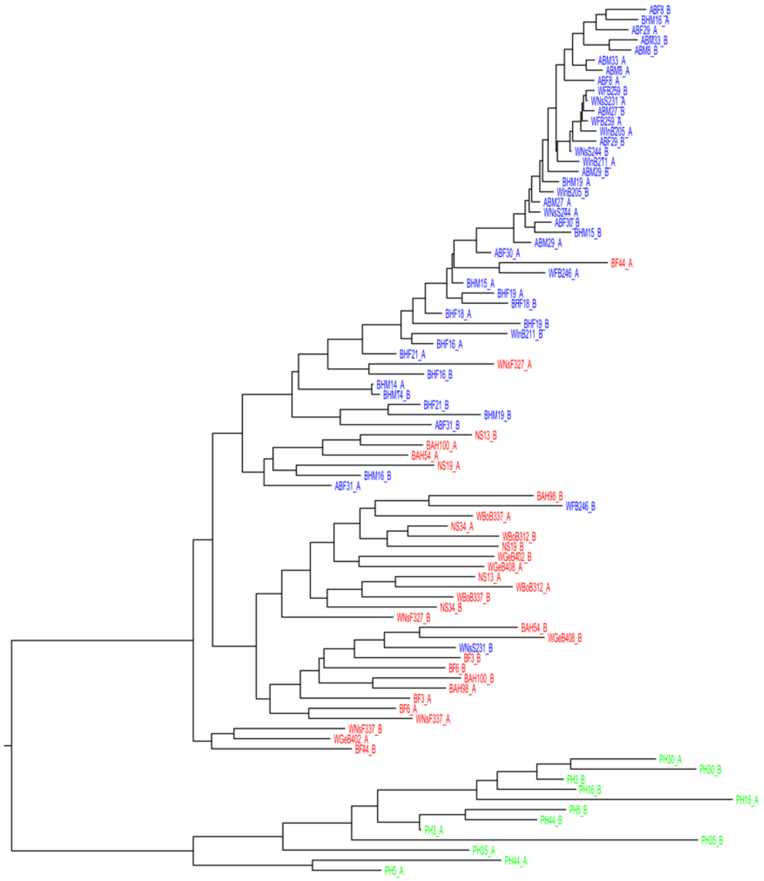
Figure S3.2 Loci associated with salinity identified in our previous study (Martinez Barrio et al., 2016)**. (A)** Comparison of delta allele frequencies (dAF) in Baltic Sea vs. NE Atlantic and Baltic Sea vs. NW Atlantic populations. **(B)** Comparison of delta allele frequencies (dAF) in Baltic Sea vs. NE Atlantic and NE Atlantic vs. NW Atlantic populations. **(C)** Heat map showing copy number variation partially overlapping the *HCE* gene. Orientation of transcription is marked with an arrow; population samples and salinity at sampling locations are indicated to the right; abbreviations are explained in Table 3.1. The figure is adapted from Martinez Barrio et al. (Martinez Barrio et al., 2016).


Scaffold1420 (TSHR)
Scaffold481 (HERPUD2)
Spring-spawners (NE & NW Atlantic)
Autumn-spawners (NE & NW Atlantic)
Pacific herring
ABF30_A
BHM14_A
WInB205_A
ABM29_B
WNsS231_B
ABM33_A
WFB259_B
WInB211_B
ABF30_B
WNsS244_A
ABF8_A
ABM8_A
ABF8_B
ABF29_A
BHF16_A
BHF21_A
WNsS231_A
WInB205_B
WNsS244_B
ABM29_A
WInB211_A
BHF19_B
ABF29_B
BHM16_A
ABM8_B
BHM19_B
BHF19_A
BHM14_B
BHM15_B
ABF31_A
ABF31_B
WFB246_B
WFB259_A
BHF16_B
BHM15_A
BHM16_B
ABM27_B
BHF21_B
BHM19_A
ABM27_A
BHF18_B
ABM33_B
BHF18_A
WFB246_A
BAH98_B
WBoB312_A
WNsF337_B
WBoB312_B
WBoB337_A
BAH100_B
WNsF327_B
BF6_A
BAH98_A
WGeB408_B
NS34_B
WGeB402_A
BF6_B
NS34_A
BF3_B
BF44_B
WBoB337_B
WGeB408_A
BAH54_A
BAH54_B
WGeB402_B
NS19_B
WNsF327_A
WNsF337_A
NS13_A
BF3_A
BAH100_A
BF44_A
NS19_A
NS13_B
PH44_A
PH44_B
PH16_A
PH3_B
PH30_A
PH16_B
PH30_B
PH35_A
PH5_B
PH3_A
PH5_A
PH35_B
WNsS244_B
BHF21_A
WInB211_A
BHF19_A
BHF18_A
BHF16_A
WInB205_A
ABM29_B
ABM33_A
ABM8_B
BHF18_B
BHF21_B
ABM8_A
BHM16_A
BHF19_B
ABF8_A
ABF29_B
ABM27_A
WNsS244_A
ABF29_A
ABM27_B
ABM29_A
ABM33_B
BHF16_B
BHM16_B
ABF8_B
ABF31_A
ABF31_B
BHM14_B
BHM14_A
BHM15_A
BHM19_B
BHM15_B
WFB259_A
WFB246_A
WInB211_B
WNsS231_A
WFB259_B
ABF30_A
ABF30_B
WFB246_B
WNsS231_B
BAH54_B
BF6_A
NS34_A
NS34_B
NS13_B
WNsF337_A
WGeB408_B
WBoB337_A
WBoB312_A
WBoB312_B
BAH98_B
NS13_A
WGeB402_B
WBoB337_B
WGeB408_A
WInB205_B
WNsF327_A
WGeB402_A
BF3_A
WNsF327_B
NS19_A
BAH54_A
BF6_B
WNsF337_B
BAH100_A
BAH98_A
BAH100_B
BF44_B
BHM19_A
BF3_B
NS19_B
BF44_A
PH30_B
PH44_A
PH35_A
PH16_B
PH3_A
PH30_A
PH44_B
PH3_B
PH35_B
PH16_A
PH5_B
PH5_A

Scaffold1440 *(SOX11)*



Scaffold312*(SYNE2/ESR2A)*



Scaffold90 *(CALM1)*

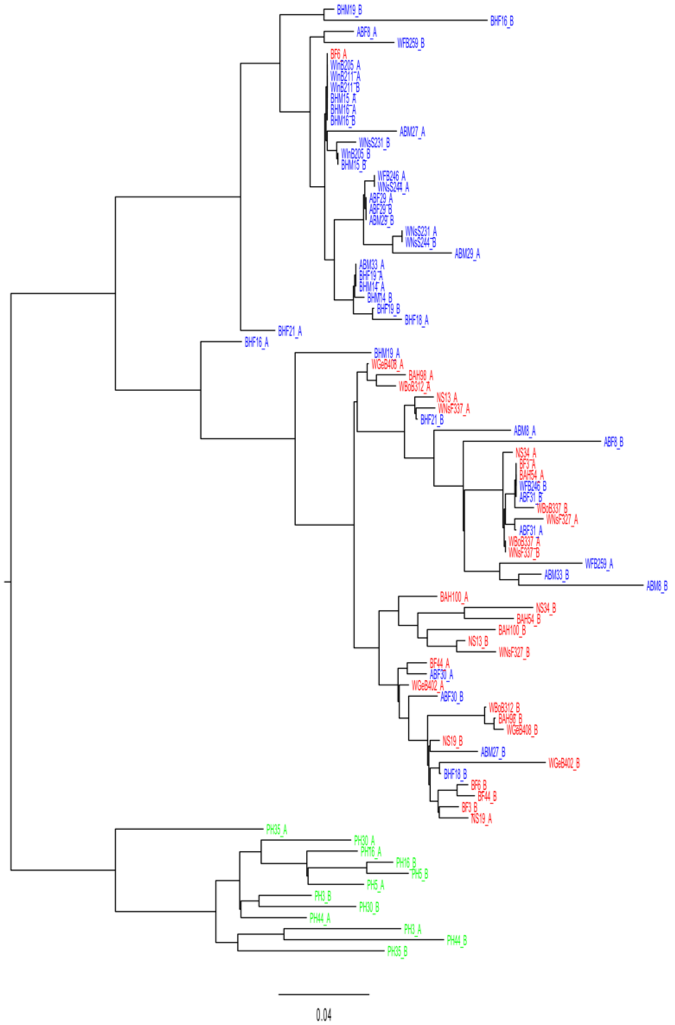


Figure S3.3 Haplotype trees for four loci showing strong differentiation between autumn- and spring-spawners across the North Atlantic Ocean and Baltic Sea. Haplotypes were deduced from individual whole genome resequencing; A and B refer to the two haplotypes from the same individual. Candidate genes at each locus are indicated.

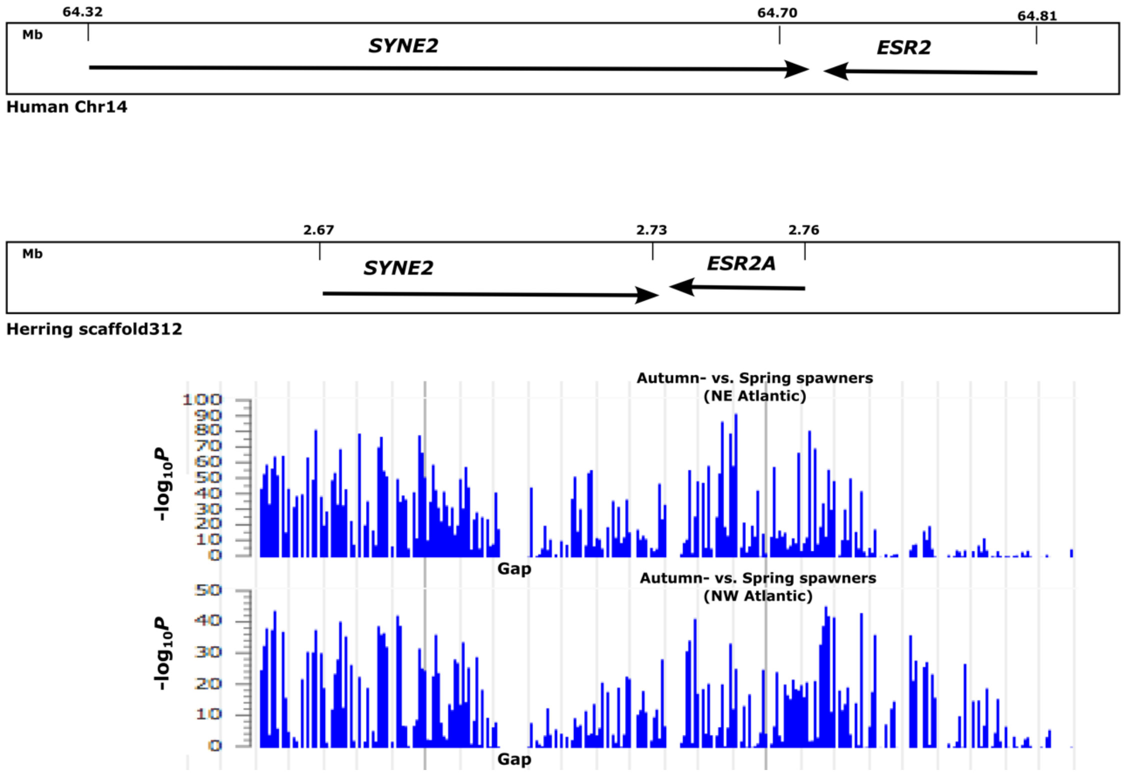


Figure S3.4 Organization of *SYNE2* and *ESR2* genes in human and Atlantic herring genomes. The genetic differentiation in this region between autumn- and spring-spawning herring populations in the NE and NW Atlantic is shown.

3.13 References

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

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bondesson, M., Hao, R., Lin, C.-Y., Williams, C., & Gustafsson, J.-Å. (2015). Estrogen receptor signaling during vertebrate development. *Biochim Biophys Acta*, *1849*(2), 142–151. https://doi.org/http://dx.doi.org/10.1016/j.bbagrm.2014.06.005

Bradbury, I. R., Hubert, S., Higgins, B., Borza, T., Bowman, S., Paterson, I. G., … Bentzen, P. (2010). Parallel adaptive evolution of Atlantic cod on both sides of the Atlantic Ocean in response to temperature. *Proceedings of the Royal Society of London B: Biological Sciences*, *277*(1701), 3725–3734. https://doi.org/10.1098/rspb.2010.0985

Browning, S. R., & Browning, B. L. (2016). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, *81*(5), 1084–1097. https://doi.org/10.1086/521987

Department of Fisheries & Oceans Canada. (2012). *Assessment of Atlantic herring in the southern Gulf of St. Lawrence (NAFO Div. 4T). Available online at http://www.dfo-mpo.gc.ca/csas-sccs/Publications/SAR-AS/2012/2012_014-eng.pdf.*

Department of Fisheries and Oceans Canada. (2011). Canadian fisheries statistics 2008, Available online at http://www.dfo-mpo.gc.ca/stats/commercial/cfs/2008/CFS2008_e.pdf.

DePristo, M. a, Banks, E., Poplin, R., Garimella, K. V, Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498. https://doi.org/10.1038/ng.806

Felsenstein, J. (1989). PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*, *5*, 164–166.

Gröger, J. P., Kruse, G. H., & Rohlf, N. (2009). Slave to the rhythm: how large-scale climate cycles trigger herring (Clupea harengus) regeneration in the North Sea. *ICES Journal of Marine Science: Journal Du Conseil* . https://doi.org/10.1093/icesjms/fsp259

Hanon, E. A., Lincoln, G. A., Fustin, J.-M., Dardente, H., Masson-Pévet, M., Morgan, P. J., & Hazlerigg, D. G. (2015). Ancestral TSH mechanism signals summer in a photoperiodic mammal. *Current Biology*, *18*(15), 1147–1152. https://doi.org/10.1016/j.cub.2008.06.076

Hoekstra, H. E., & Nachman, M. W. (2003). Different genes underlie adaptive melanism in different populations of rock pocket mice. *Molecular Ecology*, *12*(5), 1185–1194. https://doi.org/10.1046/j.1365-294X.2003.01788.x

Hohenlohe, P. a, Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. a, & Cresko, W. a. (2010). Population Genomics of Parallel Adaptation in Threespine Stickleback using Sequenced RAD Tags. *PLoS Genetics*, *6*(2), e1000862. https://doi.org/10.1371/journal.pgen.1000862

Iles, T. D., & Sinclair, M. (1982). Atlantic Herring: Stock Discreteness and Abundance. *Science*, *215*(4533), 627–633. https://doi.org/10.1126/science.215.4533.627

Johannessen, A., Skaret, G., Langard, L., Slotte, A., Husebo, A., & Ferno, A. (2014). The dynamics of a metapopulation: changes in life-history traits in resident herring that co-occur with oceanic herring during spawning. *PloS One*, *9*(7), e102462. https://doi.org/10.1371/journal.pone.0102462

Johnston, I. A., Vieira, V. L. A., & Temple, G. K. (2001). Functional consequences and population differences in the developmental plasticity of muscle to temperature in Atlantic herring Clupea harengus. *Marine Ecology Progress Series*, *213*, 285–300. https://doi.org/10.3354/meps213285

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., … Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, *484*(7392), 55–61. https://doi.org/10.1038/nature10944

Kim, H.-D., Choe, H. K., Chung, S., Kim, M., Seong, J. Y., Son, G. H., & Kim, K. (2011). Class-C SOX transcription factors control GnRH gene expression via the intronic transcriptional enhancer. *Molecular Endocrinology*, *25*(7), 1184–1196. https://doi.org/10.1210/me.2010-0332

King, T. L., Kalinowski, S. T., Schill, W. B., Spidle, A. P., & Lubinski, B. A. (2001). Population structure of Atlantic salmon (Salmo salar L.): a range-wide perspective from microsatellite DNA variation. *Molecular Ecology*, *10*(4), 807–821.

Lamichhaney, S., Barrio, A. M., Rafati, N., Sundstrom, G., Rubin, C.-J., Gilbert, E. R., … Andersson, L. (2012). Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. *Proceedings of the National Academy of Sciences*, *109*(47), 19345–19350. https://doi.org/10.1073/pnas.1216128109

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Luo, Y., Widmer, A., & Karrenberg, S. (2015). The roles of genetic drift and natural selection in quantitative trait divergence along an altitudinal gradient in Arabidopsis thaliana. *Heredity*, *114*, 220–228.

Martinez Barrio, A., Lamichhaney, S., Fan, G., Rafati, N., Pettersson, M., Zhang, H., … Andersson, L. (2016). The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. *ELife*, *5*(MAY2016), 1–32. https://doi.org/10.7554/eLife.12081

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

McPherson, A., O'Reilly, P. T., & Taggart, C. T. (2004). Genetic Differentiation, Temporal Stability, and the Absence of Isolation by Distance among Atlantic Herring Populations. *Transactions of the American Fisheries Society*, *133*(2), 434–446. https://doi.org/10.1577/02-106

McPherson, A., Stephenson, R. L., O'Reilly, P. T., Jones, M. W., & Taggart, C. T. (2001). Genetic diversity of coastal Northwest Atlantic herring populations: implications for management. *Journal of Fish Biology*, *59*(SUPPL. A), 356–370. https://doi.org/10.1006/jfbi.2001.1769

McQuinn, I. H. (1997). Metapopulations and the Atlantic herring. *Reviews in Fish Biology and Fisheries*, *7*(3), 297–329. https://doi.org/10.1023/A:1018491828875

Melamed, P., Savulescu, D., Lim, S., Wijeweera, A., Luo, Z., Luo, M., & Pnueli, L. (2012). Gonadotrophin-Releasing Hormone signalling downstream of Calmodulin. *Journal of Neuroendocrinology*, *24*(12), 1463–1475. https://doi.org/10.1111/j.1365-2826.2012.02359.x

Miller, M. R., Brunelli, J. P., Wheeler, P. A., Liu, S., Rexroad, C. E., Palti, Y., … Thorgaard, G. H. (2012). A conserved haplotype controls parallel adaptation in geographically distant salmonid populations. *Molecular Ecology*, *21*(2), 237–249. https://doi.org/10.1111/j.1365-294X.2011.05305.x

Nakao, N., Ono, H., Yamamura, T., Anraku, T., Takagi, T., Higashi, K., … Yoshimura, T. (2008). Thyrotrophin in the pars tuberalis triggers photoperiodic response. *Nature*, *452*(7185), 317–322.

Nei, M. (1987). *Molecular Evolutionary Genetics*. New York: Columbia Univ Press.

Ono, H., Hoshino, Y., Yasuo, S., Watanabe, M., Nakane, Y., Murai, A., … Yoshimura, T. (2008). Involvement of thyrotropin in photoperiodic signal transduction in mice. *Proceedings of the National Academy of Sciences, USA*, *105*(47), 18238–18242. https://doi.org/10.1073/pnas.0808952105

Pearse, D E, & Pogson, G. H. (2000). Parallel evolution of the melanic form of the California legless lizard, Anniella pulchra, inferred from mitochondrial DNA sequence variation. *Evolution; International Journal of Organic Evolution*, *54*(3), 1041–1046.

Pearse, Devon E, Miller, M. R., Abadia-Cardoso, A., & Garza, J. C. (2014). Rapid parallel evolution of standing variation in a single, complex, genomic region is associated with life history in steelhead/rainbow trout. *Proceedings. Biological Sciences / The Royal Society*, *281*(1783), 20140012. https://doi.org/10.1098/rspb.2014.0012

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE*, *5*(3), e9490.

Pritchard, J. K., & Di Rienzo, A. (2010). Adaptation - not by sweeps alone. *Nature Reviews. Genetics*, *11*(10), 665–667. https://doi.org/10.1038/nrg2880

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and

Population-Based Linkage Analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Reid, R. N., Cargnelli, L. M., Griesbach, S. J., Packer, D. B., Johnson, D. L., Zetlin, C., … Berrien, P. L. (1999). Atlantic Herring, Clupea harengus, Life History and Habitat Characteristics, Available online at http://www.nefsc.noaa.gov/publications/tm/tm126/tm126.pdf.

Rubin, C.-J., Zody, M. C., Eriksson, J., Meadows, J. R. S., Sherwood, E., Webster, M. T., … Andersson, L. (2010). Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature*, *464*(7288), 587–591. https://doi.org/10.1038/nature08832

Ryman, N., & Palm, S. (2006). POWSIM: a computer program for assessing statistical power when testing for genetic differentiation. *Molecular Ecology Notes*, *6*(3), 600–602. https://doi.org/10.1111/j.1471-8286.2006.01378.x

Savolainen, O., Lascoux, M., & Merilä, J. (2013). Ecological genomics of local adaptation. *Nature Reviews. Genetics*, *14*(11), 807–820. https://doi.org/10.1038/nrg3522

Schlötterer, C., Tobler, R., Kofler, R., & Nolte, V. (2014). Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. *Nature Reviews Genetics*, *15*(11), 749–763. https://doi.org/10.1038/nrg3803

Sinclair, M., & Tremblay, M. J. (1984). Timing of Spawning of Atlantic Herring (Clupea harengus harengus) Populations and the Match–Mismatch Theory. *Canadian Journal of Fisheries and Aquatic Sciences*, *41*(7), 1055–1065. https://doi.org/10.1139/f84-123

Stahl, G. (1983). Differences in the amount and distribution of genetic variation between natural populations and hatchery stocks of Atlantic salmon. *Aquaculture*, *33*(1–4), 23–32. https://doi.org/http://dx.doi.org/10.1016/0044-8486(83)90383-6

Stephenson, R. L., Melvin, G. D., & Power, M. J. (2009). Population integrity and connectivity in Northwest Atlantic herring: a review of assumptions and evidence. *ICES Journal of Marine Science*, *66*(8), 1733–1739. https://doi.org/10.1093/icesjms/fsp189

Stern, D. L. (2013). The genetic causes of convergent evolution. *Nat Rev Genet*, *14*(11), 751–764.

Temple, G. K., Cole, N. J., & Johnston, I. A. (2001). Embryonic temperature and the relative timing of muscle-specific genes during development in herring (Clupea harengus L.). *Journal of Experimental Biology*, *204*(21), 3629–3637.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., … DePristo, M. A. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. In *Current Protocols in Bioinformatics* (Vol. 11, pp. 11.10.1-11.10.33). https://doi.org/10.1002/0471250953.bi1110s43

Watabe, S. (1999). Myogenic regulatory factors and muscle differentiation during ontogeny in fish. *Journal of Fish Biology*, *55*(A), 1–18. https://doi.org/10.1111/j.1095-8649.1999.tb01042.x

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, *38*(6), 1358. https://doi.org/10.2307/2408641

CHAPTER 4. ADAPTATION TO SEASONAL REPRODUCTION AND THERMAL MINIMA-RELATED FACTORS DRIVES FINE-SCALE DIVERGENCE DESPITE GENE FLOW IN ATLANTIC HERRING POPULATIONS

4.1 Abstract

High connectivity and low potential for local adaptation have been common assumptions for most marine species, given their usual high fecundity and dispersal capabilities. Recent genomic studies, however, have disclosed unprecedented levels of population subdivision in what were previously presumed to be panmictic or nearly panmictic species. Here we analyzed neutral and putatively adaptive genetic variation at the whole-genome level in Atlantic herring (*Clupea harengus* L.) spawning aggregations distributed across the reproductive range of the species in North America. We uncovered genetic population structure at putatively adaptive loci, against the backdrop of low genetic differentiation at neutral loci. Our results revealed an intricate pattern of population subdivision associated with two overlapping axes of divergence: a temporal axis determined by seasonal reproduction, and a spatial axis defined by a latitudinal cline establishing a steep north-south genetic break. Genetic-environment association analyses indicated that winter sea-surface temperature is the best predictor of the spatial structure observed (explained 58.1% of the total genetic variance). Thousands of outlier SNPs distributed along specific parts of the genome spanning numerous candidate genes underlined each pattern of differentiation (seasonal reproduction: 14 724, latitudinal cline: 6 595), forming so-called “genomic regions or islands of divergence”. Our results indicate that timing of reproduction and latitudinal spawning location are features under natural selection leading to local adaptation in the herring. Our study highlights the importance of preserving adaptive and neutral intraspecific diversity, and the utility of an integrative seascape genomics approach for disentangling intricate patterns of intraspecific diversity in highly dispersive and abundant marine species.

4.2 Introduction

Population subdivision and connectivity are important topics in evolutionary and conservation biology, because they can help elucidate how local adaptation arises (Barrett

& Hoekstra, 2011; Lewontin, 2002) and can guide management plans aiming to protect intraspecific genetic diversity, a determinant factor for population persistence and adaptive divergence in changing environments (Allendorf et al., 2010). Yet, the difficulty in determining the relative importance of genetic drift, gene flow, and selection in shaping contemporary patterns of intraspecific genetic diversity, remains a major challenge (Ravinet et al., 2017). The increased power for assessing neutral and putatively adaptive genetic variation with next-generation sequencing (NGS) technologies (Nosil & Feder, 2012) is helping to uncover unprecedented levels of genetic structure in what were previously presumed to be panmictic or nearly panmictic species.

Marine species are outstanding examples of such paradigm shifts, as they have often been expected and observed to exhibit low levels of population structure and low divergence potential (Palumbi, 1994), given their high fecundity and dispersal capabilities (Hauser & Carvalho, 2008). Recent genomic studies revealing fine-scale structuring are challenging this view [e.g., Atlantic cod (*Gadus morhua*) (Bradbury et al., 2013); Atlantic herring (*Clupea harengus*) (Martinez Barrio et al., 2016); American lobster (*Homarus americanus*) (Benestan et al., 2015)]. Various mechanisms by which population structure could arise have been proposed, including: oceanographic barriers, isolation-by-distance, larval and adult behavior, recent evolutionary history (e.g. historical vicariance and secondary contact), and natural selection (Palumbi, 1994). There is great interest in understanding how natural selection can lead to population divergence and local adaptation, especially under the homogenizing effect of gene flow (Tigano & Friesen, 2016) because of its direct relationship with fitness, effective population size, population persistence, and evolution. However, the genetic basis of adaptive traits remains largely unknown (Barrett & Hoekstra, 2011). Genome scans performed with NGS methods are helping to identifying loci associated with adaptive phenotypes (Jones et al., 2012; Tavares et al., 2018). Such loci typically show elevated genetic divergence that is interpreted as a signature of selection. Nevertheless, disentangling genomic signatures of selection from signatures of demographic history has been limiting (Hoban et al., 2016). Species that are widely distributed are often exposed to diverse ecological habitats where selection can result in local adaptation (Yeaman & Whitlock, 2011).

Therefore, highly fecund marine species inhabiting heterogeneous environments offer ideal candidates for the study of ecological adaptation, since in these the effect of genetic drift is minuscule and the effectiveness of natural selection is greater.

Atlantic herring is an abundant marine schooling pelagic fish that has colonized diverse environments throughout the North Atlantic, including open ocean and the brackish waters of the Baltic Sea. These characteristics, together with the increasing availability of genomic resources, make this species ideal for investigating the genetic basis and mechanisms involved in ecological adaptation. Juveniles and adults undertake annual migrations between feeding, overwintering, and spawning areas. Herring matures at 3-4 years of age and can live to 20+ years (Benoît et al., 2018). Spawning occurs mostly in spring and fall seasons at predictable times and locations near shore, which suggests strong spawning site fidelity (McQuinn, 1997; Stephenson et al., 2009; Wheeler & Winters, 1984a). Atlantic herring plays an important role in the marine ecosystem, feeding on plankton and being preyed upon by numerous marine fish, birds and mammals. It also sustains large fisheries throughout the North Atlantic (FAO, 2019), some of which have experienced severe periods of decline and signs of recovery in the last century (Britten, Dowd, & Worm, 2016; Engelhard & Heino, 2004; Overholtz, 2002; Simmonds, 2007). The ecological, economic, and cultural importance of herring has therefore motivated research on this species for more than a century (Stephenson et al., 2009); however, its complex life history has made the description of its population structure elusive (Iles & Sinclair, 1982).

Numerous studies have examined the population structure of herring using different genetic tools and at various spatial scales, mostly in the northeast (NE) Atlantic. Such studies have observed low levels of population differentiation at neutral loci (e.g. Andersson et al., 1981; André et al., 2011; Jorgensen et al., 2005). The expansion of these studies to the use of thousands of single nucleotide polymorphisms (SNPs) derived from various genomic techniques have revealed significant genetic differentiation at putatively adaptive loci in relation to environmental gradients (Guo et al., 2016; Lamichhaney et al., 2012; Limborg et al., 2012). Moreover, the recent development of a high-quality genome assembly for the Atlantic herring allowed the identification of many millions of SNPs and a breakthrough in the possibility to study the genetic basis of ecological adaptation in this species (Martinez Barrio et al., 2016). A few studies have addressed this question in the northwest (NW) Atlantic (Kerr, Fuentes-Pardo, Kho, McDermid, & Ruzzante, 2019; Lamichhaney et al., 2017; McPherson et al., 2004); while they provided important insight on population structuring with seasonal reproduction and within the southern region, and reported temporal stability of genomic divergence between spring and fall spawners, they were limited by scarce sampling.

In the NW Atlantic, herring spawn from Cape Cod to southern Labrador (Bourne, Mowbray, Squires, & Koen-Alonso, 2018; Sinclair & Iles, 1989) between April and November, but spawning peaks in spring and fall. Spring- and fall-spawners are therefore the main spawning types recognized in the region. The relative abundance of each reproductive strategy varies geographically: in the north (northern Newfoundland) spring-spawners were historically more abundant, at mid-range (Gulf of St. Lawrence) both strategies were common, and in the southern extreme (Bay of Fundy, Scotian Shelf, Gulf of Maine) fall-spawners predominate (Melvin et al., 2009). Changes in the prevalence of these components have been observed in the last decade; in particular, a significant decline of spring-spawners and a moderate abundance of fall-spawners in the Gulf of St. Lawrence (McDermid, Swain, Turcotte, Robichaud, & Surette, 2018) and Newfoundland (Bourne et al., 2018). Such changes have been attributed to varying elevated fishing mortality, declines in weight-at-age, and environmental conditions (Melvin et al., 2009), suggesting that the effects of climate change on population persistence of Atlantic herring are important. The concerning population declines (Britten et al., 2016) emphasize the need to disentangle the population structure of NW Atlantic herring.

Here, we study neutral and adaptive variation of adult herring collected from 14 spawning grounds distributed across the species' reproductive range in the NW Atlantic. The two overarching questions were: *i*) What are the spatial scale and pattern of population structuring in herring and what is the genetic basis of such structuring, and *ii*)

What is the potential functional effect of variant sites underlying population divergence and which mechanisms and environmental variables are associated with population structure patterns? We used whole-genome re-sequencing of pools of individuals [Pool-seq, (Schlötterer et al., 2014)] and individual genotyping along with multivariate statistical approaches, machine learning algorithms, and oceanographic information, to address these questions. Considering the particular attributes of the NW Atlantic Ocean (DFO, 1997; Townsend, Thomas, Mayer, Thomas, & Quinlan, 2004) and the importance of environment for shaping population divergence in herring, we predict that some of the divergent genomic regions exclusively found in Canada may be strongly associated with local environmental conditions. Our results provide insight into how population divergence arises in the presence of gene flow via temporal and spatial isolation and will help inform management and conservation practices.

4.3 Materials and Methods

4.3.1 Sample collection and DNA extraction

Adult herring were collected from 14 inshore spawning aggregations distributed across Atlantic Canada and the Gulf of Maine (N per aggregate = 48-50, total of 697 individuals), which represent most of the reproductive range of the species in the NW Atlantic (Fig. 4.1A and Table 4.1). Sampling took place during the local spawning peak in the spring and fall seasons from 2012 to 2016. Spawning sites correspond to areas with recurrent annual spawning.. Because of the presumed spawning site fidelity and the mixing of populations during the non-spawning seasons, we targeted individuals in reproductive condition to assess population definition. Individual muscle or fin tissue samples were preserved in 95% ethanol at -20 ºC until processing. DNA was isolated from the tissue samples using a standard phenol chloroform protocol (Sambrook & Russel 2006). DNA concentration (in ng/µl) was measured in triplicates using the Quant-iT PicoGreen dsDNA assay (Thermo Fisher Scientific, U.S.) and the Roche LightCycler 480 Instrument (Roche Molecular Systems, Inc., Germany). DNA integrity was verified with 0.8% agarose gel electrophoresis using 0.5x TBE buffer and a 1Kb molecular weight ladder.

4.3.2 Pool-sequencing and read quality filtering

Genome-wide patterns of genetic variation and population allele frequencies were assessed for each spawning aggregation using the Pool-seq approach. This method consists of performing whole-genome sequencing of pools of individuals using a single barcoded library, which implies that only population level data is recovered (individual genotype information is lost). In our case, each pool comprised equal amounts of DNA of ~50 individuals collected on the same spawning ground (the terms spawning aggregation and sampling site will be interchangeably used hereafter). We aimed to include in the same pool only DNA from "ready-to-spawn" and "actively spawning" individuals [gonadal maturity stage 5 and 6, respectively, (Bucholtz, Tomkiewicz, & Dalskov, 2008)]. Yet, in some spawning aggregations (BDO-S, NDB-S, NDB-F, TRB-F, and ME4-F, see pie charts in Fig. 4.1A) 25-50% of individuals were in "maturing" (stage 4) or "resting" (stage 8) condition at the time of sampling. The designation of "S" or "F" in the location name thus only reflects the season of collection and not necessarily the actual spawning season of all fish included in the pool. Individual DNA were normalized to a common concentration and pooled to a single tube using the liquid handling robot epmotion 5407 (Eppendorf, Germany). Sequencing library preparation and shotgun sequencing were outsourced. In brief, a single TruSeq Nano Illumina DNA library was built for each DNA pool (i.e. spawning aggregation). AMPURE beads were used for fragment size selection, targeting an insert size of ~550 bp. The 14 pooled-DNA libraries were sequenced using paired-end 126-bp reads on an Illumina Hiseq-2500 sequencer in two batches (5 libraries in 2015, 11 in 2016). Target read depth of coverage per pool was 40-50x, for an estimated herring genome size of ~850 Mb (Martinez Barrio et al., 2016).

Quality of raw sequence reads of each pool was checked using FastQC v0.11.5 (Simon Andrews, 2010), and jointly evaluated for the 14 pools with MultiQC v.1.3 (Ewels, Magnusson, Lundin, & Käller, 2016). Low quality bases (Phred score <20) and Illumina adapters were trimmed-off the reads, and reads shorter than 40 bp were removed from the dataset using Trimmomatic v.0.36 (Bolger et al., 2014) [*parameters:*

ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10 SLIDINGWINDOW:5:20 MINLEN:40]. High quality paired-reads remaining after filtering were used for downstream analysis.

4.3.3 Read mapping, SNP calling and filtering

We adapted the Genome Analysis Toolkit (GATK) Best Practices workflow (Van der Auwera et al., 2013) to variant discovery in Pool-seq data and to our computing infrastructure. For this we first obtained a stitched version of the herring genome for optimal SNP caller performance in the computer cluster available. Then, sequence reads of each pool were independently aligned against the stitched herring genome using the Burrows-Wheeler Aligner (BWA) v0.7.12-r1039 [*default parameters, MEM algorithm*] (Li, 2013). SNP calling was performed using GATK v3.8 (McKenna et al., 2010) (see Fig. S4.1). Lastly, the raw variant calls were filtered using GATK (Fig. S4.2), Popoolation2, and custom python scripts (See Supporting Information for details). In Pool-seq applications, population allele frequencies are derived from the total read counts supporting a variant site. Read coverage though, can be biased by sequencing and read mapping artifacts (Dohm, Lottaz, Borodina, & Himmelbauer, 2008; Kolaczkowski, Kern, Holloway, & Begun, 2011). To control for these factors and minimize their potential effect on population allele frequency calculation, we applied the allele count correction proposed by (Feder, Petrov, & Bergland, 2012; Kolaczkowski et al., 2011). Details on the application of this correction method and population allele frequencies estimation can be found in the Supporting Information.

4.3.4 Population structure

Based on the population allele frequencies, we examined genetic structure among spawning aggregations with a Neighbor-Joining (NJ) tree and with pairwise F_{ST} estimates. We computed pairwise Nei (1972) genetic distance with *Gendist* and built a NJ tree with *Neighbor*, both programs implemented in the package PHYLIP v3.697 (Baum, 1989). Bootstrapping was performed using the program *Seqboot* of PHYLIP, and the consensus tree was visualized with FigTree (Rambaut, 2007). We estimated unbiased F_{ST} for pools ($\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$) between all pairs of spawning aggregations using the R package *poolfstat* (Hivert, 2018). This algorithm computes F-statistics equivalent to Weir &

Cockerham (1984) estimates, while accounting for random sampling of chromosomes that may occur during DNA pooling and sequencing in Pool-seq applications.

4.3.5 Outlier loci detection and genome-wide patterns of differentiation

To identify loci potentially under selection, we performed genome scans for outlier loci detection using Principal Component Analysis (PCA), as implemented in the R package *pcadapt* v.4.0.2 (Luu, Bazin, & Blum, 2017). This algorithm assumes that divergent loci highly correlated to population structure are likely under selection. Outlier loci are detected based on the Mahalanobis distance calculated from the correlation coefficients between SNPs and a selected number of *K* principal components (PCs) (i.e. PCA loadings).

We performed a genome scan for the first 13 PCs (default is *K*=number of pools-1, 14-1= 13) using a minor allele frequency (MAF) of 0.05. Loci with Benjamini-Hochberg (BH) adjusted *P*-values ≤0.01 were considered candidates for being under selection. To identify which PCs explained the greatest proportion of genomic variance, we examined the scree plot generated by *pcadapt*, as well as the allele frequency patterns revealed in heatmaps made with the R package *ComplexHeatmap* (Gu, Eils, & Schlesner, 2016). The heatmaps depicted population allele frequencies (standardized to the major allele) of the 200 outlier loci most correlated to each PC (ranked by *P*-value in ascending order). We further explored the loci driving genomic differentiation in the herring by performing, with *pcadapt*, component-wise genome scans for the PCs exhibiting distinctive allele frequency patterns. To examine the distribution of outlier loci across the herring genome, for each informative PC we obtained Manhattan plots depicting the genomic position of outlier SNPs and their respective significance association value (–$\log_{10}P$-value) using the R package *qqman* (Turner, 2014).

4.3.6 Identification of the most informative outlier loci

We ranked outlier loci based on their importance for classification to each of the categories (or classes) of distinctive genomic patterns of differentiation in herring. For this we used random forest (RF), a supervised learning algorithm implemented in the R package *randomForest* (Liaw & Wiener, 2002). For the seasonal reproductive pattern, classes corresponded to spring or fall. For the latitudinal pattern, classes were northern (SIL-S, SPH-S, NTS-S, LAB-F, BLS-F, NDB-S, NDB-F, TRB-F, MIR-F, BDO-S, SCB-F), intermediate (MUS-F, GEB-F), and southern (ME4-F) regions. The RF model was based on 50 individual genotypes per spawning aggregation simulated from population allele frequencies using the R function *sample.geno* implemented in *pcadapt* v3.0.4. For the RF runs, the parameter *mtry* was set to default (equals to sqrt(*p*), where *p* is the number of loci); *ntree* was set to 1,000,000; and *sampsize* was set to 2/3 of the class with the lower sample size. From a scatter-plot of importance values generated by the random forest classifier (Mean Decrease in Accuracy, MDA), loci before the point where the differences between importance values level-off ("elbow method") were considered the most important (Goldstein, Hubbard, Cutler, & Barcellos, 2010).

4.3.7 Validation of a subset of outlier SNPs related to seasonal reproduction and to latitudinal divergence

We validated some of the top candidate loci detected with Pool-seq data that showed strong association with seasonal reproduction and latitudinal divergence with individual genotypes. For this, we genotyped 240 individuals (30 individuals from 8 locations) in 40 SNPs related to seasonal reproduction and 90 SNPs related to latitude using the Agena MassARRAY SNP genotyping platform (Agena Bioscience, Inc.). These SNPs were chosen considering these criteria: (*i*) top ranked based on importance values (Mean Decrease in Accuracy, MDA) obtained from the random forest algorithm (as described in the previous section), (*ii*) had ≥150 bp of flanking sequence for primer design, (*iii*) did not fall within or a few bases away from repetitive regions and had fewer than 4 flanking SNPs, (iv) when two or more top ranked SNPs were located within the same scaffold, the ones separated by ≥ 1Kb were kept, in an attempt to minimize redundancy in the panel). The application of these filters and the retrieval and preparation of DNA sequences for primer design for the Agena platform were performed with custom R scripts. A quality control of raw SNP genotypes was performed using PLINK (Purcell et al., 2007), in which SNPs and individuals with more than 20% missing data, and SNPs with minor allele frequency (MAF) lower than 0.01 were removed. We obtained a heatmap plot using the R function *heatmap.2* of the R package *gplots* for the visual inspection of individual genotype patterns. File format conversions required for missing data filtering and heatmap plotting were conducted with PGDSpider (Lischer & Excoffier, 2012) and a custom python script (data was transformed to PLINK format, then to VCF file format, and finally to 0,1,2 format).

4.3.8 Functional annotation of outlier loci

We investigated the potential effect on gene function of outlier SNPs associated with seasonal reproduction and the latitudinal cline using SNPeff v4.11 (build 2015-10-03) (Cingolani et al., 2012b) [*default parameters*]. This program determines the position of a SNP with respect to the constituents of a nearby gene within 5Kb (i.e. exons, introns, 5'-UTR region, etc.), and predicts its putative effect on gene and protein composition (i.e. synonymous and missense mutations, premature stop codon, etc., a complete list of effects is described in the program documentation). Variants located beyond 5Kb of a gene were annotated as 'intergenic'. We based this analysis on the current herring genome assembly and annotations (Martinez Barrio et al., 2016). Further, we separately examined gene ontology (GO) terms of the genes annotated to the outlier loci most strongly associated with seasonal reproduction and the latitudinal cline (-log$_{10}$*P*-value ≥ 7, equivalent to *P*-value ≤ 1x10^{-7}, lower threshold commonly used for significant association in human GWAS, (Fadista, Manning, Florez, & Groop, 2016; Panagiotou & Ioannidis, 2012). Details of the analysis performed on the GO terms can be found in the Supporting Information.

4.3.9 Genetic-Environment Association analysis

We performed redundancy analysis (RDA) and random forest (RF) regressions to identify environmental variables significantly associated with spatial patterns of population divergence.

The environmental dataset used for these analyses consisted of sea surface temperature (SST), sea bottom temperature (SBT), and sea surface salinity (SSS) for winter, spring, summer and fall seasons, for a total of 12 oceanographic variables. These variables are relevant in population structuring of numerous marine species in the NW Atlantic (Stanley et al., 2018).

To obtain environmental measures for each sampling location, we acquired monthly data layers of SST, SBT, and SSS between 2008-2017 from NEMO 2.3 (Nucleus for European Modelling of the Ocean), an oceanographic model developed by the Bedford Institute of Oceanography, Canada. A detailed description of oceanic (Madec, Delecluse, Imbard, & Levy, 1998) and sea ice (Fichefet & Maqueda, 1997) model components can be found in Wang, Brickman, Greenan, & Yashayaev (2016) and Brickman, Hebert, & Wang (2018). Data layers were converted to an ASCII grid with a NAD83 projection (ellipse GRS80), they had a nominal resolution of 1/12o (~5km^2), and a uniform land mask. Four seasonal bins, corresponding to winter (January-February-March), spring (April-May-June), summer (July-August-September), and fall (October-November-December), were averaged across 9 years in order to capture long-term trends of oceanographic variation. Data extraction for the 14 geo-referenced locations was conducted using custom R scripts (Stanley et al., 2018). Environmental data were standardized to zero mean and unit variance in R for downstream analysis. Collinearity between environmental variables was estimated with pairwise correlation coefficients computed with the function *pairs.panels* of the R package *psych* (Revelle, 2018) (Fig. S4.11), and with variance inflation factors (VIF) obtained from RDA models built with the R package *vegan* (Dixon, 2003). Prior to RDA, the most collinear variables were removed based on biological/ecological criteria (Forester, Lasky, Wagner, & Urban, 2018). Subsequently, remaining collinear variables were identified and removed one by one in consecutive RDA runs based on their VIF. The variable with the highest VIF was discarded in each run until all variables had a VIF < 5, following recommendations by (Zuur, Ieno, & Elphick, 2010).

For RDA, we used the reduced environmental data as constraining variables for the population allele frequencies of the top 500 outlier loci exhibiting the latitudinal pattern. RDA runs were performed with the R package *vegan*, following Jeffery et al. (2018) and Lehnert et al. (2018). Environmental variables that best explained genetic variance were identified using a bi-directional stepwise permutational ordination method (1 000 iterations) implemented in the R function *ordistep*. Significance of the overall

RDA model and of selected environmental variables was assessed with analysis of variance (ANOVA) using 1 000 permutations. In order to estimate the proportion of the genetic variance independently explained by environment, geographic distance, or both, we performed variance partitioning using partial redundancy analysis (pRDA), either conditioned on geographic distance (Cartesian coordinates) or selected environmental variables, respectively. Cartesian coordinates of each location, equivalent to the pairwise least-cost geographic distance between locations accounting for land as barrier, were obtained with the R function *CartDist* (Stanley & Jeffery, 2017). Concordance between Cartesian and geographic coordinates was assessed with a linear regression (Fig. S4.3).

For RF regressions, we used the population allele frequencies of each outlier locus as single response vectors and the 12 standardized environmental variables as predictors. A RF regression was performed for each outlier locus with the R package *randomForest*, as described in Lehnert et al. (2018) and Sylvester et al. (2018). Default parameters for regression were applied to the RF runs (*mtry* = *p*/3, where *p* is the total number of predictors, or environmental variables in this case), except that *ntree* was set to 10,000. The selected number of trees to grow per run (*ntree*) assured Mean Decrease in Accuracy (MDA) convergence, as demonstrated in a pilot test that compared MDA of predictors of 3 independent RF runs (correlation coefficient *r* = 0.9999, Fig. S4.4). Environmental variables were then ranked based on their relative importance to explain genetic variance from the averaged MDA values across loci, and the mean residual square error (MSE) of each location averaged across loci.

4.3.10 Isolation-by-distance pattern test

To evaluate whether global (all loci) and latitude-related population structure (subset of loci) corresponded to an isolation-by-distance (IBD) pattern, we determined the significance of the association between geographic and genetic distances for all possible pairs of sampled spawning sites using Mantel tests (Mantel 1967) with 9999 permutations, implemented in the R package *ade4* (Dray & Dufour, 2007). Genetic distances were linearized ($\hat{F}_{\mathrm{ST}} = \frac{\hat{F}_{\mathrm{ST}}}{1-\hat{F}_{\mathrm{ST}}}$) (Rousset, 1997) with $\hat{F}_{ST}$ computed using all SNPs identified across the genome, in the first case, or solely outlier SNPs strongly associated with latitudinal divergence, for the latter. Geographic distances were estimated with the R package *CartDist* (Stanley & Jeffery, 2017) as the least-coast oceanic distance in Km considering land as barrier.

4.4 Results

4.4.1 Pool-sequencing

A total of ~800 GB of raw sequence data were obtained. After quality filtering and adapter trimming, 6 119 940 640 reads of optimal quality (Phred score > 20) were available for the genomic analysis. Read mapping statistics indicated that > 98.8% of read-pairs were correctly aligned to the stitched version of the herring reference genome (mapping quality MQ > 48, median insert size of 527 bp) (Table S4.1), confirming that misalignment errors, if present, were negligible. Average read depth of coverage per pool ranged between 25x to 44x and varied between sequencing batches [2015 batch mean 28.7 ± 4.0, 2016 batch mean 36.9 ± 2.6 (Table S4.1). We monitored the potential effect of coverage variation in downstream analysis, in particular for collections with lower coverage (TRB-S, NTS-S, and GEB-F). Variant calling resulted in 11 154 328 raw SNPs of which 2 189 380 passed quality filters and were retained for further analysis.

4.4.2 Population structure

As observed in our previous study (Lamichhaney et al., 2017), spawning aggregations in the NW Atlantic clustered according to reproductive season in a Neighbor-Joining tree, with spring and fall spawning collections forming separate groups (Fig. 4.1B), although a few exceptions were observed. BDO-S sample was in an intermediate position with respect to these two main clusters, and a spring-collected sample in Newfoundland (NDB-S) clustered with the fall group, suggesting it may be composed of a large proportion of fall spawners. A closer examination of the fall group revealed clustering according to latitude. Southern collections in the Scotian Shelf (MUS-F, GEB-F), Bay of Fundy (SCB-F), and Gulf of Maine (ME4-F) were separated from northern collections in the Gulf of St. Lawrence (MIR-F, BLS-F), Newfoundland (TRB-F, NDB-S, NDB-F) and

Labrador (LAB-F). Such separation suggests genetic differences may exist between herring inhabiting these two geographic regions.

The pairwise fixation index F_{ST} for pools ($\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$) ranged between 0.012 and 0.043, indicating low levels of genetic structure among the 14 spawning aggregations studied (Fig. 4.1C, pairwise F_{ST} values in Table S4.2). Nevertheless, three clear patterns of subtle genetic differentiation were noticeable: i) between spring and fall spawners (SIL-S, SPH-S, NTS-S, vs. others, $\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ 0.022-0.043), ii) within spring spawners, the sample from the NW of the Gulf of St. Lawrence (SIL-S) was the most genetically distinguishable ($\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ ~0.030), and iii) within fall spawners, the two southernmost collections (GEB-F and ME4-F) were the most divergent ($\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ 0.020-0.031). In general, the largest genetic differentiation was observed between spring spawners and the most southern collections ($\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ ~0.040). Interestingly, the two spring-collected samples BDO-S and NDB-S (two samples presumably containing both spring and fall spawning individuals, see below) exhibited similar levels of differentiation ($\hat{F}_{\mathrm{ST}}^{\mathrm{pool}}$ 0.022-0.033) with samples comprising solely spring spawners (SIL-S, SPH-S, NTS-S) as with samples comprising solely fall spawners.

4.4.3 Outlier loci detection and genome-wide patterns of differentiation

A PCA-based whole-genome scan for the identification of SNPs putatively under selection revealed two main axes of genomic differentiation in NW Atlantic herring: spawning season, and geographic origin according to latitude. In a PCA plot based on 2,189,380 SNPs (Fig. 4.1D), spring and fall spawning herring were distinguishable along the first principal component (PC1) (36% of variance explained). PC2 distinguished two collections, German Bank (GEB-F) and Northumberland Strait (NTS-S) from the rest (Fig. S4.5). These two collections exhibited the shallowest average sequencing coverage, suggesting this axis (PC2) is largely reflecting an artefact of sequencing. PC2 was therefore ignored (Fig. S4.5). On PC3, the southernmost collections, distributed on the Scotian Shelf, Bay of Fundy and Maine (MUS-F, SCB-F, GEB-F, ME4-F), were differentiated from the aggregations in the Gulf of St. Lawrence, Newfoundland, and

Labrador (30% of the variance explained) that formed a tight cluster. The sample from Maine (ME4-F) was the most differentiated of all, followed by German Banks (GEB-F), the southernmost location sampled on the Scotian Shelf. Along PC1, BDO-S and SIL-S were positioned in between the spring and fall spawners, BDO-S being closer to the fall samples and SIL-S to the spring samples. NBD-S clustered tightly with the fall spawners. In general, with the exception of the two southernmost samples (GEB-F and ME4-F), fall spawning aggregations grouped more closely together than the spring spawning ones, suggesting that more genetic differences may exist among the spring spawners than among fall spawners included in this study.

In PC1, a total 14 724 outlier SNPs were detected (with Benjamini-Hochberg-adjusted *P*-values and FDR ≤ 0.01). A Manhattan plot depicting significance values (–$\log_{10}P$-value) of outlier loci for this PC disclosed numerous "peaks" or regions of divergence across the genome, spanning about 18 scaffolds and numerous genes (Fig. 4.2A). The top SNPs of these scaffolds were in the proximity of genes with known function in reproduction (± 5Kb), such as *TSHR*, *ESRA*, *HERPUD2*, *CALM* (Martinez Barrio *et al.* 2016). Moreover, a new set of candidate genes linked to seasonal reproduction were *ISO3*, *SERTM1*, *SIPA1L1*, *CAMKK1, TMEM150C, CBLB, ENTPD5, KCNJ6, LPAR6* and *GPR119*, as they were near top outlier loci in the unique islands of differentiation only observed in the NW Atlantic (Lamichhaney et al., 2017). A heatmap depicting standardized population allele frequencies of the top 200 outlier loci from the scaffolds identified with RF (ranked in descending order by -$\log_{10}P$-value) distinguished aggregations by spawning season (Fig. 4.2B), with fall spawners fixed for one allele and almost all spring spawners fixed for the alternative allele. The exceptions to this observation were three aggregations sampled in spring, BDO-S, SIL-S, and NDB-S. The first two collections exhibited allele frequencies around 0.5, while NDB-S showed population allele frequencies consistent with fall spawners. These results indicate that BDO-S and SIL-S either correspond to a mixture of spring and fall spawning individuals or to hybrids or both, and that NDB-S should be considered as a sample of fall spawners, suggesting possible mislabeling.

In PC3, a total of 6 595 outlier loci were detected (with BH-adjusted *P*-values and FDR ≤ 0.01). A Manhattan plot for this PC disclosed four main regions of divergence across the genome, corresponding to scaffolds 44, 122, 869 and 958, and a small number of outlier loci from other scaffolds (Fig. 4.2C). The top SNPs in the four main scaffolds were located within 5Kb of the genes *FAM129B, FNBP1, SH3GLB2*, and *GPR107.* A heatmap representing standardized population allele frequencies of the top 200 outlier loci from the scaffolds identified with RF (ranked in descending order by -log$_{10}$*P*-value) revealed contrasting genetic patterns according to latitude (Fig. 4.2D). In northern collections, including Labrador (LAB-F), Newfoundland (NDB-S, NDB-F, TRB-F, SPH-S), Gulf of St. Lawrence (BLS-F, SIL-S, MIR-F, NTS-S), Bras D'Or lake (BDO-S), and inner Bay of Fundy (SCB-F), one allele was close to fixation; in the southernmost collection, in Maine (ME4-F), the alternative allele was in high frequency; and in intermediate southern collections along the Scotian Shelf (MUS-F, GEB-F) allele frequencies were around 0.5. An extended examination of population allele frequencies of the 14,724 outlier SNPs detected in PC1 (Fig. S4.8), revealed that additional SNPs from the four scaffolds showing the latitudinal pattern were present in PC1 and showed the same pattern as the ones found in PC3 (3 378). Thus, these SNPs were removed from the PC1 set and added to the ones detected in PC3, for a total of 11,346 SNPs associated with seasonal reproduction and 9 973 SNPs associated with latitude.

A closer examination of the genomic distribution of outlier SNPs revealed that seasonal reproduction-related outliers exhibited varying levels of significance (–log$_{10}$*P*-value up to 30) (Fig. 4.2A), were confined to a particular region within a scaffold (around 50-500 Kb) and spanned a given set of genes (Fig. S4.6). In contrast, latitude-related outliers showed similar significance values (–log$_{10}$*P*-value ~15) (Fig. 4.2C), were widely spread along scaffolds (covering between 480 Kb to 4.75 Mb) and spanned numerous genes (Fig. S4.7).

4.4.4 Validation of a subset of outlier SNPs related to seasonal reproduction and to latitudinal divergence

A total of 230 individuals (NDB-F: 30, NDB-S: 29, SIL-S: 27, NTS-S: 30, BDO-S: 28, MUS-F: 29, GEB-F: 27, ME4-F: 30) and 52 and 74 SNPs related to seasonal reproduction and latitudinal divergence, respectively, passed the missing rate and MAF quality filters. Heatmaps depicting individual SNP genotypes for each of the two panels (Fig. S4.9) confirmed the overall patterns of population allele frequencies of the two axes of divergence detected with Pool-seq data (Fig. 4.2B,D), seasonal reproduction and latitude.

The SNP panel discriminating spawning season revealed that the spring-collected samples SIL-S and BDO-S corresponded to a mixture of spring and fall spawners and putative hybrids, the latter defined as heterozygous individuals at many of the loci showing a high degree of fixation between groups. SIL-S comprised an even proportion of pure fall spawners and putative hybrids with a few pure spring spawners, whereas BDO-S comprised mostly pure fall spawners and a few hybrids and spring spawners. The other spring-collected samples, NTS-S, consisted of mostly pure spring spawners and a few putative hybrids, while NDB-S corresponded to pure fall spawners. In contrast, all the fall-collected samples genotyped (NDB-F, MUS-F, GEB-F and ME4-F) corresponded to pure fall spawners, with a few heterozygous loci.

The SNP panel discriminating by latitude confirmed northern samples were characterized by high frequency of one allele, while the alternative allele had greater frequency in the southernmost sample (in Maine), although putative hybrids were present in both cases in varying proportions. Intermediate locations (BDO-S, MUS-F, GEB-F) exhibited a genotypic cline of increasing proportion of putative hybrids towards the south.

4.4.5 Functional annotation of outlier loci

A total of 2,977 and 1,257 outlier SNPs associated with seasonal reproduction and latitudinal divergence, respectively, were annotated with respect to a neighboring gene (within 5Kb). For both cases, the majority of outlier SNPs were located within introns and intergenic regions, or 5Kb upstream or downstream of genes (Fig. 4.3A). A small number of outlier SNPs were predicted as synonymous (~2%) or missense variants (1.6% and 0.9%, for spawning- and latitude-related outliers, respectively).

Excluding intergenic variants and genes that did not correspond to an orthologous gene in zebrafish, a list of 298 and 182 genes associated with seasonal reproduction and latitudinal divergence in herring, respectively, resulted from the annotated outlier loci. For seasonal reproduction-related genes, 126 had a GO term in the biological process category, 109 in the cellular component category, and 120 in the molecular function category (Fig. S4.10A). For latitude-related genes, 90 had a GO term in the biological process category, 72 in the cellular component category, and 80 in the molecular function category; considered together, close to half of the genes lacked GO classification. A comprehensive description of particular functions within the three GO categories and the number of genes in each of them is presented in Fig. S4.10B).

The overrepresentation enrichment analysis (ORA) of both sets of candidate genes did not reach statistical significance (FDR of 5%) (Table S4.3 and S4.4), likely due to the large number of genes lacking GO annotation (Fig. S4.10). However, a closer examination of the top GO terms with *P*-value < 0.05 (ranked in ascending *P*-values from ORA, Table S4.3 and S4.4, GO terms indicated with an asterisk), suggested that seasonal reproduction-related candidate genes may participate in biological processes such as metabolism of lipids, cell adhesion, biosynthesis of cellular products, peptidyl-aminoacid modification, protein complex biogenesis, inositol lipid-mediated signaling, developmental maturation, regulation of developmental process, and cellular component organization (Fig. 4.3B-top, Table S4.3). These genes might primarily act in cellular components such the endoplasmic reticulum and the whole membrane (Fig. 4.3B-middle) and play a molecular function related to cell adhesion molecule and protein binding and lipid transporter and transferase activities (Fig. 4.3B-bottom). The top GO terms of candidate genes associated with latitudinal divergence were all involved in embryological and organ development processes (Fig. 4.3C-top, Table S4.4). These genes might act in cellular components such phosphatase complex, collagen trimer, and in the extracellular region (Fig. 4.3C-middle), and participate in sulfur compound binding and hydrolase and isomerase activities (Fig. 4.3C-bottom).

4.4.6 Genome-Environment Association analysis

Collinearity among several of the environmental variables examined and redundancy analyses (See Supporting Information) allowed us to reduce the environmental data set to just three variables: summer SBT, winter SST, and spring SSS. Winter SST (Win_SST) (Fig. 4.4A) was the environmental variable that best explained the genetic variance of outlier loci exhibiting the latitudinal cline ($F = 16.7$, $p = 0.001$, from *ordistep* function) in the RDA approach (Fig. 4.4B). Other temperature or salinity variable in the reduced environmental dataset were statistically insignificant(from ANOVA with 1 000 permutations, significance value = 0.05). Spawning aggregations were separated according to Win_SST on RDA axis 1, which explained 58.1% of the total genetic variance ($R^2 = 0.58$, adjusted $R^2 = 0.55$). pRDA however, showed that the Win_SST-based RDA model was no longer significant when the effect of geographic distance between sites was removed from the model. A variance partitioning analysis revealed that the interaction between environment and geographic distance explained the greatest proportion of clinal genetic variation (44.9%).

In agreement with RDA results, RF regressions also indicated that Win_SST was the most important environmental variable (MDA = 23.5), followed by Fall_SST (MDA = 21.8) (Fig. 4.4C). The other temperature variables had lower importance (MDA < 10), and salinity measures were the least important of all (MDA < 5). ME4-F, the southernmost spawning aggregation sampled, exhibited the highest mean square error (MSE = 0.21), followed by SCB-F and MUS-F (MSE ~ 0.05), whereas the other 10 collections had lower MSE, below 0.03 (Fig. 4.4D).

A closer examination of the map of the NW Atlantic depicting average Win_SST over the last 9 years and the predominant population allele frequency of the 14 sites studied (Fig. 4.4A), revealed that herring in “northern” collections in the Bay of Fundy, the Gulf of St. Lawrence, and Newfoundland and Labrador were characterized by being exposed to temperatures below zero (-2 ºC), whereas in “southern” collections they were mainly exposed to temperatures above zero (>2 ºC).

4.4.7 Isolation-by-distance test

The Mantel test showed there is not a significant linear relationship between geographic and genetic distances for all loci across the genome ($R^2 = 0.04$), whereas there is a significant linear relationship ($R^2 = 0.30$) between geographic distance and genetic differentiation when only looking at outlier SNPs exhibiting the latitudinal break in population allele frequencies between northern and southern collections (Fig. 4.5).

4.5 Discussion

Here we described patterns of genetic variation at the whole-genome level in Atlantic herring populations distributed across the reproductive range of the species in North America. This study represents the most comprehensive assessment of this kind in the region to date. We uncovered fine-scale population structure at outlier loci putatively under selection, despite low differentiation at selectively neutral loci. This observation is consistent with previous genetic work on herring in both, the NE (Guo et al., 2016; Lamichhaney et al., 2012; Limborg et al., 2012; Martinez Barrio et al., 2016; Teacher et al., 2013) and the NW Atlantic (Lamichhaney et al., 2017; McPherson et al., 2004, 2001). The large population sizes, high potential for gene flow, and minute effect of genetic drift explain the low genetic differentiation observed at neutral loci (Palumbi, 1994). These conditions also favor the more efficient action of natural selection, which seems to be shaping the genetic differences observed at outlier loci.

While prior genomic studies disclosed genetic structure with seasonal reproduction and salinity (Lamichhaney et al., 2012; Martinez Barrio et al., 2016), and others suggested structuring along the salinity/temperature gradient in the Baltic Sea from a dozens of markers (Gaggiotti et al., 2009; Guo et al., 2016; Limborg et al., 2012), here we successfully disentangled two main overlapping axes of divergence supported by thousands of outlier SNPs: seasonal reproduction and a latitudinal cline defining a north-south genetic break. Our genetic-environment association analyses indicated that winter sea-surface temperature is the best predictor of the spatial structure observed. These results: demonstrate for the first time that herring from the north (Labrador, Newfoundland, Gulf of St. Lawrence and Bay of Fundy) are genetically distinguishable from the ones in the south (Scotian Shelf and Maine) regardless of their spawning season; indicating that thermal-minima related factors are likely driving latitudinal genetic differentiation; and provide additional evidence supporting the recently described multispecies biogeographic break in eastern Nova Scotia (Stanley et al. 2018).

Outlier SNPs exhibited remarkable clustering, forming so-called “genomic regions of divergence” (Nosil et al., 2009; Turner, Hahn, & Nuzhdin, 2005), and extreme allele frequency differences (i.e. alternative alleles were close to fixation in either spring- or fall-spawning, or in northern- or the southernmost populations). Theory predicts that formation of genomic regions of divergence (Schluter, 2009; Wu, 2001) and fixation of different alleles conducive to opposing phenotypes often result from natural selection acting in contrasting directions between environments (Vitti et al., 2013). Considering the heterogeneous environmental properties of the Northwest Atlantic (Melvin et al., 2009; Townsend et al., 2004) and having discarded an effect of genetic drift and an isolation-by-distance pattern, we conclude that disruptive selection may be the main evolutionary force involved in population structuring in the region.

A few exceptions to the allele fixation pattern were observed in both axes of divergence. In seasonal reproduction outliers, two aggregations sampled in spring, BDO-S and SIL-S, exhibited allele frequencies around 0.5 at SNPs being close to fixation for opposite alleles in other populations of spring- and fall-spawning herring. This observation suggests these collections either correspond to a mixture of spring- and fall-spawners, or to a unique population where allele diversity is favored. Individual genotypes of a subset of diagnostic SNPs of spawning time confirmed BDO-S and SIL-S comprised a mixture of spring and fall spawners and putative hybrids (i.e. heterozygous individuals at many of the loci showing a high degree of fixation between groups). In latitude-related outliers, intermediate allele frequencies were observed in MUS-F and GEB-F, two locations in southwestern Nova Scotia, mid-range in the latitudinal cline. Interestingly, these locations are few kilometers south of the biographic barrier described in the NW Atlantic (Stanley et al., 2018). Environmental conditions in the NW Atlantic vary between years in relation to oceanographic global trends (Townsend et al., 2004). It is possible then that populations in southwestern Nova Scotia experience significant inter-annual environmental fluctuations during winter months, depending on the strengthening either of the warm Gulf Stream flowing north or of the cold Labrador Current flowing south. Under these dynamic circumstances, it is possible that balancing selection may be maintaining polymorphism at these loci. Additional studies including an extended sampling in the southern region could be used to test this hypothesis.

A closer examination of the genomic regions of divergence revealed they vary in size and genomic location between the two axes of divergence. Seasonal reproduction-related outliers were distributed across 18 scaffolds in which they spanned about 50-500 Kb and a given set of genes. In contrast, latitude-related outliers were mostly spread in four scaffolds, covering a larger extension, from 480 Kb to 4.75 Mb, and larger number of genes. The observation that latitude-related outliers were widely distributed and consistently divergent across four large scaffolds suggests that they could be located within a chromosomal rearrangement. If this were the case, the expectation would be that populations from the north were homozygous for one state of the variant, the ones in southwest Nova Scotia were polymorphic, and in the Gulf of Maine were homozygous for the alternative state of the variant. Further research supported by a linkage map, not described yet for herring, is required for the evaluation of this hypothesis.

A bioinformatic evaluation of the functional effect of outlier SNPs disclosed that, for both axes of divergence, the majority of SNPs were located within introns, intergenic regions, and 5Kb upstream or downstream of genes, and a smaller proportion corresponded to missense mutations (1.6% and 0.9%, for spawning- and latitude-related outliers, respectively). Mutations in introns can modify regulatory domains, intron-exon boundaries and RNA splicing (Pagani & Baralle, 2004); missense mutations result in a different amino acid; and mutations in regulatory elements can modify gene expression (Epstein, 2009; Metzger et al., 2016; Nei, 2007). While at this point is not possible to trace a direct link between single SNPs and gene function or identify causal mutations, our observations suggest that single base changes in introns, protein-coding, and regulatory regions may be involved in adaptive divergence in NW Atlantic herring, in agreement with previous observations in the NE Atlantic (Martinez Barrio et al., 2016).

Gene annotation of top outlier SNPs confirmed that *TSHR*, *HERPUD2*, *SOX1*, *SOX11A*, *SYNE1*, *SYNE2*, and *ESR2A* are candidate genes related to seasonal reproduction. These genes have a known function in reproduction and were previously linked to spawning time in NE Atlantic herring (Lamichhaney et al., 2017; Martinez Barrio et al., 2016). We discovered an additional set of candidate genes, *ISO3*, *SERTM1*, *SIPA1L1*, *CAMKK1, TMEM150C, CBLB, ENTPD5, KCNJ6, LPAR6* and *GPR119*, corresponding to the genomic regions of differentiation uniquely observed in the NW Atlantic (Lamichhaney et al., 2017), hence, they can potentially be involved in local adaptation. Candidate genes related with the latitudinal cline are *FAM129B, FNBP1, SH3GLB2*, and *GPR107*.

A qualitative examination of the top ranked GO terms indicated that candidate genes related to seasonal reproduction may be involved in biological processes such as metabolism of lipids, biosynthesis of cellular products, developmental maturation, regulation of developmental process, and cellular component organization. Similarly, latitude-related candidate genes may participate in embryological and organ development processes. These observations suggest that outlier SNPs underlying the two axes of divergence may be involved in different physiological pathways, and that natural selection along the latitudinal cline likely acts on early life stages, in agreement with the proposed hypothesis for the multispecies climatic cline (Stanley et al., 2018). It is likely that early life stages experience selection along the latitudinal cline given that larval retention areas are in the proximity of spawning grounds (Stephenson et al., 2009). If selection would act on juveniles or adults, which are highly migratory, then the pattern should not coincide with spawning locations.

We provide genetic evidence that suggests timing of reproduction and latitudinal spawning location are features under disruptive selection leading to local adaptation. Several characteristics of herring biology and ecology seem to support this. For instance, (*i*) spawning occurs at predictable times and locations, the timing differs among geographic regions (Stephenson et al., 2009), and there is no evidence indicating that individual fish can switch spawning season (Melvin et al., 2009); (*ii*) herring spawn once a year and exhibits spawning site fidelity (Wheeler & Winters, 1984); (*iii*) spring- and fall-spawners differ in morphometric characters, in life-history traits (fecundity, egg size and growth), and in phenotypic traits (number of vertebrae and otolith shape) (Baxter, 1959; Cushing, 1967; Messieh, Anthony, & Sinclair, 1985); growth rate, otolith shape, and vertebral counts seem to be largely influenced by genetic factors (Berg et al., 2018); (*iv*) early life stages spawned in different seasons and locations experience contrasting environmental conditions (e.g. in the Gulf of St. Lawrence, eggs released by spring spawners hatch after 30 days at 5°C, while eggs of fall spawners hatch after 10 days at 15°C; in Nova Scotia, eggs of fall spawners hatch in 11 days at 10°C) (Scott & Scott, 1988); (*v*) larval retention areas occur near spawning grounds and are stable over time, in predictable patterns related to oceanographic conditions (Stephenson et al., 2009); and (*vi*) genetic differences between spring- and fall-spawners are temporally stable (Kerr et al., 2019). From this, we then infer that timing of reproduction and latitudinal spawning location can be adaptive strategies to increase offspring survival, particularly at vulnerable early life stages, in environments that vary seasonally and geographically. When timing of reproduction is largely heritable, the resulting temporal assortative mating may reduce gene flow between individuals breeding at different times (Hendry & Day, 2005). In herring, gene flow may be limited between early spring-spawners and late fall-spawners even if they are in sympatry (as their gonads are not ripe at the same time, as we observed in our samples). How do hybrids occur? We hypothesize hybridization could happen between late spring-spawners and early fall-spawners at geographic areas where both reproductive strategies coexist (e.g. in the Gulf of St. Lawrence), and when the onset of gonadal maturation coincides (likely temperature driven). Hybrids would survive then, if they can cope with the local environmental conditions.

Although disruptive selection is a strong candidate for explaining latitudinal divergence in herring, other mechanisms are possible. For example, additional biotic or abiotic factors that covariate with temperature may be the actual drivers of adaptation. Pre- or post-zygotic reproductive incompatibilities that coincide with latitude (but are not dependent on) can result in the observed spatial genetic discontinuity (Bierne, Welch, Loire, Bonhomme, & David, 2011). The current latitudinal break may actually reflect historical vicariance (Bradbury et al., 2010), not contemporary population dynamics (Palumbi, 1994). Further studies are required to evaluate these alternative hypotheses.

Even though valuable information was obtained through this study, there were some limitations. In Pool-seq individual information is missed, thus it is not possible to correct accidental mixing of individuals with different origin/spawning season. To avoid this, we selected maturing and ripe fish collected in known spawning grounds during the local peak of reproduction. Despite these precautions, we found evidence of some mixed aggregations (SIL-S and BDO-S). Moreover, in the over-representation enrichment analysis statistical significance was not reached. This outcome may have been influenced by the restriction that only herring candidate genes with a zebrafish ortholog could be included, and that half of the total genes mapped to zebrafish lacked a GO term. We expect with a more complete reference genome and annotations, along with functional experiments, a better functional characterization of outlier loci will be achieved.

Our findings have several implications and potential applications in fisheries. Firstly, our results support the maintenance of separate management of spring- and fall-spawning components currently in place across most of the region. Secondly, management units should be revised in order to protect the functional intraspecific biodiversity revealed in this study, specifically considering a climate change scenario as spring-spawners seem to be less resilient to a warming ocean (Melvin et al., 2009). Thirdly, as we now have the molecular tools to distinguish herring spawning in spring or autumn and in northern and southern regions, a subset of outlier SNPs reported here can be used for genetic monitoring of stock composition already at the larval stage and out of breeding seasons to minimize the risk of overexploitation of vulnerable components within mixed stocks. And lastly, the current herring population models could be revised as none of them are in complete agreement with our genetic data, as similarly noted by McPherson et al. (2004). For instance, the discrete population concept proposes that gene flow is limited, hybrids have reduced fitness, and local populations are reproductively isolated by fixed spawning time, natal homing, spawning site fidelity, and larval retention areas with particular hydrographic features (Sinclair, 1988; Sinclair & Iles, 1989). While our data agrees with most of this, the presence of numerous putative hybrids suggests that gene flow may be more extensive than expected under this model and they are viable. In the dynamic balance population concept there is significant gene flow, no stable population structure, no fixed spawning time, no philopatry, no larval retention areas, as populations respond to changing environmental conditions (Smith & Jamieson, 1986). The temporal and spatial structuring we observed is opposite to this model. And in the metapopulation concept (adopted migrant) there is repeated homing to traditional spawning grounds defined by hydrographic features, migration and homing patterns are socially transmitted, and significant gene flow can occur as vagrants are adopted by non-natal local populations (McQuinn, 1997). This model implies an isolation-by-distance pattern and that spawning time is not genetically determined (it is learned), contrary to our observations.

In summary, our results confirm that Atlantic herring is a system that provides ideal conditions for the study of ecological adaptation with gene flow in the wild (Lamichhaney et al., 2017; Martinez Barrio et al., 2016), and provide insight into patterns and mechanisms of genomic divergence and local adaptation despite gene flow in an abundant and highly dispersive marine fish.

4.6 Acknowledgments

We thank staff at Fisheries and Oceans Canada, the Maine Department of Resources, Comeau's Seafoods Ltd, Cape Breeze Seafoods Ltd, and fishers Crystal and Donald Kent, Gordon McKay, and many others for their valuable contribution in sample collection. Thanks to Gregory McCracken for assistance in sampling and DNA extraction, to Gavin Douglas, Emma Sylvester, Sarah Lehnert, Ryan Stanley, Simone Fior, Alan Bergland, Miguel Carneiro, Michael Blum, and Mathieu Gautier for data analysis assistance. Many thanks to Ryan Stanley, David Brickman, and Zeliang Wang from DFO for obtaining and providing the oceanographic data layers used in this study. Library preparation and shotgun sequencing was performed in The Centre for Applied Genomics of the Hospital for Sick Children, Canada. Computations were conducted on the supercomputer Mp2 of the University of Sherbrooke, managed by Calcul Québec and

Compute Canada and funded by the Canada Foundation for Innovation (CFI), the ministère de l'Économie, de la science et de l'innovation du Québec (MESI) and the Fonds de recherche du Québec - Nature et technologies (FRQ-NT). A.P.F.P. and D.E.R. thank the Killam Trust. A.P.F.P. thanks the Vanier Canada Graduate Scholarship, the President's Award of Dalhousie University, the Nova Scotia Graduate Scholarship, the Lett Fund and a Strategic grant to D.E.R for graduate studies funding. This study was funded by NSERC Discovery and Strategic grants to D.E.R.

4.7 Author contributions

D.E.R. and A.P.F.P. designed and conceived the study; C.B., R.S., K.E., L.P., and J.L.M provided herring samples; A.P.F.P contributed to tissue collection and processing, and performed lab work and bioinformatics data analysis; D.E.R and L.A. contributed to the interpretation of results; A.P.F.P. wrote the manuscript with input from D.E.R. and L.A. All authors vetted and approved the manuscript before submission. Abbreviations of names as described in the statement of co-authorship (page 7).

4.8 Tables

Table 4.1 Characteristics of the 14 herring spawning aggregations included in this study.

Locality	**Code**	**Sample size (N)**	**Geographic coordinates (longitude, latitude)**		**Sampling (dd/mm/yy)**	**Season**	**Salinity (PPM)**	**Sequencing year**
Seven Islands	SIL-S	50	-66.33	50.09	06/06/2012	Spring	35	2016
Stephenville	SPH-S	48	-57.94	49.73	30/05/2012	Spring	35	2016
Northumberland Strait	NTS-S	50	-64.12	46.30	14/05/06	Spring	35	2015
Labrador	LAB-F	50	-55.50	52.25	24/08/2014, 22/08/2015	Fall	35	2016
Blanc Sablon	BLS-F	49	-57.31	51.38	13/08/2014	Fall	35	2016
Notre Dame Bay	NDB-S	50	-55.44	49.55	03/05/2015	Spring	35	2016
Notre Dame Bay	NDB-F	50	-55.47	49.55	26/10/2015	Fall	35	2016
Trinity Bay	TRB-F	50	-53.47	47.84	28/09/2014	Fall	35	2015
Miramichi	MIR-F	50	-63.96	47.04	25/08/2014	Fall	35	2016
Bras D'Or lake	BDO-S	50	-60.85	45.93	20/04/2016	Spring	25	2016
Scots Bays	SCB-F	50	-64.92	45.17	24/08/2015	Fall	35	2016
Musquodoboit	MUS-F	50	-63.10	44.63	28/10/2015	Fall	35	2016
German Banks	GEB-F	50	-66.33	43.45	28/08/2014	Fall	35	2015
Maine fishing area 514	ME4-F	50	-70.41	42.09	19/10/2015	Fall	35	2016

4.9 Figures

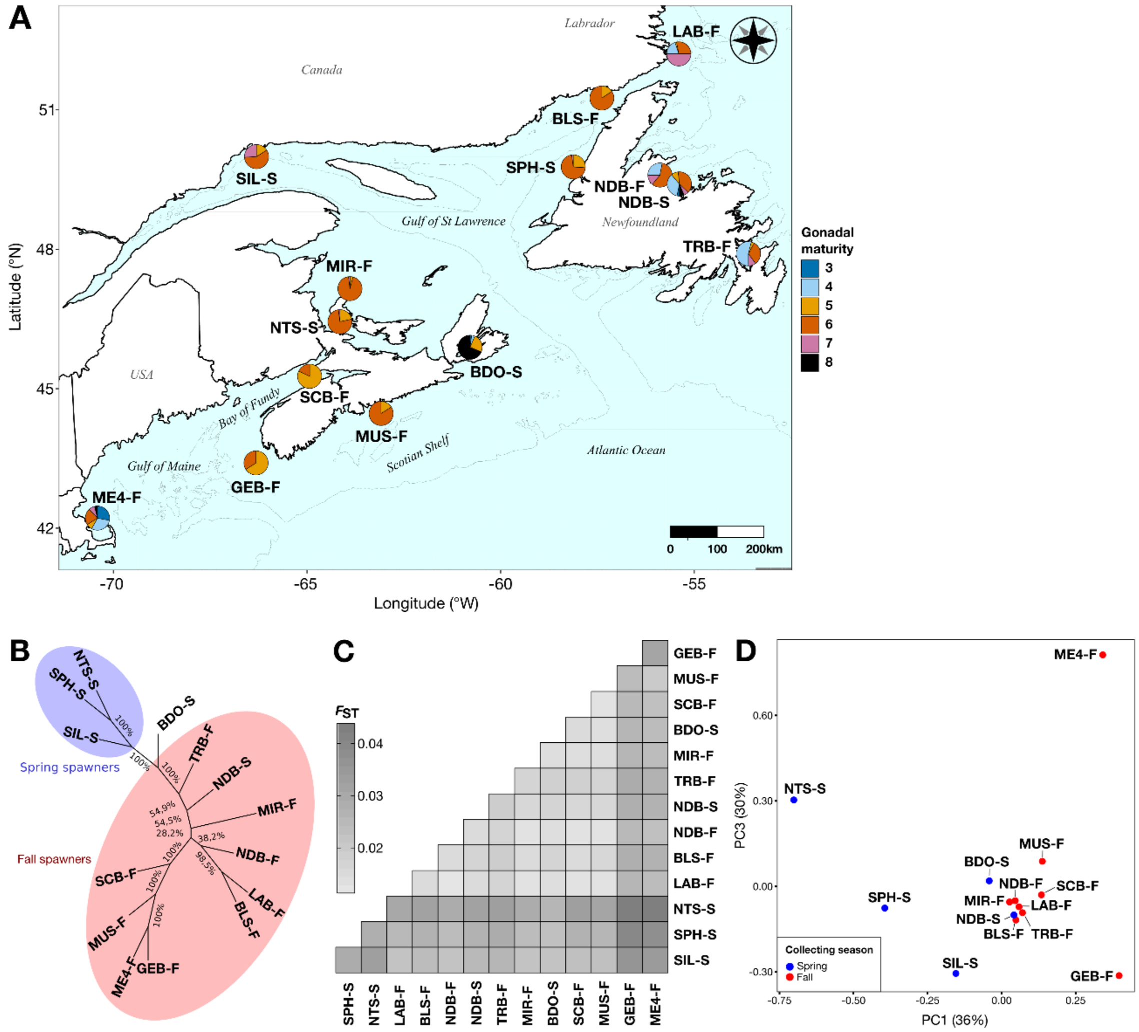


Figure 4.1 Geographic location and population structure among 14 spawning aggregations in the NW Atlantic. (A) Map depicting sampling locations in the Northwest Atlantic. Location names as described in Table 4.1. Pie charts indicate the proportion of individuals in a given gonadal maturity stage for each spawning aggregation. Dark and light blue: maturing individuals (stages 3 and 4, respectively), light and dark orange: ready-to-spawn and actively spawning individuals (stages 5 and 6, respectively), pink: spent individuals (recently spawned) (stage 7), and black: individuals with resting gonads (stage 8) (Bucholtz et al., 2008). **(B)** NJ phylogenetic tree based on Nei's distance calculated from population allele frequencies of 2 189 380 SNPs (percent bootstrap support is shown for all branches, based on 1000 bootstrapping). The collections clustered according to reproductive season into two main groups, spring (blue oval) and fall (red oval) spawners with a few exceptions. BDO-S was in an intermediate position between these two groups. The spring-collected sample NDB-S was closer to fall

spawners. Within the fall spawners, collections clustered depending on the latitude, forming the southern and northern sub-groups. **(C)** Heatmap depicting pairwise F_{ST} estimates based on population allele frequencies of 2 189 380 SNPs (values presented in Table S4.2). Samples are ordered by collecting season with "S" indicating spring and "F" fall. Within season of collection, samples are ordered by latitude. Shading represents the degree of genomic divergence. Pairwise F_{ST} ranged between 0.012 and 0.043, indicating overall and varying low levels of population genetic structure. The most significant genomic differentiation was observed between spring and fall spawning aggregations. Within spring spawners, the location SIL-S appeared as the most differentiated, and within fall spawners, the greatest genetic divergence was observed between the two southernmost collections (GEB-F and ME4-F) and all others. Notably, two spring-collected samples (BDO-S and NDB-S) showed similar patterns of differentiation with respect to spring spawners as other fall spawners. Collection name abbreviations are as defined in Table 4.1. **(D)** Plot of principal components 1 and 3 explaining 36% and 30%, respectively, of the genetic variation among 14 herring spawning aggregations in the NW Atlantic (based on 2 189 380 SNPs). Each dot represents a spawning aggregation. Colors indicate spawning season, blue for spring, red for fall, and yellow for mix. Aggregations were distinguishable by spawning season along PC1 and by geographic origin along PC3.

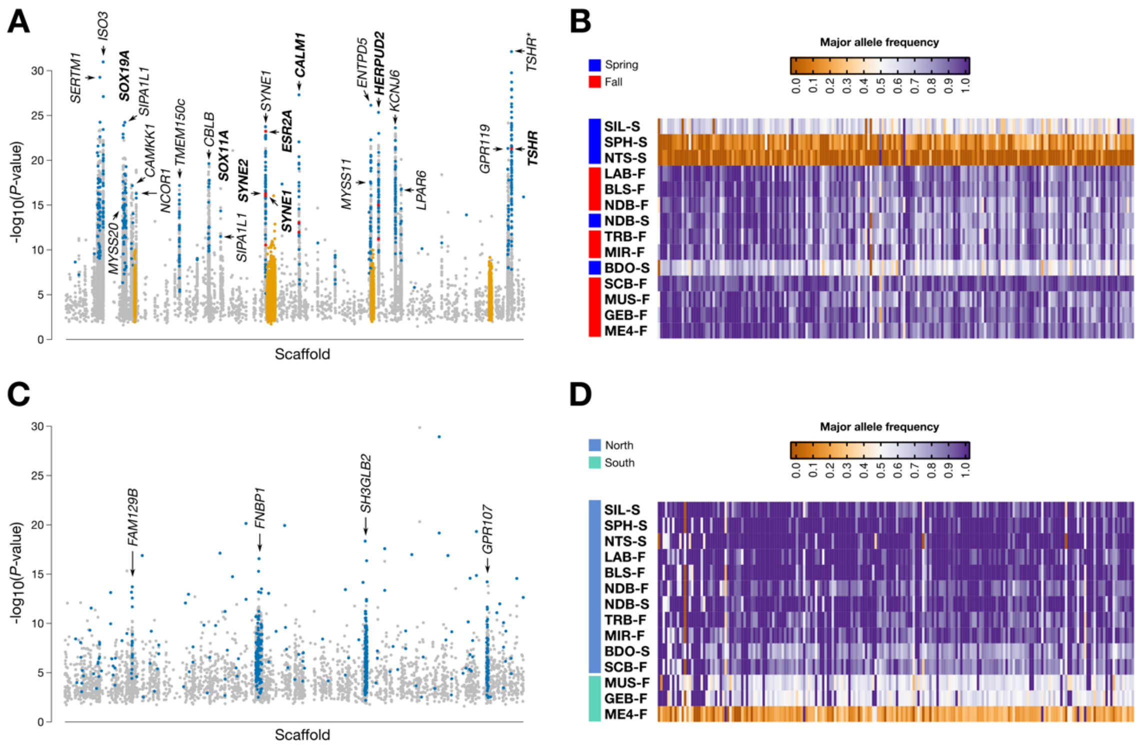


Figure 4.2 Genome-wide patterns of differentiation associated with seasonal reproduction and latitude for 14 Atlantic herring spawning aggregations in the Northwest Atlantic. (A-C) Manhattan plots depicting the genomic position of outlier SNPs and their respective significance association value (–log10*P*-value) obtained with *pcadapt*, **(A)** for PC1 and **(C)** for PC3. Each dot of the Manhattan plots represents a single SNP locus. For the purpose of visualization, only outlier loci per PC are displayed (14 726 for PC1 and 6 570 for PC3). The top-ranked 500 SNPs based on importance values from a RF classifier are highlighted in blue. SNPs reported in (Lamichhaney et al., 2017) as highly associated with seasonal reproduction are emphasized in red in (A). SNPs within the four scaffolds showing the latitudinal pattern in PC3 but present in PC1 were denoted in yellow (A). When available, annotation of the closest gene 5Kb upstream or downstream of both, the top SNP per scaffold and the SNPs reported in our previous study, are shown. The SNP annotated as *TSHR** falls within the first exon of the *TSHR* gene (unpublished Leif Andersson com. pers.). **(B and D)** Heatmaps depicting standardized population allele frequencies of the top 200 outlier loci distinguishing collections by **(B)** seasonal reproduction (PC1), and **(D)** latitude (PC3). Each row in the heatmaps corresponds to a collection site and each column to a SNP. SNPs were ranked in descending order based on their significance association value, -log$_{10}$*P*-value (from left to right). Cell colors represent the population allele frequency of the major allele; thus, purple indicates fixation of the major allele (allele frequency of 1) whereas orange represents fixation of the minor allele (major allele frequency of 0).

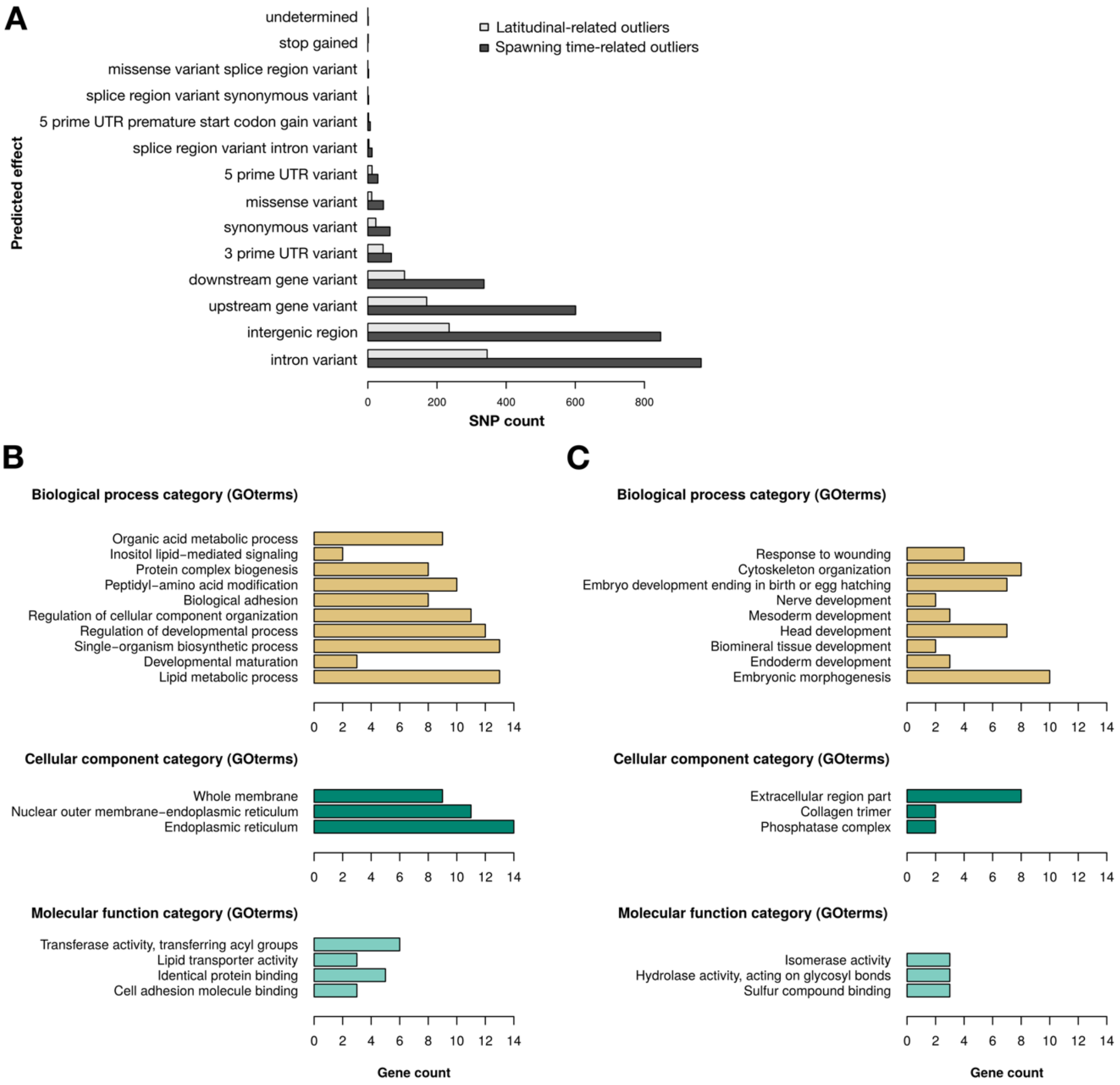


Figure 4.3 Functional characterization of outlier SNPs. (**A**) Functional classification of outlier SNPs associated with seasonal reproduction (dark gray) and with latitude (light gray), counts. (**B-C**) Bar plots showing the relative proportion of genes in each of the top GO terms associated with (**B**) seasonal reproduction and (**C**) latitude, for each biological category (i.e. biological process, cellular component, and molecular function). Top GO terms corresponded to the ones with *P*-value < 0.05 (ranked in ascending order based on their *P*-values obtained from ORA, Table S4.3 and S4.4, GO terms with an asterisk). Gene counts are indicated within parenthesis.

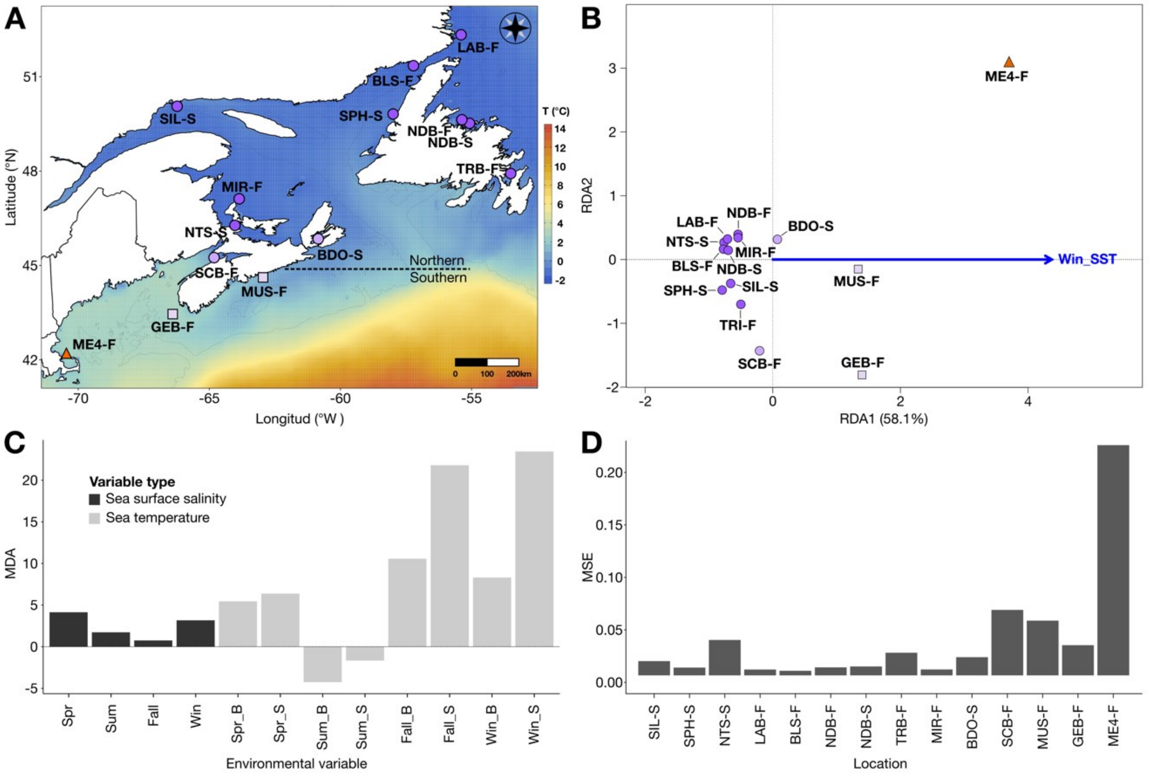


Figure 4.4 Genetic-environment association analysis. (A) Map depicting winter sea surface temperature averaged between 2008-2017 and the predominant population allele frequencies at diagnostic SNPs in the 14 spawning aggregations included in this study. A purple circle represents the prevalent alleles fixed in northern collections, a light purple square corresponds to southern collections with intermediate allele frequencies, and an orange tringle shows fixation of the minor alleles in the southernmost collection in Maine. **(B)** Redundancy analysis plot based on population allele frequencies of the top 500 outlier loci, with respect to latitudinal cline, ranked with a random forest classifier. Each circle corresponds to a spawning aggregation and their color indicates the predominant population allele frequency; labelling as in Fig. 4.4A. The vector of the most significant environmental variable is shown in blue, in this case it is surface sea temperature in winter months (January-February-March) (F = 16.7, p = 0.001). The length of the vector indicates its level of correlation with genetic variance. RDA1 explained 58.1% of the genetic variance, RDA2 is shown just for plotting purposes. **(C-D)** Random forest regression results. **(C)** Mean decrease in accuracy (MDA) for 12 environmental variables averaged across all runs and loci. Bars in light gray correspond to temperature values and bars in dark gray to surface salinity. Within temperature measures, surface is denoted with "S" whereas bottom is indicated with "B". Average sea surface winter temperature is the most important variable explaining genetic variance, followed by Fall sea surface temperature. Salinity variables were the least important. **(D)** Mean squared error (MSE) for each location averaged across all runs and loci. The southernmost collection had the largest MSE, followed by the collection in the Bay of Fundy and in the Scotian Shelf.

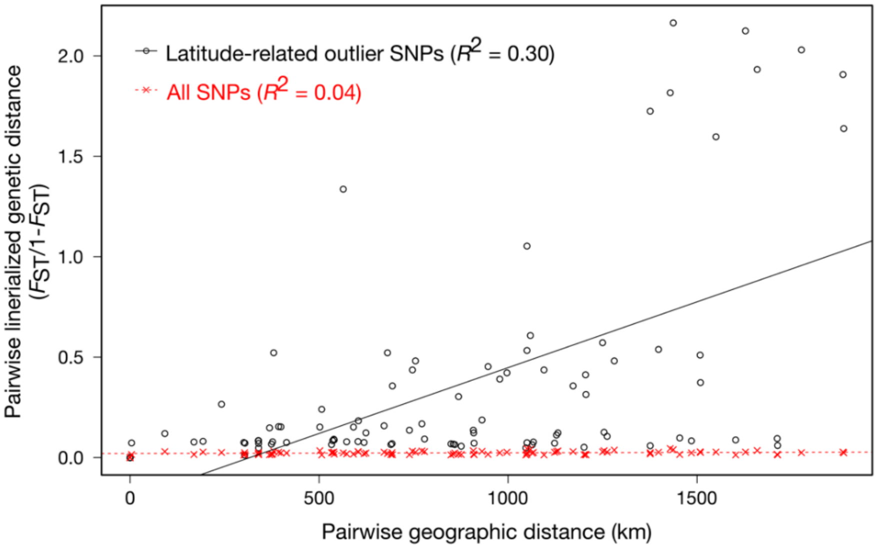


Figure 4.5 Isolation-by-distance (IBD) test for 14 NW Atlantic herring populations based on all neutral and outlier SNPs or on latitude-related SNPs only. Regression between linearized genetic distance (F_{ST}/1-F_{ST}), calculated from either 2,189,371 SNPs (red "X"s) or from 6,595 latitude-related outlier SNPs (open black circles), and geographic distance (in km) between pairs of populations. The dashed red line and the continuous black line correspond to the best fit line in each case. R^2 values indicate the correlation between geographic and genetic distance matrices used in the Mantel test (Mantel's test for all SNPs: $P < 0.001$, $R^2 = 0.04$, 9999 replicates; for latitude-related SNPs only: $P < 0.001$, $R^2 = 0.30$, 9999 replicates). Note the IBD pattern is only observed in the latitude-related outlier SNPs, not in all SNPs.

4.10 Supporting Information

Extended Materials and Methods

Read mapping, SNP calling and filtering

We adapted the Genome Analysis Toolkit (GATK) Best Practices workflow (Van der Auwera et al., 2013) to variant discovery in Pool-seq data and to our computing infrastructure. Firstly, a stitched version of the herring genome was obtained for optimal SNP caller performance in the computer cluster available. As GATK works best with small numbers of contigs (<100), we joined the 145,282 contigs of the current herring genome assembly into 94 super-scaffolds using the python script *ScaffoldStitcher.py* (O'Connor Lab, 2016). A long spacer (2000 N's) was inserted between contigs to minimize the occurrence of unspecific paired-read mapping (an excessively large insert size, > 2000 bp, would help diagnose an affected read pair).

Secondly, sequence reads of each pool were independently aligned against the stitched herring genome using the Burrows-Wheeler Aligner (BWA) v0.7.12-r1039 [*default parameters, MEM algorithm*] (Li, 2013). Read mapping summary statistics per pool were obtained from the resulting binary alignment map (BAM) files using Qualimap v2.2.1 (Okonechnikov et al., 2015), and jointly assessed for all pools with MultiQC v1.3. We then sorted reads, marked PCR duplicates, and added read groups to each BAM file using Picard tools v2.10.2 (Broad Institute, 2018), and obtained an index file for each BAM file with SAMtools v1.5 (Li & Durbin, 2009a).

Thirdly, SNP calling was performed using GATK v3.8 (McKenna et al., 2010) (a diagram summarizing SNP calling implementation is shown in Fig. S4.1). For this, we obtained an intermediate Genomic Variant Call Format (gVCF) file for each pool with the GATK-*HaplotypeCaller* algorithm (Poplin et al., 2017) [*parameters: -ERC GVCF -ploidy 10 --max_genotype_count 286 -newQual -mbq 20 -minPruning 5 --read_filter OverclippedRead.* Note that actual pool ploidy was ~100, but 10 was used as this was the maximum ploidy successfully used by GATK-users at the time of run. Larger numbers exceeded algorithm capacity]. We then performed a joint SNP calling on the 14 gVCF files using the GATK-*GenotypeGVCFs* algorithm [*parameters: -T GenotypeGVCFs -newQual*] (Broad Institute, 2014).

Lastly, the raw variant calls were filtered using GATK, Popoolation2, and custom python scripts. With GATK, we kept only biallelic SNPs present in all pools and SNPs with the most reliable sequence support as described by GATK-annotations. We removed SNPs with GATK annotations beyond cutoff values established from density plots made with the R package *ggplot* (Fig. S4.2) [*filters: FS>60.0, SOR>3.0, MQ<40.0, MQRankSum<-12.5, ReadPosRankSum<-8.0*] (a description of each annotation can be found in (Broad Institute, 2016). With Popoolation2, we retained SNPs with a minimum coverage of 20x, a maximum coverage below the 2% of the empirical coverage distribution of each pool, and a minimum count of 4 reads supporting the minor allele across populations. With these filters we aimed to keep variants supported by a large number of reads, to exclude variants likely resulting from sequencing errors, and to exclude spurious SNPs likely falling within copy number variations (CVNs) or repetitive regions characterized by extremely high coverage, respectively. With custom python scripts we removed monomorphic SNPs that are not informative for allele frequencies and F_{ST} calculation.

Allele count correction and population allele frequencies estimation

In Pool-seq applications, population allele frequencies are derived from the total read counts supporting a variant site. Read coverage, however, can vary across the genome due to non-biological factors during sequencing and read alignment. For example, some genomic regions may accumulate more reads by chance during sequencing, because of lower efficiency of Illumina technology at GC regions (Dohm et al., 2008) or due to inherent limitations of read aligners, in particular at repetitive regions. In addition, chromosome sampling during DNA pooling and sequencing may result in pseudo-replication, or the sequencing of a chromosome more than once (Feder et al., 2012; Kolaczkowski et al., 2011). To control for these factors and minimize their potential effect on the estimation of population allele frequencies, some correction methods of raw read counts have been proposed. Two of the most common methods are: 1) random subsampling of bases to a target coverage, implemented in Popoolation2 (Kofler, Pandey, & Schlötterer, 2011a); and 2) rescaling of read counts to the effective number of chromosomes/alleles in a pool, also known as the effective sample size n_{eff} (Feder et al., 2012; Kolaczkowski et al., 2011). For the first method, it has not been discussed in the literature how to select an optimal target coverage, nor its effect in downstream analysis. In contrast, the second method has a strong theoretical support and its utility has been demonstrated (Bergland, Behrman, O'Brien, Schmidt, & Petrov, 2014; Feder et al., 2012; Kolaczkowski et al., 2011; Wiberg, Gaggiotti, Morrissey, & Ritchie, 2017). Therefore, prior to estimating population allele frequencies, we applied the n_{eff} allele count correction to the raw read counts using a custom python script implementing this formula:

$$n_{\mathrm{eff}} = \frac{(n * CT) - 1}{n + CT}$$

where CT = read depth, and the number of chromosomes/alleles in the pool (n) is $2N$ (for diploid species). With a custom python script, we then computed population allele frequencies as the ratio of the reference allele read count to the total reads supporting a variant site.

Functional annotation of outlier loci

We excluded from this analysis SNPs predicted to be located in intergenic regions as no gene was reliably assigned to them. We also evaluated whether candidate genes were enriched in three biological categories, molecular function, biological process, or cellular component, using an overrepresentation enrichment analysis (ORA) of GO terms implemented in *Webgestalt* (Wang, Duncan, Shi, & Zhang, 2013). ORA statistically evaluates the number of genes reported in a particular pathway among a gene list of interest, for example, resulting from GWAS or gene expression studies (Khatri, Sirota, & Butte, 2012). For ORA we used the zebrafish (*Danio rerio*) homologs of the herring candidate genes related with seasonal reproduction or latitudinal divergence, the non-redundant functional database for each of the three biological categories considered, and the genome protein coding database as reference set. *P*-values were adjusted for multiple testing with the Benjamini-Hochberg method (Benjamini & Hochberg, 1995). Significantly enriched GO terms were identified using a false discovery rate (FDR) of 5%. In addition, we examined the top GO terms with *P*-value < 0.05 of the two candidate gene sets (i.e. seasonal reproduction and latitudinal cline) for each of the three biological categories considered. GO terms were ranked in ascending order based on their *P*-values obtained from ORA.

Genome-Environment Association analysis

Environmental variable reduction was conducted before performing environment association analysis, given the high collinearity observed between surface and bottom seasonal temperatures (Spring, $r = 0.9$; Fall, $r = 0.92$; Winter, $r = 0.86$; Summer, $r = 0.68$) and among all seasonal surface salinity measures ($r = 0.79\text{-}0.96$) (Fig. S4.11). Since herring is a pelagic fish that spawns near shore in relatively shallow waters (50-100 m deep), we only retained surface temperatures for most seasons except for Summer, for which we kept both, surface and bottom measures, as their $r < 0.7$. A cutoff value of $|r| > 0.7$ is generally used for removal of collinear variables (Forester et al., 2018). In addition, we only retained spring salinity (Spr_SSS) as it represents the general variation in salinity of other seasons. In summary, the reduced environmental dataset used in RDA had 6 variables (spring, summer, winter, and fall sea surface temperature; summer sea bottom sea temperature; and spring sea surface salinity). This dataset was further reduced to 3 variables (summer sea bottom temperature, winter sea surface temperature, and spring sea surface salinity) based on VIF values of RDA runs.

Supplementary tables

Table S4.1 Read mapping summary statistics of the Pool-seq data of 14 herring spawning aggregations included in this study. Abbreviations, SD: Standard deviation, MQ: Mapping quality.

Locality	**Code**	**Number of reads**		**%GC**	**Mapped reads, both in pair**	**Mean MQ**	**Median insert size**	**Median coverage**	**Sequence batch**
Seven Islands	SIL-S	426105370		43	99.50	48.28	546	37	2016
Stephenville	SPH-S	460698330		43	99.06	48.43	505	41	2016
Northumberland Strait	NTS-S	391232819		44	98.98	47.21	531	*28*	*2015*
Labrador	LAB-F	449046570		43	98.60	48.33	534	41	2016
Blanc Sablon	BLS-F	436962768		43	98.71	48.3	532	37	2016
Notre Dame Bay	NDB-F	447042754		43	98.97	48.38	533	42	2016
Notre Dame Bay	NDB-S	422028368		43	98.99	48.17	517	36	2016
Trinity Bay	TRB-F	411425306		43	98.85	46.91	545	*33*	*2015*
Miramichi	MIR-F	454079824		43	98.13	48.37	536	43	2016
Bras D'Or lake	BDO-S	441958372		43	98.74	48.35	526	40	2016
Scots Bays	SCB-F	475694602		43	98.89	48.32	538	44	2016
Musquodoboit	MUS-F	445447932		43	98.83	48.36	533	41	2016
German Banks	GEB-F	377288622		43	98.60	48.33	534	*25*	*2015*
Maine fishing area 514	ME4-F	492521602		44	98.85	47.39	478	41	2016
	Total	6119940640	**Average**	43.1	98.8	48.1	527.7	37.8	
			SD	0.4	0.3	0.5	17.7	5.7	

Table S4.2 Pairwise $\hat{F}_{ST}^{pool}$ for 14 herring populations in the northwest Atlantic.

	SIL-S	SPH-S	NTS-S	LAB-F	BLS-F	NDB-F	NDB-S	TRB-F	MIR-F	BDO-S	SCB-F	MUS-F	GEB-F	ME4-F
SIL-S	0	0.029	0.033	0.022	0.024	0.022	0.025	0.026	0.022	0.022	0.023	0.023	0.036	0.034
SPH-S	0.029	0	0.028	0.025	0.027	0.024	0.027	0.029	0.024	0.023	0.027	0.026	0.04	0.038
NTS-S	0.033	0.028	0	0.03	0.032	0.029	0.032	0.033	0.029	0.026	0.033	0.031	0.043	0.043
LAB-F	0.022	0.025	0.03	0	0.015	0.012	0.015	0.017	0.013	0.014	0.013	0.013	0.026	0.024
BLS-F	0.024	0.027	0.032	0.015	0	0.015	0.017	0.019	0.015	0.017	0.015	0.016	0.028	0.027
NDB-F	0.022	0.024	0.029	0.012	0.015	0	0.015	0.017	0.013	0.014	0.013	0.013	0.025	0.023
NDB-S	0.025	0.027	0.032	0.015	0.017	0.015	0	0.019	0.016	0.017	0.016	0.016	0.028	0.027
TRB-F	0.026	0.029	0.033	0.017	0.019	0.017	0.019	0	0.017	0.018	0.018	0.017	0.029	0.027
MIR-F	0.022	0.024	0.029	0.013	0.015	0.013	0.016	0.017	0	0.014	0.014	0.014	0.026	0.024
BDO-S	0.022	0.023	0.026	0.014	0.017	0.014	0.017	0.018	0.014	0	0.015	0.014	0.026	0.024
SCB-F	0.023	0.027	0.033	0.013	0.015	0.013	0.016	0.018	0.014	0.015	0	0.013	0.025	0.023
MUS-F	0.023	0.026	0.031	0.013	0.016	0.013	0.016	0.017	0.014	0.014	0.013	0	0.024	0.02
GEB-F	0.036	0.04	0.043	0.026	0.028	0.025	0.028	0.029	0.026	0.026	0.025	0.024	0	0.031
ME4-F	0.034	0.038	0.043	0.024	0.027	0.023	0.027	0.027	0.024	0.024	0.023	0.02	0.031	0

Table S4.3 Overrepresentation analysis of outlier SNPs associated with seasonal reproduction. Gene Ontology (GO) terms are listed for each of the three biological categories, biological process, cellular component, and molecular function. Analysis performed with Webgestalt, only outliers with a significance value -$\log_{10}P$-value > 7.0 were included. Top GOs (*P*-value < 0.05) are indicated with an asterisk.

Category		Gene set	Description	C	O	E	R	*P*-value	FDR
Biological process	*	GO:0006629	lipid metabolic process	383	13	5.879	2.211	0.006	1.000
	*	GO:0021700	developmental maturation	32	3	0.491	6.108	0.013	1.000
	*	GO:0044711	single-organism biosynthetic process	454	13	6.969	1.865	0.022	1.000
	*	GO:0050793	regulation of developmental process	415	12	6.370	1.884	0.025	1.000
	*	GO:0051128	regulation of cellular component organization	384	11	5.894	1.866	0.033	1.000
	*	GO:0022610	biological adhesion	247	8	3.791	2.110	0.036	1.000
	*	GO:0018193	peptidyl-amino acid modification	342	10	5.250	1.905	0.037	1.000
	*	GO:0070271	protein complex biogenesis	250	8	3.837	2.085	0.038	1.000
	*	GO:0048017	inositol lipid-mediated signaling	23	2	0.353	5.665	0.048	1.000
	*	GO:0006082	organic acid metabolic process	311	9	4.774	1.885	0.049	1.000
		GO:0001655	urogenital system development	99	4	1.520	2.632	0.065	1.000
		GO:0071822	protein complex subunit organization	280	8	4.298	1.861	0.066	1.000
		GO:0044087	regulation of cellular component biogenesis	101	4	1.550	2.580	0.069	1.000
		GO:0006325	chromatin organization	245	7	3.761	1.861	0.082	1.000
		GO:0048646	anatomical structure formation involved in morphogenesis	449	11	6.892	1.596	0.083	1.000
		GO:0006259	DNA metabolic process	254	7	3.899	1.795	0.095	1.000
		GO:0006839	mitochondrial transport	36	2	0.553	3.619	0.105	1.000
		GO:0050808	synapse organization	36	2	0.553	3.619	0.105	1.000
		GO:0051239	regulation of multicellular organismal process	419	10	6.431	1.555	0.109	1.000
		GO:0048699	generation of neurons	478	11	7.337	1.499	0.116	1.000
Cellular component	*	GO:0005783	endoplasmic reticulum	401	14	6.236	2.245	0.003	0.079
	*	GO:0042175	nuclear outer membrane-endoplasmic reticulum membrane network	276	11	4.292	2.563	0.003	0.079
	*	GO:0098805	whole membrane	261	9	4.059	2.218	0.018	0.313
		GO:0031975	Envelope	320	9	4.976	1.809	0.057	0.643
		GO:0019867	outer membrane	59	3	0.917	3.270	0.063	0.643
		GO:0048475	coated membrane	35	2	0.544	3.675	0.102	0.868
		GO:0031982	Vesicle	275	7	4.276	1.637	0.133	0.890
		GO:0005773	Vacuole	129	4	2.006	1.994	0.140	0.890
		GO:0005694	Chromosome	236	6	3.670	1.635	0.159	0.898
		GO:0005739	Mitochondrion	484	10	7.526	1.329	0.214	1.000
		GO:0005667	transcription factor complex	110	3	1.711	1.754	0.244	1.000
		GO:1903293	phosphatase complex	20	1	0.311	3.215	0.270	1.000
		GO:0070603	SWI/SNF superfamily-type complex	21	1	0.327	3.062	0.281	1.000
		GO:0099023	tethering complex	25	1	0.389	2.572	0.325	1.000
		GO:0098588	bounding membrane of organelle	433	8	6.733	1.188	0.359	1.000
		GO:0097458	neuron part	217	4	3.374	1.185	0.439	1.000

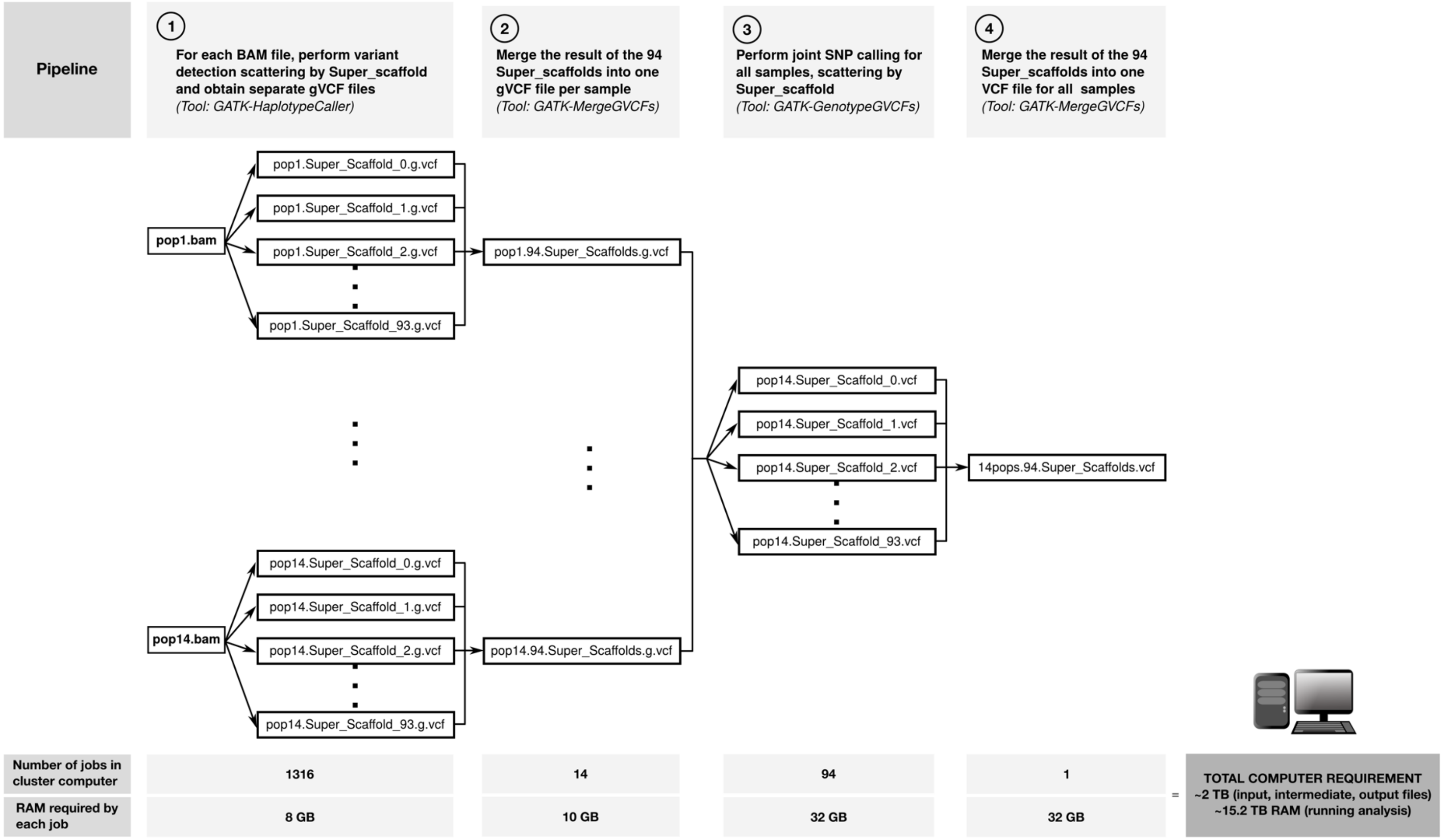
Category		Gene set	Description	C	O	E	R	*P*-value	FDR
		GO:0042579	Microbody	37	1	0.575	1.738	0.441	1.000
		GO:0044815	DNA packaging complex	44	1	0.684	1.462	0.500	1.000
		GO:1990204	oxidoreductase complex	45	1	0.700	1.429	0.508	1.000
		GO:0044421	extracellular region part	367	6	5.707	1.051	0.512	1.000
Molecular function	*	GO:0050839	cell adhesion molecule binding	28	3	0.383	7.835	0.006	0.342
	*	GO:0042802	identical protein binding	91	5	1.244	4.018	0.008	0.342
	*	GO:0005319	lipid transporter activity	41	3	0.561	5.351	0.018	0.424
	*	GO:0016746	transferase activity, transferring acyl groups	156	6	2.133	2.813	0.020	0.424
		GO:0000287	magnesium ion binding	62	3	0.848	3.539	0.053	0.679
		GO:0005506	iron ion binding	106	4	1.449	2.760	0.057	0.679
		GO:0046983	protein dimerization activity	264	7	3.610	1.939	0.069	0.679
		GO:0070405	ammonium ion binding	34	2	0.465	4.302	0.078	0.679
		GO:0016765	transferase activity, transferring alkyl or aryl (other than methyl) groups	35	2	0.479	4.179	0.082	0.679
		GO:0019842	vitamin binding	36	2	0.492	4.063	0.086	0.679
		GO:0044877	macromolecular complex binding	279	7	3.815	1.835	0.087	0.679
		GO:0043177	organic acid binding	38	2	0.520	3.849	0.095	0.679
		GO:0008565	protein transporter activity	42	2	0.574	3.482	0.112	0.742
		GO:0005509	calcium ion binding	362	8	4.950	1.616	0.121	0.746
		GO:0019904	protein domain specific binding	49	2	0.670	2.985	0.144	0.827
		GO:0008289	lipid binding	220	5	3.008	1.662	0.182	0.980
		GO:0000981	RNA polymerase II transcription factor activity, sequence-specific DNA binding	225	5	3.077	1.625	0.194	0.982
		GO:0048037	cofactor binding	135	3	1.846	1.625	0.281	1.000
		GO:0030246	carbohydrate binding	77	2	1.053	1.899	0.284	1.000
		GO:0000988	transcription factor activity, protein binding	141	3	1.928	1.556	0.303	1.000

Table S4.4 Overrepresentation analysis of outlier SNPs associated with latitude. Gene Ontology (GO) terms are listed for each of the three biological categories, biological process, cellular component, and molecular function. Analysis performed with Webgestalt, only outliers with a significance value -log$_{10}$*P*-value > 7.0 were included. Top GOs (*P*-value < 0.05) are indicated with an asterisk.

Category		**Gene set**	**Description**	**C**	**O**	**E**	**R**	***P*-value**	**FDR**
Biological process	*	GO:0048598	embryonic morphogenesis	397	10	4.281	2.336	0.010	1
	*	GO:0007492	endoderm development	45	3	0.485	6.183	0.012	1
	*	GO:0031214	biomineral tissue development	22	2	0.237	8.431	0.023	1
	*	GO:0060322	head development	270	7	2.911	2.404	0.026	1
	*	GO:0007498	mesoderm development	63	3	0.679	4.416	0.030	1
	*	GO:0021675	nerve development	26	2	0.280	7.134	0.032	1
	*	GO:0009792	embryo development ending in birth or egg hatching	288	7	3.105	2.254	0.035	1
	*	GO:0007010	cytoskeleton organization	372	8	4.011	1.994	0.047	1
	*	GO:0009611	response to wounding	127	4	1.369	2.921	0.048	1
		GO:0006818	hydrogen transport	79	3	0.852	3.522	0.053	1
		GO:0016311	dephosphorylation	192	5	2.070	2.415	0.056	1
		GO:0007049	cell cycle	389	8	4.194	1.907	0.058	1
		GO:0031099	regeneration	83	3	0.895	3.352	0.060	1
		GO:0001503	ossification	42	2	0.453	4.416	0.075	1
		GO:0010035	response to inorganic substance	42	2	0.453	4.416	0.075	1
		GO:0007417	central nervous system development	354	7	3.817	1.834	0.087	1
		GO:0001501	skeletal system development	158	4	1.704	2.348	0.091	1
		GO:0044087	regulation of cellular component biogenesis	101	3	1.089	2.755	0.095	1
		GO:1901135	carbohydrate derivative metabolic process	444	8	4.788	1.671	0.105	1
		GO:0048568	embryonic organ development	301	6	3.246	1.849	0.106	1
Cellular component	*	GO:1903293	phosphatase complex	20	2	0.215	9.289	0.019	0.672
	*	GO:0005581	collagen trimer	27	2	0.291	6.881	0.034	0.672
	*	GO:0044421	extracellular region part	367	8	3.951	2.025	0.040	0.672
		GO:0048471	perinuclear region of cytoplasm	38	2	0.409	4.889	0.062	0.795
		GO:0005667	transcription factor complex	110	3	1.184	2.533	0.114	1.000
		GO:0031984	organelle subcompartment	68	2	0.732	2.732	0.166	1.000
		GO:0031300	intrinsic component of organelle membrane	72	2	0.775	2.580	0.181	1.000
		GO:1905360	GTPase complex	27	1	0.291	3.440	0.254	1.000
		GO:1905368	peptidase complex	49	1	0.528	1.896	0.413	1.000
		GO:0005739	mitochondrion	484	6	5.211	1.152	0.423	1.000
		GO:0030054	cell junction	222	3	2.390	1.255	0.431	1.000
		GO:0031975	envelope	320	4	3.445	1.161	0.456	1.000
		GO:0005794	Golgi apparatus	331	4	3.563	1.123	0.483	1.000
		GO:0098588	bounding membrane of organelle	433	5	4.661	1.073	0.506	1.000
		GO:0043235	receptor complex	67	1	0.721	1.386	0.519	1.000
		GO:0098552	side of membrane	73	1	0.786	1.272	0.549	1.000
		GO:0005856	cytoskeleton	485	5	5.221	0.958	0.611	1.000
		GO:0098796	membrane protein complex	295	3	3.176	0.945	0.625	1.000

Category		Gene set	Description	C	O	E	R	*P*-value	FDR
		GO:0019898	extrinsic component of membrane	96	1	1.033	0.968	0.650	1.000
		GO:1990904	ribonucleoprotein complex	324	3	3.488	0.860	0.690	1.000
Molecular function	*	GO:1901681	sulfur compound binding	51	3	0.523	5.736	0.015	0.825
	*	GO:0016798	hydrolase activity, acting on glycosyl bonds	71	3	0.728	4.120	0.036	0.825
	*	GO:0016853	isomerase activity	77	3	0.790	3.799	0.044	0.825
		GO:0098531	transcription factor activity, direct ligand regulated sequence-specific DNA binding	39	2	0.400	5.000	0.060	0.825
		GO:0005198	structural molecule activity	282	6	2.892	2.075	0.069	0.825
		GO:0016757	transferase activity, transferring glycosyl groups	215	5	2.205	2.268	0.069	0.825
		GO:0016746	transferase activity, transferring acyl groups	156	4	1.600	2.500	0.076	0.825
		GO:0005509	calcium ion binding	362	7	3.713	1.885	0.077	0.825
		GO:0016788	hydrolase activity, acting on ester bonds	397	7	4.072	1.719	0.111	1.000
		GO:0003707	steroid hormone receptor activity	59	2	0.605	3.305	0.122	1.000
		GO:0048037	cofactor binding	135	3	1.385	2.167	0.161	1.000
		GO:0003690	double-stranded DNA binding	156	3	1.600	1.875	0.215	1.000
		GO:0008066	glutamate receptor activity	26	1	0.267	3.750	0.235	1.000
		GO:0019205	nucleobase-containing compound kinase activity	28	1	0.287	3.482	0.251	1.000
		GO:0004896	cytokine receptor activity	35	1	0.359	2.786	0.303	1.000
		GO:0005539	glycosaminoglycan binding	36	1	0.369	2.709	0.311	1.000
		GO:0001067	regulatory region nucleic acid binding	191	3	1.959	1.532	0.312	1.000
		GO:0016874	ligase activity	119	2	1.220	1.639	0.346	1.000
		GO:0009055	electron carrier activity	46	1	0.472	2.120	0.379	1.000
		GO:0016773	phosphotransferase activity, alcohol group as acceptor	492	6	5.046	1.189	0.392	1.000

Supplementary figures


Pipeline
1
For each BAM file, perform variant detection scattering by Super_scaffold and obtain separate gVCF files
(Tool: GATK-HaplotypeCaller)
2
Merge the result of the 94 Super_scaffolds into one gVCF file per sample
(Tool: GATK-MergeGVCFs)
3
Perform joint SNP calling for all samples, scattering by Super_scaffold
(Tool: GATK-GenotypeGVCFs)
4
Merge the result of the 94 Super_scaffolds into one VCF file for all samples
(Tool: GATK-MergeGVCFs)
pop1.bam
pop1.Super_Scaffold_0.g.vcf
pop1.Super_Scaffold_1.g.vcf
pop1.Super_Scaffold_2.g.vcf
pop1.Super_Scaffold_93.g.vcf
pop1.94.Super_Scaffolds.g.vcf
pop14.bam
pop14.Super_Scaffold_0.g.vcf
pop14.Super_Scaffold_1.g.vcf
pop14.Super_Scaffold_2.g.vcf
pop14.Super_Scaffold_93.g.vcf
pop14.94.Super_Scaffolds.g.vcf
pop14.Super_Scaffold_0.vcf
pop14.Super_Scaffold_1.vcf
pop14.Super_Scaffold_2.vcf
pop14.Super_Scaffold_93.vcf
14pops.94.Super_Scaffolds.vcf
=
TOTAL COMPUTER REQUIREMENT
~2 TB (input, intermediate, output files)
~15.2 TB RAM (running analysis)

Number of jobs in cluster computer	1316	14	94	1
RAM required by each job	8 GB	10 GB	32 GB	32 GB

Figure S4.1 Workflow to perform SNP calling on Pool-seq data using GATK 3.8 in a cluster computer. Note that a newer GATK version (4.x) has been released, thus these steps may not apply.

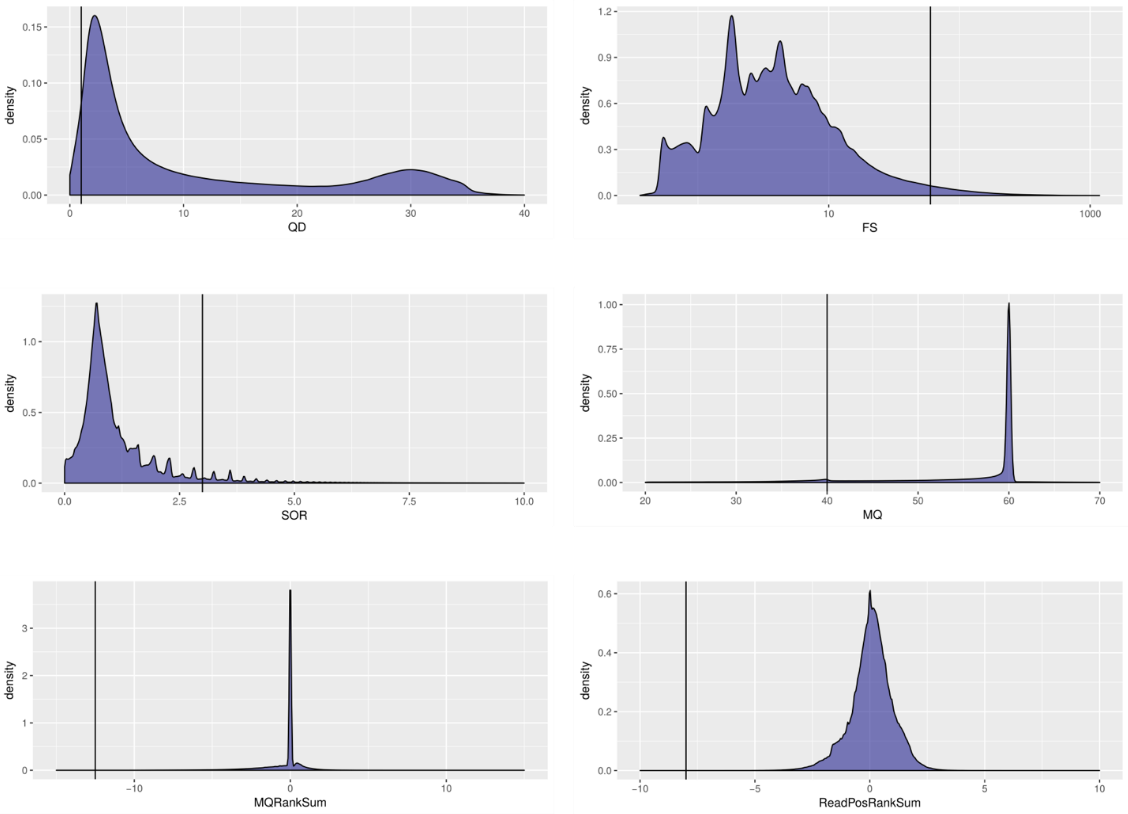


Figure S4.2 Density plots of GATK variant annotations used as reference to determine cutoff values to apply hard filters to raw SNP calls. Plots obtained with the R package *ggplot*. The black vertical line shows the cutoff value used. Abbreviations: QualByDepth (QD) 2.0, FisherStrand (FS) 60.0, StrandOddsRatio (SOR) 3.0, RMSMappingQuality (MQ) 40.0, MappingQualityRankSumTest (MQRankSum) -12.5, ReadPosRankSumTest (ReadPosRankSum) -8.0. A complete explanation of each of these filters can be found in (Broad Institute, 2016).

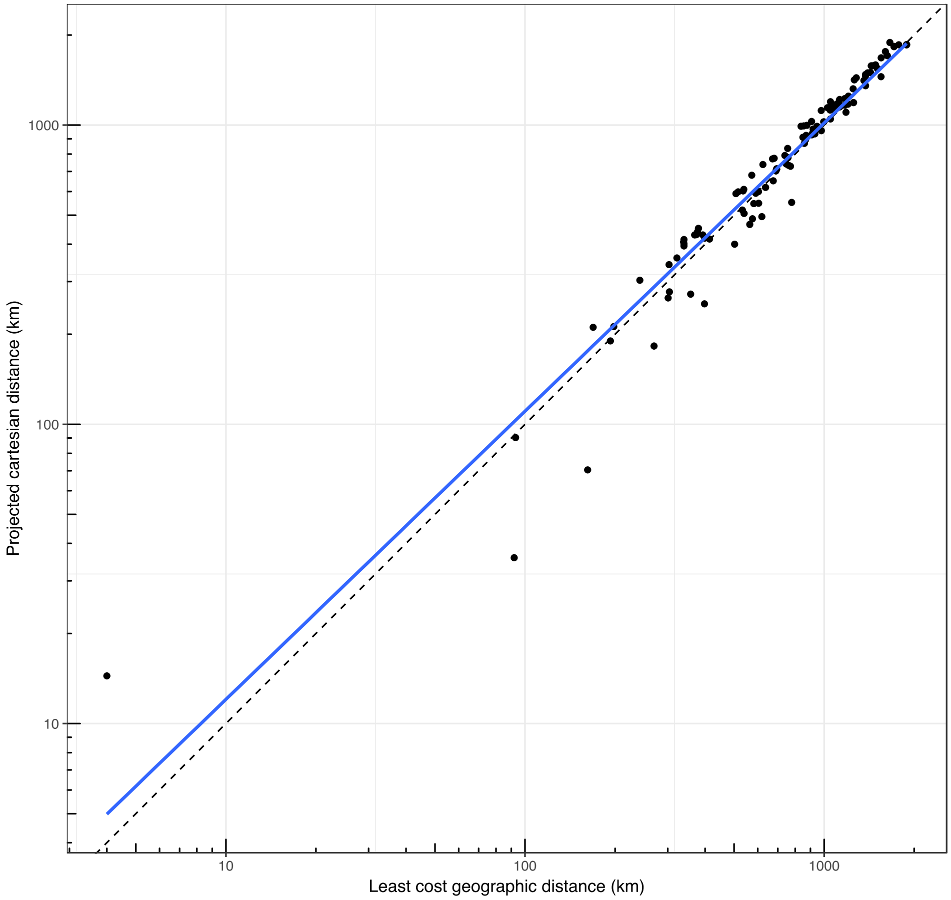


Figure S4.3 Linear regression between Cartesian and geographic coordinates using least-cost distances between 14 Atlantic herring spawning sites in the NW Atlantic. Distances were estimated with the R function *CartDist* (Stanley & Jeffery, 2017). Reprojection stress was good (< 0.05): 0.0448.

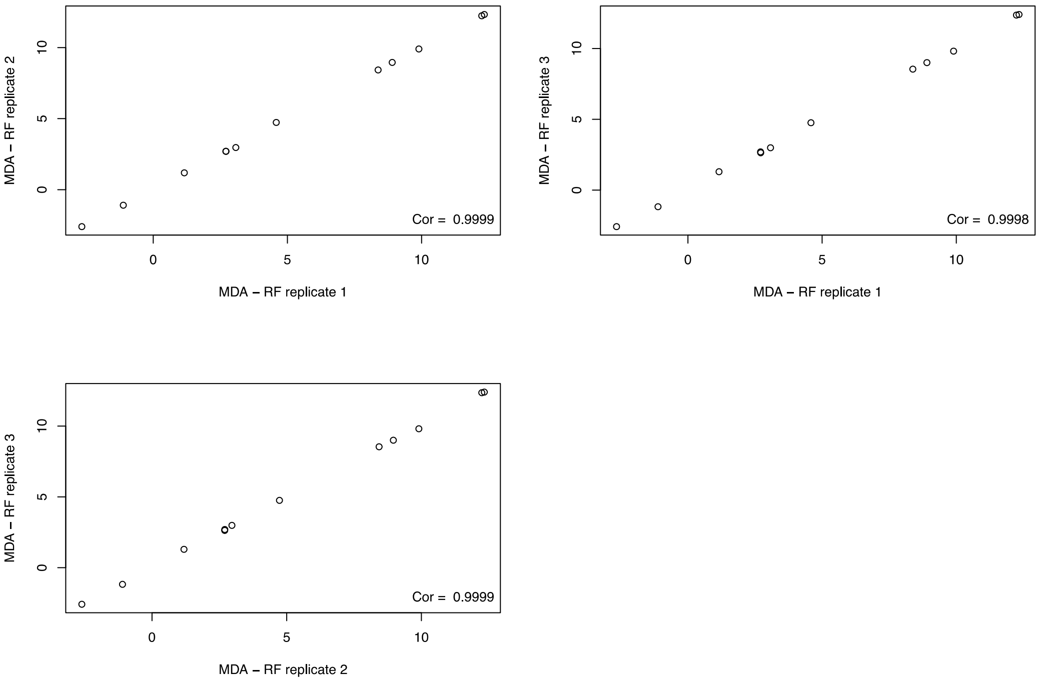


Figure S4.4 Correlation of Mean Decrease in accuracy (MDA) values between pairs of replicates of three random forest regression runs. Plots indicate very high correlation in MDA between runs (Cor = 0.9999).

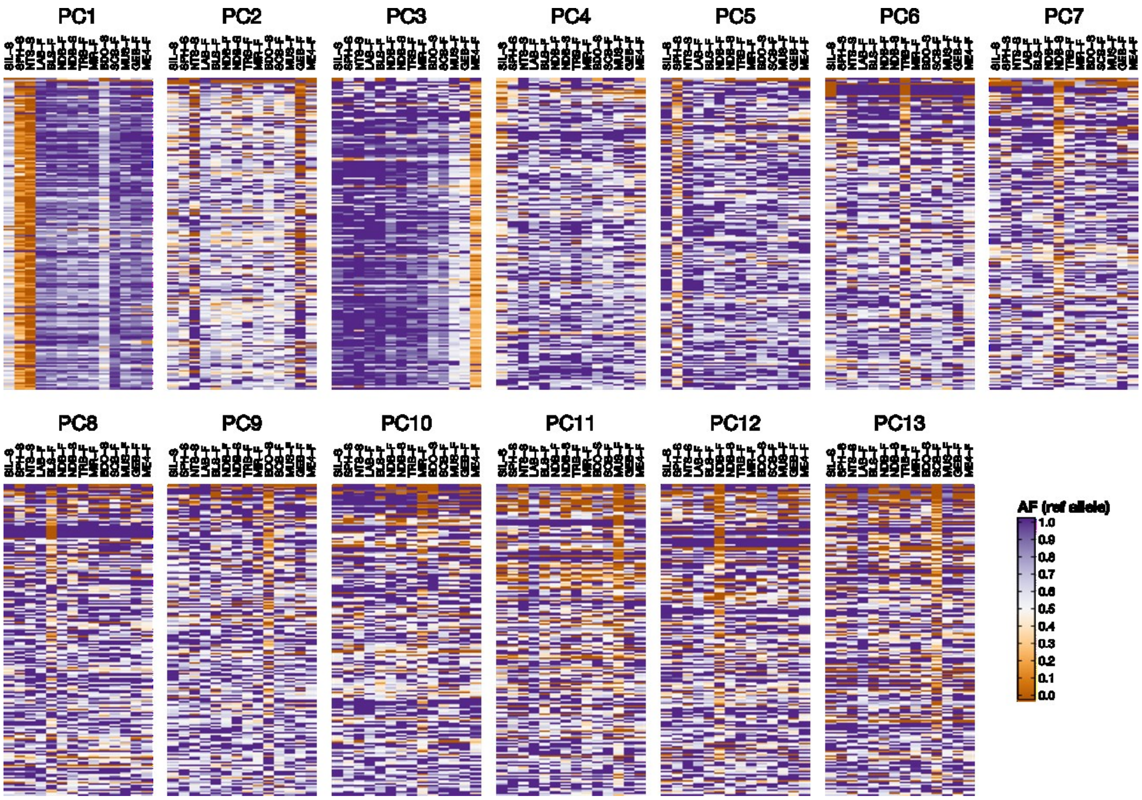


Figure S4.5 Heatmaps representing standardized population allele frequencies (major allele) of the top 200 outlier SNPs detected with *pcadapt* R package (BH-FDR = 0.01) for PCs 1 to 13.

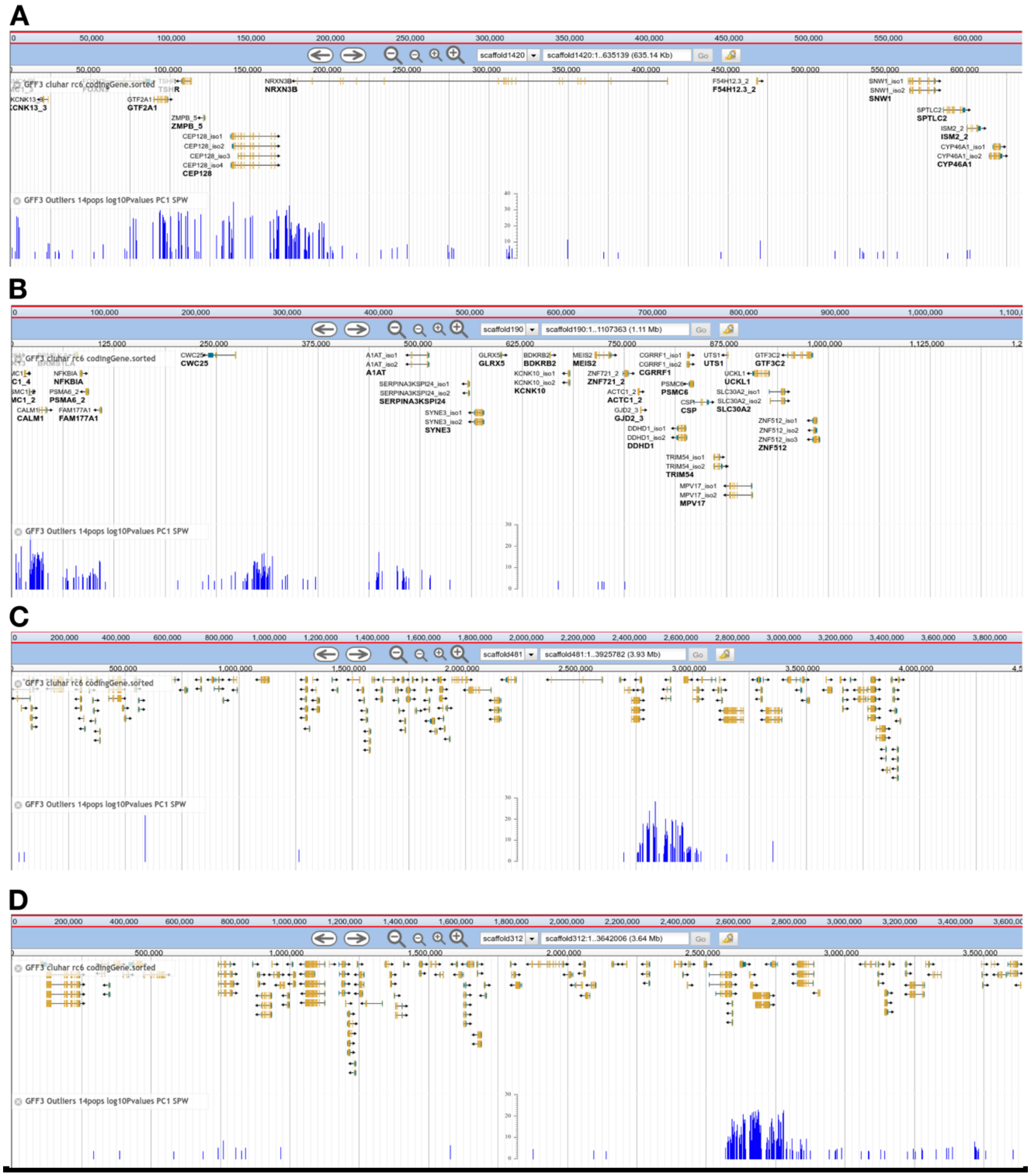

A
scaffold1420
scaffold1420:1..635139 (635.14 Kb)
GFF3 cluhar rc6 codingGene.sorted
NRXN3B
GTF2A1
ZMPB_5
CEP128
F54H12.3_2
SNW1
SPTLC2
ISM2_2
CYP46A1
GFF3 Outliers 14pops log10Pvalues PC1 SPW
B
scaffold190
scaffold190:1..1107363 (1.11 Mb)
CWC25
NFKBIA
PSMA6_2
CALM1
FAM177A1
A1AT
SERPINA3KSPI24
SYNE3
GLRX5
BDKRB2
KCNK10
MEIS2
ZNF721_2
ACTC1_2
GJD2_3
DDHD1
CGRRF1
PSMC6
CSP
UTS1
UCKL1
SLC30A2
TRIM54
MPV17
GTF3C2
ZNF512
GFF3 Outliers 14pops log10Pvalues PC1 SPW
C
scaffold481
scaffold481:1..3925782 (3.93 Mb)
GFF3 cluhar rc6 codingGene.sorted
GFF3 Outliers 14pops log10Pvalues PC1 SPW
D
scaffold312
scaffold312:1..3642006 (3.64 Mb)
GFF3 cluhar rc6 codingGene.sorted
GFF3 Outliers 14pops log10Pvalues PC1 SPW

Figure S4.6 Genomic distribution of outlier loci detected in PC1 in four scaffolds showing the seasonal reproduction pattern of divergence in NW Atlantic herring. Significance values (-log$_{10}$*P*-values) of outlier loci are represented as lines in blue. The position of annotated genes is shown on top of the significance values. **(A)** scaffold 1420, **(B)** scaffold 190, **(C)** scaffold 481, **(D)** scaffold 312. Note the clustering of outlier loci in restricted regions within each scaffold. Images obtained with JBrowse genome browser (Buels et al., 2016).

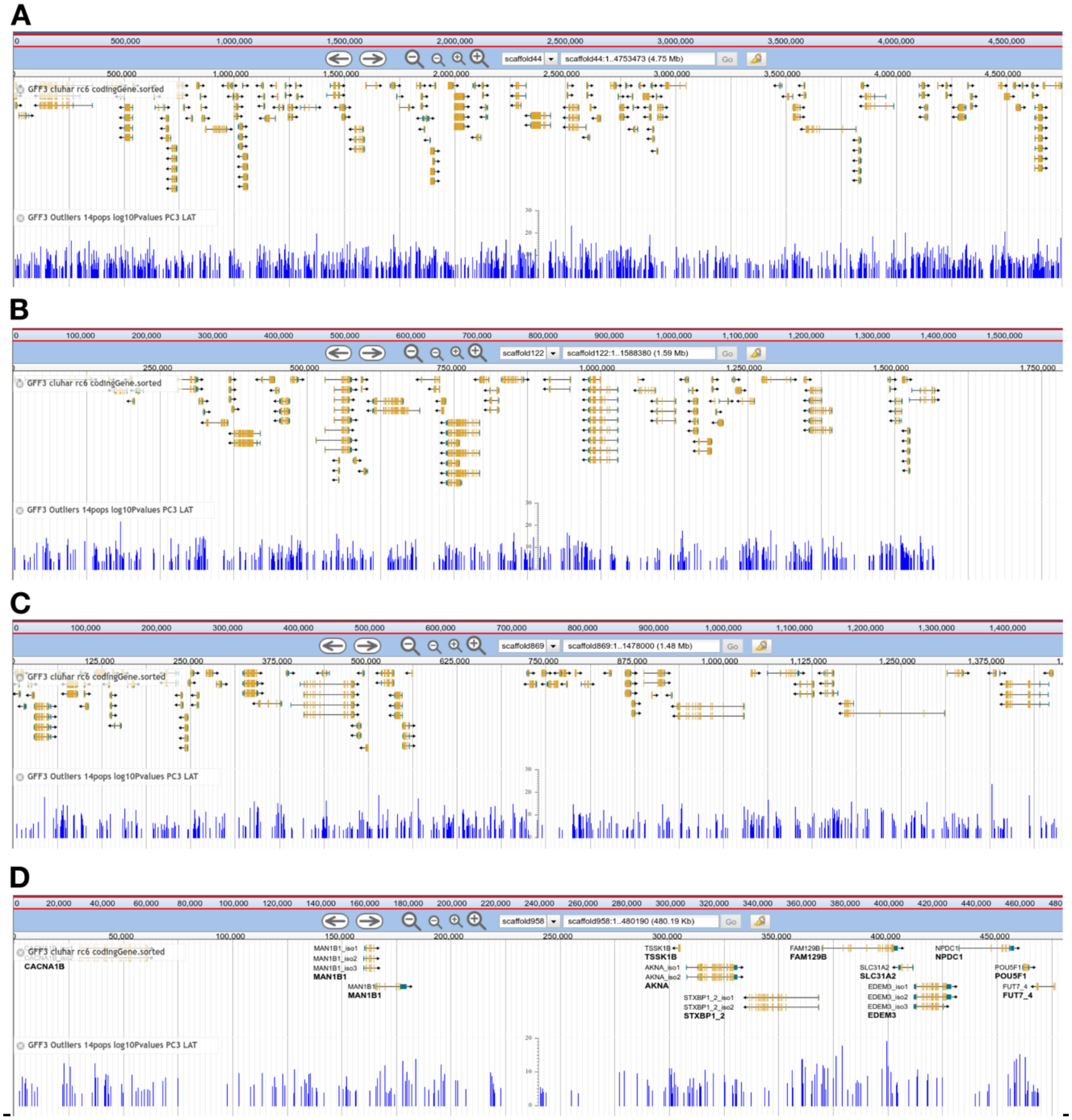

A
scaffold44
scaffold44:1..4753473 (4.75 Mb)
GFF3 cluhar rc6 codingGene.sorted
GFF3 Outliers 14pops log10Pvalues PC3 LAT
B
scaffold122
scaffold122:1..1588380 (1.59 Mb)
GFF3 cluhar rc6 codingGene.sorted
GFF3 Outliers 14pops log10Pvalues PC3 LAT
C
scaffold869
scaffold869:1..1478000 (1.48 Mb)
GFF3 cluhar rc6 codingGene.sorted
GFF3 Outliers 14pops log10Pvalues PC3 LAT
D
scaffold958
scaffold958:1..480190 (480.19 Kb)
GFF3 cluhar rc6 codingGene.sorted
CACNA1B
MAN1B1_iso1
MAN1B1_iso2
MAN1B1_iso3
MAN1B1
TSSK1B
AKNA_iso1
AKNA_iso2
AKNA
STXBP1_2_iso1
STXBP1_2_iso2
STXBP1_2
FAM129B
SLC31A2
EDEM3_iso1
EDEM3_iso2
EDEM3_iso3
EDEM3
NPDC1
POU5F1
FUT7_4
GFF3 Outliers 14pops log10Pvalues PC3 LAT

Figure S4.7 Genomic distribution of outlier loci detected in PC3 from the four scaffolds showing the latitudinal pattern of divergence in NW Atlantic herring. Significance values (-log$_{10}$$P$-values) of outlier loci are represented as lines in blue. The position of annotated genes is shown on top of the significance values. **(A)** scaffold 44, **(B)** scaffold 122, **(C)** scaffold 869, **(D)** scaffold 958. Note the wide distribution of outlier loci across all scaffolds. Images obtained with JBrowse genome browser (Buels et al., 2016).

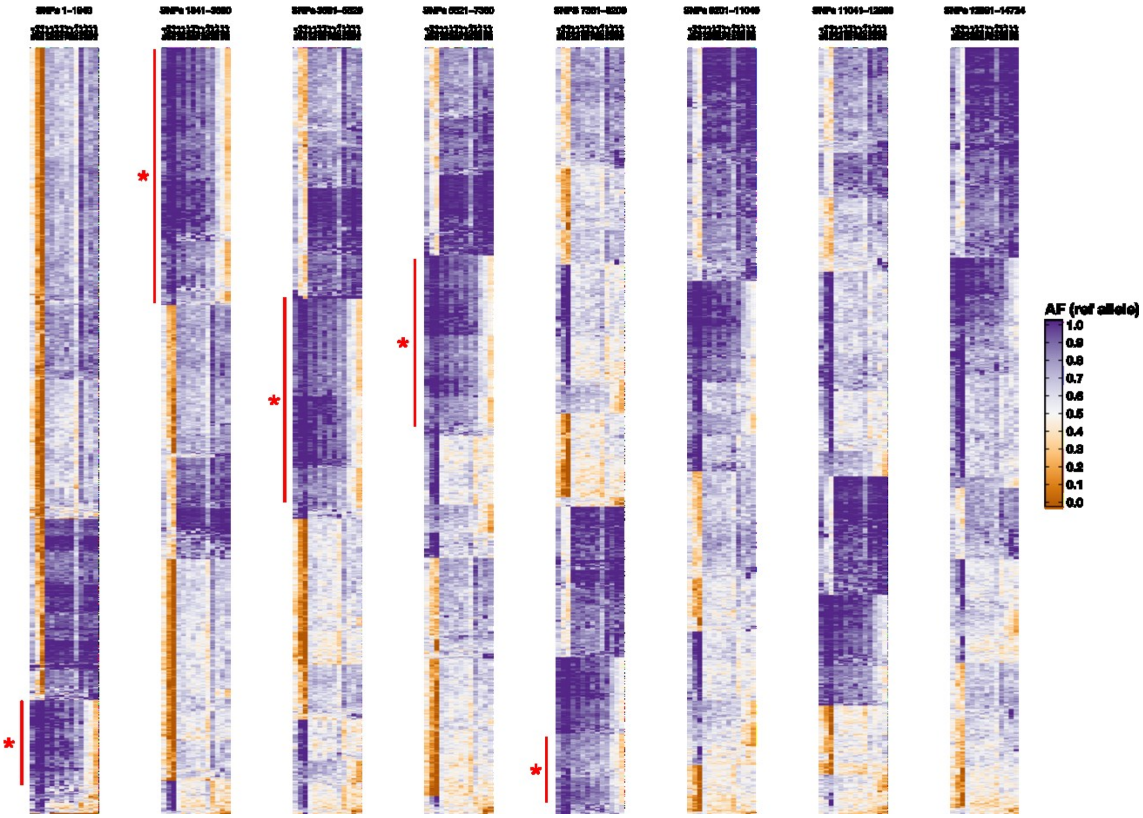


Figure S4.8 Population allele frequencies of the 14,724 outlier SNPs detected in PC1. Extended analysis from Fig. 4.2B. SNPs were ranked by -log$_{10}$*P*-value (descending order) obtained from the genome scan performed with the R package *pcadapt*. The total number of SNPs was split in eight groups for practicality. Clustering by rows (SNPs) was used to facilitate the visualization of blocks of SNPs showing the latitudinal pattern detected in PC3, which are denoted with a red asterisk. Each row is a SNP and each column is a sampling location (SIL-S, SPH-S, NTS-S, LAB-F, BLS-F, NDB-F, NDB-M, TRB-F, MIR-F, BDO-M, SCB-F, MUS-F, GEB-F, ME4-F), abbreviations are as in Table 4.1.

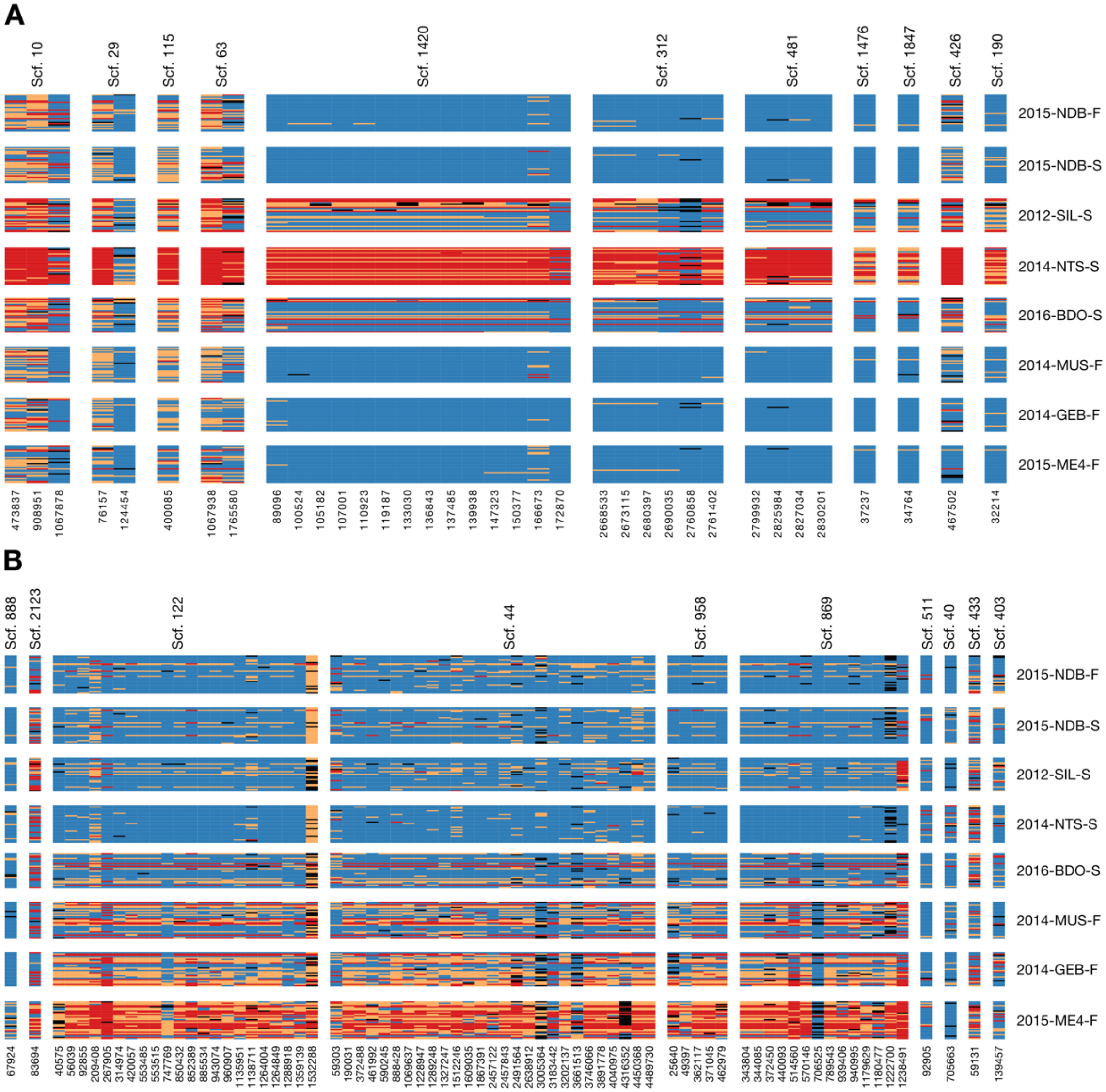


Figure S4.9 Heatmaps depicting standardized individual genotypes for a subset of diagnostic SNPs. A total of 230 individuals (~28 in each location) were genotyped in SNP markers related with **(A)** seasonal reproduction (n=52), and **(B)** latitudinal divergence (n=37). Each row corresponds to an individual and each column corresponds to a single SNP. SNPs from the same scaffold are grouped same as individuals from the same location. Therefore, different scaffolds are separated by vertical white spaces while collections are distinguished by horizontal white spaces. The labels at the top of each block indicate scaffold number (Scf.) whereas the number at the bottom show the corresponding base pair position of the variant. Cell colors represent the individual genotypes, blue for homozygous A/A, yellow for heterozygous, red for homozygous B/B, and black for missing genotype.

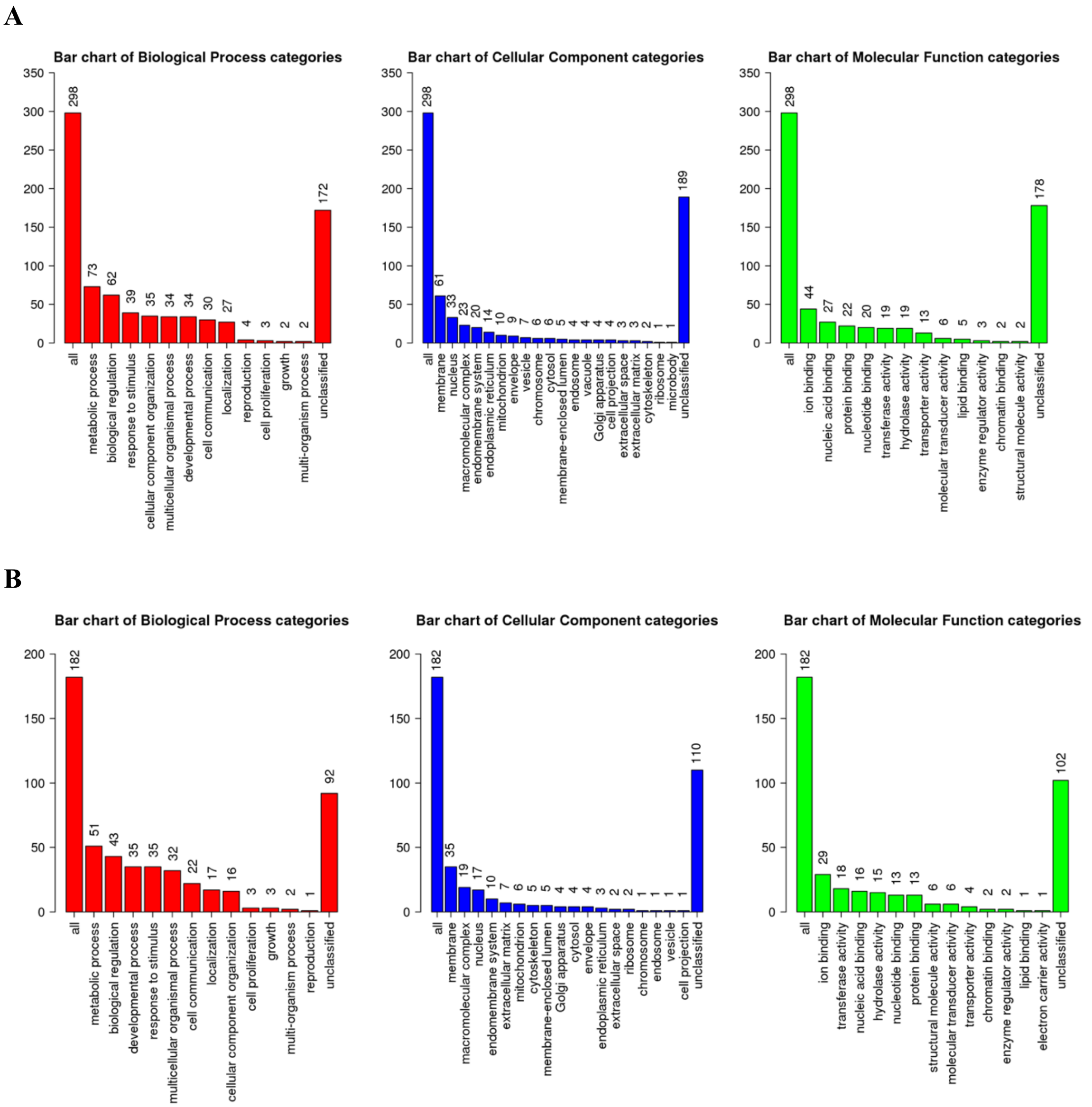


Figure S4.10 Summary of GO terms corresponding to candidate genes associated with seasonal reproduction and with latitudinal divergence in NW Atlantic herring. Each color represents the GO term categories for biological processes (red), cellular components (blue), and molecular function (green). (**A**) For spawning-related genes, and for (**B**) latitude-related genes. Plots generated by Webgestalt (Wang et al., 2013).

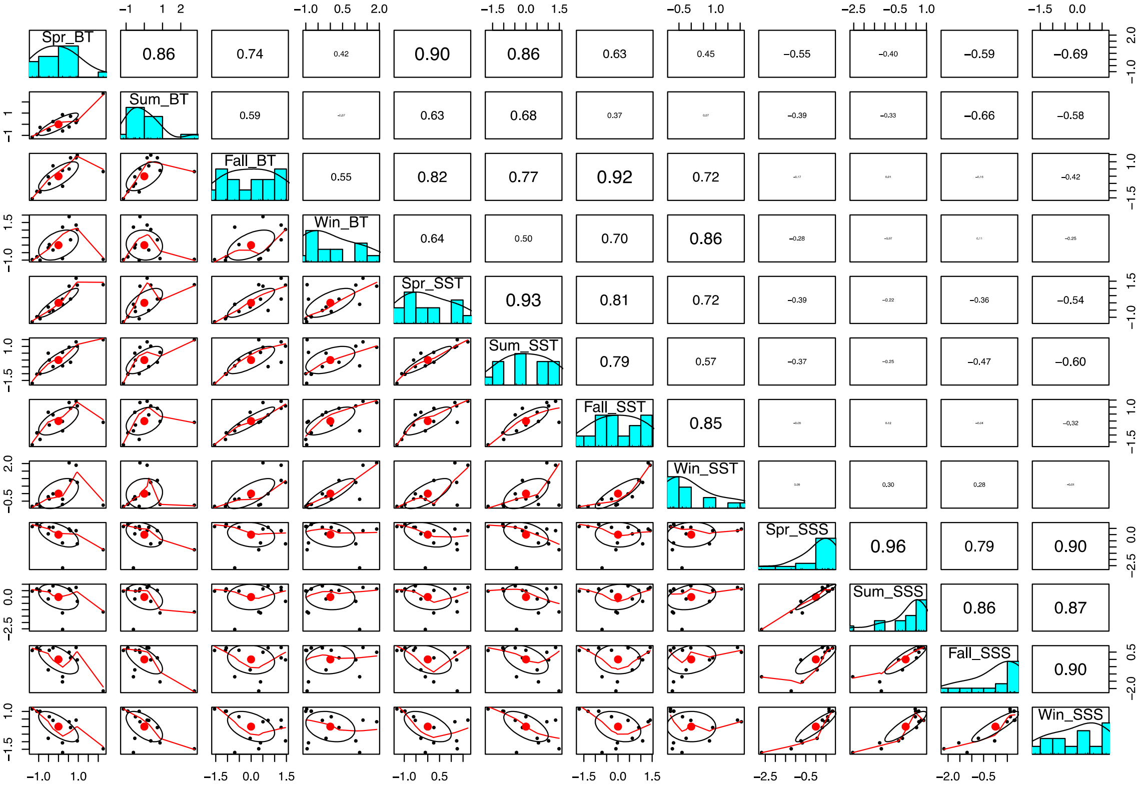


Figure S4.11 Pairwise correlation of 12 environmental variables considered for environment association analyses. Below the diagonal a scatterplot for each pairwise comparison is shown, and above the diagonal the correspondent Pearson correlation coefficients are presented; their font size reflects their magnitude. Plot obtained with the function *pairs.panels* of the R package *psych*.

4.11 References

Allendorf, F. W., Hohenlohe, P. A., & Luikart, G. (2010). Genomics and the future of conservation genetics. *Nature Reviews Genetics*, *11*(10), 697–709. https://doi.org/10.1038/nrg2844

Andersson, L., Ryman, N., Rosenberg, R., & Ståhl, G. (1981). Genetic variability in Atlantic herring (Clupea harengus harengus): description of protein loci and population data. *Hereditas*, *95*(1), 69–78. https://doi.org/10.1111/j.1601-5223.1981.tb01330.x

André, C., Larsson, L. C., Laikre, L., Bekkevold, D., Brigham, J., Carvalho, G. R., … Ryman, N. (2011). Detecting population structure in a high gene-flow species, Atlantic herring (Clupea harengus): direct, simultaneous evaluation of neutral vs putatively selected loci. *Heredity*, *106*(2), 270–280. https://doi.org/10.1038/hdy.2010.71

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Retrieved May 1, 2018, from http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Barrett, R. D. H., & Hoekstra, H. E. (2011). Molecular spandrels: tests of adaptation at the genetic level. *Nature Review Genetics*, *12*(11), 767–780. https://doi.org/10.1038/nrg3015

Baum, B. R. (1989). PHYLIP: Phylogeny Inference Package. Version 3.2 . Joel Felsenstein. *The Quarterly Review of Biology*, *64*(4), 539–541. https://doi.org/10.1086/416571

Baxter, I. G. (1959). Fecundities of winte-spring and summer-autumn herring spawners. *J. Cons. Int. Explor. Mer.*, *25*, 73–80.

Benestan, L., Gosselin, T., Perrier, C., Sainte-Marie, B., Rochette, R., & Bernatchez, L. (2015). RAD genotyping reveals fine-scale genetic structuring and provides powerful population assignment in a widely distributed marine species, the American lobster (Homarus americanus). *Molecular Ecology*, *24*(13), 3299–3315. https://doi.org/10.1111/mec.13245

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B*, *57*(1), 289–300. https://doi.org/10.2307/2346101

Benoît, H., Swain, D., Hutchings, J., Knox, D., Doniol-Valcroze, T., & Bourne, C. (2018). Evidence for reproductive senescence in a broadly distributed harvested marine fish. *Marine Ecology Progress Series*, *592*, 207–224. https://doi.org/10.3354/meps12532

Berg, F., Almeland, O. W., Skadal, J., Slotte, A., Andersson, L., & Folkvord, A. (2018). Genetic factors have a major effect on growth, number of vertebrae and otolith shape in Atlantic herring (Clupea harengus). *PLOS ONE*, *13*(1), e0190995. https://doi.org/10.1371/journal.pone.0190995

Bergland, A. O., Behrman, E. L., O'Brien, K. R., Schmidt, P. S., & Petrov, D. A. (2014).

Genomic Evidence of Rapid and Stable Adaptive Oscillations over Seasonal Time Scales in Drosophila. *PLoS Genetics*, *10*(11), e1004775. https://doi.org/10.1371/journal.pgen.1004775

Bierne, N., Welch, J., Loire, E., Bonhomme, F., & David, P. (2011). The coupling hypothesis: Why genome scans may fail to map local adaptation genes. *Molecular Ecology*, *20*(10), 2044–2072. https://doi.org/10.1111/j.1365-294X.2011.05080.x

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bourne, C., Mowbray, F., Squires, B., & Koen-Alonso, M. (2018). 2017 Assessment of Newfoundland east and south coast Atlantic herring (Clupea harengus) stock complexes. *DFO Can. Sci. Advis. Sec. Res. Doc. 2018/026*, v + 45 p.

Bradbury, I. R., Hubert, S., Higgins, B., Borza, T., Bowman, S., Paterson, I. G., … Bentzen, P. (2010). Parallel adaptive evolution of Atlantic cod on both sides of the Atlantic Ocean in response to temperature. *Proceedings of the Royal Society of London B: Biological Sciences*, *277*(1701), 3725–3734. https://doi.org/10.1098/rspb.2010.0985

Bradbury, I. R., Hubert, S., Higgins, B., Bowman, S., Borza, T., Paterson, I. G., … Bentzen, P. (2013). Genomic islands of divergence and their consequences for the resolution of spatial structure in an exploited marine fish. *Evolutionary Applications*, *6*(3), 450–461. https://doi.org/10.1111/eva.12026

Brickman, D., Hebert, D., & Wang, Z. (2018). Mechanism for the recent ocean warming events on the Scotian Shelf of eastern Canada. *Continental Shelf Research*, *156*, 11–22. https://doi.org/10.1016/j.csr.2018.01.001

Britten, G. L., Dowd, M., & Worm, B. (2016). Changing recruitment capacity in global fish stocks. *Proceedings of the National Academy of Sciences*, *113*(1), 134–139. https://doi.org/10.1073/pnas.1504709112

Broad Institute. (2014). Calling variants on cohorts of samples using the HaplotypeCaller in GVCF mode. Retrieved May 20, 2018, from https://software.broadinstitute.org/gatk/documentation/article.php?id=3893

Broad Institute. (2016). Understanding and adapting the generic hard-filtering recommendations. Retrieved May 20, 2018, from https://gatkforums.broadinstitute.org/gatk/discussion/6925/understanding-and-adapting-the-generic-hard-filtering-recommendations

Broad Institute. (2018). Picard tools. Retrieved May 20, 2018, from http://broadinstitute.github.io/picard/

Bucholtz, R. H., Tomkiewicz, J., & Dalskov, J. (2008). *Manual to determine gonadal maturity of herring (Clupea harengus L.). DTU Aqua-report*. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Manual+to+determine+gonadal+maturity+of+herring+(Clupea+harengus+L.)#0

Buels, R., Yao, E., Diesh, C. M., Hayes, R. D., Munoz-Torres, M., Helt, G., … Holmes,

I. H. (2016). JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biology*, *17*(1), 66. https://doi.org/10.1186/s13059-016-0924-1

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., … Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Cushing, D. H. (1967). The grouping of herring populations. *J. Mar. Biol. Ass., U.K*, *47*, 193–208.

DFO. (1997). *State of the Ocean: Northwest Atlantic*. *DFO Science Stock Status Report G0-01*. Retrieved from https://login.proxy.lib.duke.edu/login?url=https://search.proquest.com/docview/1668269133?accountid=10598%0Ahttp://pm6mt7vg3j.search.serialssolutions.com?ctx_ver=Z39.88-2004&ctx_enc=info:ofi/enc:UTF-8&rfr_id=info:sid/ProQ%3Aasfabiological&rft_val_fmt=info

Dixon, P. (2003). VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science*, *14*(6), 927–930. https://doi.org/10.1111/j.1654-1103.2003.tb02228.x

Dohm, J. C., Lottaz, C., Borodina, T., & Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, *36*(16), e105–e105. https://doi.org/10.1093/nar/gkn425

Dray, S., & Dufour, A.-B. (2007). The ade4 Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software*, *22*(4). https://doi.org/10.18637/jss.v022.i04

Engelhard, G. H., & Heino, M. (2004). Maturity changes in Norwegian spring-spawning herring before, during, and after a major population collapse. *Fisheries Research*, *66*(2–3), 299–310. https://doi.org/10.1016/S0165-7836(03)00195-4

Epstein, D. J. (2009). Cis-regulatory mutations in human disease. *Briefings in Functional Genomics and Proteomics*, *8*(4), 310–316. https://doi.org/10.1093/bfgp/elp021

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Fadista, J., Manning, A. K., Florez, J. C., & Groop, L. (2016). The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *European Journal of Human Genetics*, *24*(8), 1202–1205. https://doi.org/10.1038/ejhg.2015.269

FAO. (2019). Species Fact Sheets: Clupea harengus (Linnaeus, 1758). Retrieved from http://www.fao.org/fishery/species/2886/en

Feder, A. F., Petrov, D. A., & Bergland, A. O. (2012). LDx: Estimation of Linkage Disequilibrium from High-Throughput Pooled Resequencing Data. *PLoS ONE*, *7*(11), e48588. https://doi.org/10.1371/journal.pone.0048588

Fichefet, T., & Maqueda, M. A. M. (1997). Sensitivity of a global sea ice model to the treatment of ice thermodynamics and dynamics. *Journal of Geophysical Research: Oceans*, *102*(C6), 12609–12646. https://doi.org/10.1029/97JC00480

Forester, B. R., Lasky, J. R., Wagner, H. H., & Urban, D. L. (2018). Comparing methods for detecting multilocus adaptation with multivariate genotype–environment associations. *Molecular Ecology*, *27*(9), 2215–2233. https://doi.org/10.1111/mec.14584

Fuentes-Pardo, A. P., Bourne, C., Singh, R., Emond, K., Pinkham, L., McDermid, J. L., … Ruzzante, D. E. (2019). Adaptation to seasonal reproduction and thermal minima-related factors drives fine-scale divergence despite gene flow in Atlantic herring populations. *BioRxiv*. https://doi.org/https://doi.org/10.1101/578484

Gaggiotti, O. E., Bekkevold, D., Jørgensen, H. B. H., Foll, M., Carvalho, G. R., Andre, C., & Ruzzante, D. E. (2009). Disentangling the effects of evolutionary, demographic, and environmental factors influencing genetic structure of natural populations: Atlantic herring as a case study. *Evolution*, *63*(11), 2939–2951. https://doi.org/10.1111/j.1558-5646.2009.00779.x

Goldstein, B. A., Hubbard, A. E., Cutler, A., & Barcellos, L. F. (2010). An application of Random Forests to a genome-wide association dataset: Methodological considerations & new findings. *BMC Genetics*, *11*(1), 49. https://doi.org/10.1186/1471-2156-11-49

Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, *32*(18), 2847–2849. https://doi.org/10.1093/bioinformatics/btw313

Guo, B., Li, Z., & Merilä, J. (2016). Population genomic evidence for adaptive differentiation in the Baltic Sea herring. *Molecular Ecology*, *25*(12), 2833–2852. https://doi.org/10.1111/mec.13657

Hauser, L., & Carvalho, G. R. (2008). Paradigm shifts in marine fisheries genetics: ugly hypotheses slain by beautiful facts. *Fish and Fisheries*, *9*(4), 333–362. https://doi.org/10.1111/j.1467-2979.2008.00299.x

Hendry, A. P., & Day, T. (2005). Population structure attributable to reproductive time: isolation by time and adaptation by time. *Molecular Ecology*, *14*(4), 901–916. https://doi.org/10.1111/j.1365-294X.2005.02480.x

Hivert, V. (2018). Measuring genetic differentiation from Pool-seq data.

Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., … Whitlock, M. C. (2016). Finding the Genomic Basis of Local Adaptation: Pitfalls, Practical Solutions, and Future Directions. *The American Naturalist*, *188*(4), 379–397. https://doi.org/10.1086/688018

Iles, T. D., & Sinclair, M. (1982). Atlantic Herring: Stock Discreteness and Abundance. *Science*, *215*(4533), 627–633. https://doi.org/10.1126/science.215.4533.627

Jeffery, N. W., Bradbury, I. R., Stanley, R. R. E., Wringe, B. F., Van Wyngaarden, M., Lowen, J. Ben, … DiBacco, C. (2018). Genomewide evidence of environmentally mediated secondary contact of European green crab (Carcinus maenas) lineages in eastern North America. *Evolutionary Applications*, *11*(6), 869–882. https://doi.org/10.1111/eva.12601

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., … Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, *484*(7392), 55–61. https://doi.org/10.1038/nature10944

Jorgensen, H. B. H., Hansen, M. M., Bekkevold, D., Ruzzante, D. E., & Loeschcke, V. (2005). Marine landscapes and population genetic structure of herring (Clupea harengus L.) in the Baltic Sea. *Molecular Ecology*, *14*(10), 3219–3234. https://doi.org/10.1111/j.1365-294X.2005.02658.x

Kerr, Q., Fuentes-Pardo, A. P., Kho, J., McDermid, J. L., & Ruzzante, D. E. (2018). Temporal stability and assignment power of adaptively divergent genomic regions between herring (Clupea harengus) seasonal spawning aggregations. *Ecology and Evolution*, (October), ece3.4768. https://doi.org/10.1002/ece3.4768

Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLoS Computational Biology*, *8*(2), e1002375. https://doi.org/10.1371/journal.pcbi.1002375

Kofler, R., Pandey, R. V., & Schlötterer, C. (2011). PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics (Oxford, England)*, *27*(24), 3435–3436. https://doi.org/10.1093/bioinformatics/btr589

Kolaczkowski, B., Kern, A. D., Holloway, A. K., & Begun, D. J. (2011). Genomic Differentiation Between Temperate and Tropical Australian Populations of Drosophila melanogaster. *Genetics*, *187*(1), 245–260. https://doi.org/10.1534/genetics.110.123059

Lamichhaney, S., Barrio, A. M., Rafati, N., Sundstrom, G., Rubin, C.-J., Gilbert, E. R., … Andersson, L. (2012). Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. *Proceedings of the National Academy of Sciences*, *109*(47), 19345–19350. https://doi.org/10.1073/pnas.1216128109

Lamichhaney, S., Fuentes-Pardo, A. P., Rafati, N., Ryman, N., McCracken, G. R., Bourne, C., … Andersson, L. (2017). Parallel adaptive evolution of geographically distant herring populations on both sides of the North Atlantic Ocean. *Proceedings of the National Academy of Sciences*, *114*(17), E3452–E3461. https://doi.org/10.1073/pnas.1617728114

Lehnert, S. J., DiBacco, C., Van Wyngaarden, M., Jeffery, N. W., Ben Lowen, J., Sylvester, E. V. A., … Bradbury, I. R. (2018). Fine-scale temperature-associated genetic structure between inshore and offshore populations of sea scallop (Placopecten magellanicus). *Heredity*, 1–12. https://doi.org/10.1038/s41437-018-0087-9

Lewontin, R. C. (2002). Directions in Evolutionary Biology. *Annual Review of Genetics*, *36*(1), 1–18. https://doi.org/10.1146/annurev.genet.36.052902.102704

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv Preprint ArXiv*, *00*(00), 3. https://doi.org/arXiv:1303.3997 [q-bio.GN]

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-

Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22.

Limborg, M. T., Helyar, S., De Bruyn, M., Taylor, M. I., Nielsen, E. E., Ogden, R., … Bekkevold, D. (2012). Environmental selection on transcriptome-derived SNPs in a high gene flow marine fish, the Atlantic herring (Clupea harengus). *Molecular Ecology*, *21*(15), 3686–3703. https://doi.org/10.1111/j.1365-294X.2012.05639.x

Lischer, H. E. L., & Excoffier, L. (2012). PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, *28*(2), 298–299. https://doi.org/10.1093/bioinformatics/btr642

Luu, K., Bazin, E., & Blum, M. G. B. (2017). pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources*, *17*(1), 67–77. https://doi.org/10.1111/1755-0998.12592

Madec, G., Delecluse, P., Imbard, M., & Levy, C. (1998). *OPA8.1 Ocean general Circulation Model reference manual*. France.

Martinez Barrio, A., Lamichhaney, S., Fan, G., Rafati, N., Pettersson, M., Zhang, H., … Andersson, L. (2016). The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. *ELife*, *5*(MAY2016), 1–32. https://doi.org/10.7554/eLife.12081

McDermid, J. L., Swain, D. P., Turcotte, F., Robichaud, S. A., & Surette, T. (2018). Assessment of the NAFO Division 4T southern Gulf of St. Lawrence Atlantic herring (Clupea harengus) in 2016 and 2017. *DFO Can. Sci. Advis. Sec. Res. Doc. 2018/052*, xiv + 122 p.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

McPherson, A., O'Reilly, P. T., & Taggart, C. T. (2004). Genetic Differentiation, Temporal Stability, and the Absence of Isolation by Distance among Atlantic Herring Populations. *Transactions of the American Fisheries Society*, *133*(2), 434–446. https://doi.org/10.1577/02-106

McPherson, A., Stephenson, R. L., O'Reilly, P. T., Jones, M. W., & Taggart, C. T. (2001). Genetic diversity of coastal Northwest Atlantic herring populations: implications for management. *Journal of Fish Biology*, *59*(SUPPL. A), 356–370. https://doi.org/10.1006/jfbi.2001.1769

McQuinn, I. H. (1997). Metapopulations and the Atlantic herring. *Reviews in Fish Biology and Fisheries*, *7*(3), 297–329. https://doi.org/10.1023/A:1018491828875

Melvin, G. D., Stephenson, R. L., & Power, M. J. (2009). Oscillating reproductive strategies of herring in the western Atlantic in response to changing environmental conditions. *ICES Journal of Marine Science*, *66*(8), 1784–1792.

https://doi.org/10.1093/icesjms/fsp173

Messieh, S. N., Anthony, V., & Sinclair, M. (1985). Fecundities of Atlantic herring Clupea harengus L. populations in the Northwest Atlantic. *ICES C.M. 1985/H:8.*, 22 pp.

Metzger, B. P. H., Duveau, F., Yuan, D. C., Tryban, S., Yang, B., & Wittkopp, P. J. (2016). Contrasting Frequencies and Effects of cis - and trans -Regulatory Mutations Affecting Gene Expression. *Molecular Biology and Evolution*, *33*(5), 1131–1146. https://doi.org/10.1093/molbev/msw011

Nei, M. (1972). Genetic distance between populations. *The American Naturalist*, *1062*(949), 283–292. https://doi.org/10.1086/285153

Nei, M. (2007). The new mutation theory of phenotypic evolution. *Proceedings of the National Academy of Sciences*, *104*(30), 12235–12242. https://doi.org/10.1073/pnas.0703349104

Nosil, P., & Feder, J. L. (2012). Genomic divergence during speciation: causes and consequences. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1587), 332–342. https://doi.org/10.1098/rstb.2011.0263

Nosil, P., Funk, D. J., & Ortiz-Barrientos, D. (2009). Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, *18*(3), 375–402. https://doi.org/10.1111/j.1365-294X.2008.03946.x

O'Connor Lab. (2016). ScaffoldStitcher. Retrieved May 1, 2018, from https://bitbucket.org/dholab/scaffoldstitcher/src

Okonechnikov, K., Conesa, A., & García-Alcalde, F. (2015). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, *32*(2), btv566. https://doi.org/10.1093/bioinformatics/btv566

Overholtz, W. . (2002). The Gulf of Maine–Georges Bank Atlantic herring (Clupea harengus): spatial pattern analysis of the collapse and recovery of a large marine fish complex. *Fisheries Research*, *57*(3), 237–254. https://doi.org/10.1016/S0165-7836(01)00359-9

Pagani, F., & Baralle, F. E. (2004). Genomic variants in exons and introns: identifying the splicing spoilers. *Nature Reviews Genetics*, *5*(May), 389–396.

Palumbi, S. R. (1994). Genetic Divergence, Reproductive Isolation, and Marine Speciation. *Annual Review of Ecology and Systematics*, *25*(1), 547–572. https://doi.org/10.1146/annurev.es.25.110194.002555

Panagiotou, O. A., & Ioannidis, J. P. A. (2012). What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *International Journal of Epidemiology*, *41*(1), 273–286. https://doi.org/10.1093/ije/dyr178

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Auwera, G. A. Van der, … Banks, E. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv*, 1–22. https://doi.org/https://doi.org/10.1101/201178

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Rambaut, A. (2007). FigTree. Retrieved May 20, 2018, from http://tree.bio.ed.ac.uk/software/figtree/

Ravinet, M., Faria, R., Butlin, R. K., Galindo, J., Bierne, N., Rafajlović, M., … Westram, A. M. (2017). Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. *Journal of Evolutionary Biology*, *30*(8), 1450–1477. https://doi.org/10.1111/jeb.13047

Revelle, W. (2018). psych: Procedures for Personality and Psychological Research. Northwestern University, Evanston, Illinois, USA. Retrieved from https://cran.r-project.org/package=psych

Rousset, F. (1997). Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics*, *145*(April), 1219–1228.

Sambrook, J. & Russel D.W. (2006) Purification of Nucleic Acids by Extraction with Phenol:Chloroform. *Cold Spring Harb Protoc. doi:10.1101/pdb.prot4455*

Schlötterer, C., Tobler, R., Kofler, R., & Nolte, V. (2014). Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. *Nature Reviews Genetics*, *15*(11), 749–763. https://doi.org/10.1038/nrg3803

Schluter, D. (2009). Evidence for Ecological Speciation and Its Alternative. *Science*, *323*(5915), 737–741. https://doi.org/10.1126/science.1160006

Scott, W. B., & Scott, M. G. (1988). *Atlantic fishes of Canada. Canadian Bulletin of Fisheries and Aquatic Sciences, bulletin 219*. Toronto, CA: University of Toronto Press.

Simmonds, E. J. (2007). Comparison of two periods of North Sea herring stock management: success, failure, and monetary value. *ICES Journal of Marine Science*, *64*(4), 686–692. https://doi.org/10.1093/icesjms/fsm045

Sinclair, M. (1988). Marine Populations: an Essay on Population Regulation and Speciation (p. 252). Seattle: Washington Sea Grant/Univ. Wash. Press.

Sinclair, M., & Iles, T. D. (1989). Population regulation and speciation in the oceans. *J. Cons. Int. Explor. Mer*, *45*, 165–175.

Smith, P. J., & Jamieson, A. (1986). Stock discreteness in herrings: a conceptual revolution. *Fish. Res.*, 223–234.

Stanley, R. R. E., DiBacco, C., Lowen, B., Beiko, R. G., Jeffery, N. W., Van Wyngaarden, M., … Bradbury, I. R. (2018). A climate-associated multispecies cryptic cline in the northwest Atlantic. *Science Advances*, *4*(3), eaaq0929. https://doi.org/10.1126/sciadv.aaq0929

Stanley, R. R. E., & Jeffery, N. W. (2017). CartDist: Re-projection tool for complex marine systems. https://doi.org/10.5281/zenodo.802875

Stephenson, R. L., Melvin, G. D., & Power, M. J. (2009). Population integrity and connectivity in Northwest Atlantic herring: a review of assumptions and evidence. *ICES Journal of Marine Science*, *66*(8), 1733–1739. https://doi.org/10.1093/icesjms/fsp189

Sylvester, E. V. A., Beiko, R. G., Bentzen, P., Paterson, I., Horne, J. B., Watson, B., … Bradbury, I. R. (2018). Environmental extremes drive population structure at the northern range limit of Atlantic salmon in North America. *Molecular Ecology*, *27*(20), 4026–4040. https://doi.org/10.1111/mec.14849

Tavares, H., Whibley, A., Field, D. L., Bradley, D., Couchman, M., Copsey, L., … Coen, E. (2018). Selection and gene flow shape genomic islands that control floral guides. *Proceedings of the National Academy of Sciences*, *115*(43), 11006–11011. https://doi.org/10.1073/pnas.1801832115

Teacher, A. G., André, C., Jonsson, P. R., & Merilä, J. (2013). Oceanographic connectivity and environmental correlates of genetic structuring in Atlantic herring in the Baltic Sea. *Evolutionary Applications*, *6*(3), 549–567. https://doi.org/10.1111/eva.12042

Tigano, A., & Friesen, V. L. (2016). Genomics of local adaptation with gene flow. *Molecular Ecology*, n/a-n/a. https://doi.org/10.1111/mec.13606

Townsend, D. W., Thomas, A. C., Mayer, L. M., Thomas, M. A., & Quinlan, J. A. (2004). Oceanography of the Northwest Atlantic Shelf (1, W). In A. R. Robinson & K. H. Brink (Eds.), *The Sea: The Global Coastal Ocean: Interdisciplinary Regional Studies and Syntheses* (pp. 1–57). Harvard University Press.

Turner, S. D. (2014). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *BiorXiv*. https://doi.org/https://doi.org/10.1101/005165

Turner, T. L., Hahn, M. W., & Nuzhdin, S. V. (2005). Genomic Islands of Speciation in Anopheles gambiae. *PLoS Biology*, *3*(9), e285. https://doi.org/10.1371/journal.pbio.0030285

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., … DePristo, M. A. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. In *Current Protocols in Bioinformatics* (Vol. 11, p. 11.10.1-11.10.33). Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/0471250953.bi1110s43

Vitti, J. J., Grossman, S. R., & Sabeti, P. C. (2013). Detecting Natural Selection in Genomic Data. *Annual Review of Genetics*, *47*(1), 97–120. https://doi.org/10.1146/annurev-genet-111212-133526

Wang, J., Duncan, D., Shi, Z., & Zhang, B. (2013). WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Research*, *41*(W1), W77–W83. https://doi.org/10.1093/nar/gkt439

Wang, Z., Brickman, D., Greenan, B. J. W., & Yashayaev, I. (2016). An abrupt shift in the Labrador Current System in relation to winter NAO events. *Journal of Geophysical Research: Oceans*, *121*(7), 5338–5349. https://doi.org/10.1002/2016JC011721

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, *38*(6), 1358. https://doi.org/10.2307/2408641

Wheeler, J. P., & Winters, G. . (1984). Homing of Atlantic herring (Clupea harengus harengus) in Newfoundland waters as indicated by tagging data. *Can. J. Fish. Aquat. Sci.*, *41*, 108–117.

Wiberg, R. A. W., Gaggiotti, O. E., Morrissey, M. B., & Ritchie, M. G. (2017). Identifying consistent allele frequency differences in studies of stratified populations. *Methods in Ecology and Evolution*, *8*(12), 1899–1909. https://doi.org/10.1111/2041-210X.12810

Wu, C.-I. (2001). The genic view of the process of speciation. *Journal of Evolutionary Biology*, *14*(6), 851–865. https://doi.org/10.1046/j.1420-9101.2001.00335.x

Yeaman, S., & Whitlock, M. C. (2011). The genetic architecture of adaptation under migration-selection balance. *Evolution*, *65*(7), 1897–1911. https://doi.org/10.1111/j.1558-5646.2011.01269.x

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, *1*(1), 3–14. https://doi.org/10.1111/j.2041-210X.2009.00001.x

CHAPTER 5. MIXED COMPOSITION OF NORTHWEST ATLANTIC HERRING AGGREGATIONS OUTSIDE OF THE BREEDING SEASON REVEALED BY DIAGNOSTIC SNP PANELS

5.1 Abstract

The preservation of biologically relevant genetic diversity within species is often overlooked in fisheries management, imposing a risk for the long-term persistence of species and fisheries. The still high cost of genomic techniques is one of the limiting factors to integrate genetics into the conservation toolbox. Here we developed and demonstrated the utility of two highly informative reduced SNP panels for the identification of spawning season and latitudinal origin of northwest Atlantic herring (*Clupea harengus* L.); a highly migratory and abundant marine fish with an intricate life history and elusive definition of population structure. With these panels, we genotyped 1010 fish from 30 locations distributed across the reproductive range of the species in the northwest Atlantic. Self-assignment simulations confirmed the high accuracy of both panels (>85%). Our genetic data confirmed the southwest-northeast gradient in the prevalence of spawning types and provided evidence that reproductive time diversity may be an adaptive strategy to cope with changing environments. Temporal stability in allele frequency differences between spring and fall spawners and in fall spawners from the northern region was confirmed for the short time period covered (1-9 years), implying selective pressures may be relatively constant in ecological time scales. The analysis of hybrid indices disclosed evidence of unrestricted hybridization between reproductive and latitudinal components. Hybrids between spawning seasons represented a varying proportion of spawning aggregations (0.0-60.0%), whereas hybrids between latitudinal regions were more prevalent towards the south (21.4-55.6%), suggesting high connectivity between regions and polymorphism is advantageous in this region. Interestingly, individuals with intermediate to high admixture levels seem to be capable of spawning in either season. Analysis of mixture samples confirmed the dynamic nature of aggregations outside of the breeding season. This work highlights the complexity of the herring mating system and the importance of preserving such intraspecific diversity. Additionally, it provides a novel genetic tool with which it is possible to estimate, in

close to real time, the relative contribution of reproductive and latitudinal components to herring aggregations of presumed mixed origin sampled either during or outside the spawning seasons.

5.2 Introduction

One often overlooked consideration in fisheries management is the presence of biologically relevant genetic diversity within and between populations of the same species (Bernatchez et al., 2017; Laikre et al., 2016; Reiss, Hoarau, Dickey-Collas, & Wolff, 2009). The preservation of such biological complexity is particularly important for commercially harvested species, because it determines their resilience to extreme environmental changes and to the increased mortality due to fishing (Hilborn, Quinn, Schindler, & Rogers, 2003; Ruzzante et al., 2006; Satterthwaite & Carlson, 2015). Yet, the intraspecific diversity of numerous harvested organisms remains largely unknown. This knowledge gap can risk the long-term persistence of both species and fisheries, and in turn, of the ecosystem functions and livelihood of communities that depend on them (Reiss et al., 2009; Schindler et al., 2010). For example, uninformed fisheries targeting mixed stocks could negatively affect adaptive or less-abundant components by making them more vulnerable to overfishing (Frank & Brickman, 2000; Ruzzante et al., 2000). The resulting loss of species' biological diversity and evolutionary potential could then translate into longer time for recovery of fishing stocks or their complete depletion (Frank & Brickman, 2000; Melvin et al., 2009). Consequently, the protection of genetic diversity within species should be a central goal in wildlife conservation and fisheries management, as recognized by international agreements such the Convention on Biological Diversity (CBD; https://www.cbd.int/).

Intraspecific genetic diversity has been traditionally examined using a handful of neutral genetic markers (i.e. that do not codify for proteins), which commonly reflect patterns of differentiation due to genetic drift (i.e. random loss of alleles due to small population size) (Allendorf, 2016). With the development of Next-Generation Sequencing technologies is now possible to assess both neutral and functional genetic diversity in high genomic resolution (i.e. with thousands to millions of genetic markers)

(Nosil & Feder, 2012). Recently, the screening of markers in functional parts of the genome is revealing fine-scale structuring in species characterized by low levels of population structure at neutral markers [e.g. Atlantic herring (Fuentes-Pardo et al., 2019; Lamichhaney et al., 2012; Martinez Barrio et al., 2016), Atlantic salmon (Freamo, O'reilly, Berg, Lien, & Boulding, 2011), pearl oyster (Nayfa & Zenger, 2016)]. The still high cost of high-throughput sequencing however, limits conducting large-scale genomic studies (Shafer et al., 2015), often required for fisheries management. Highly informative and reduced panels of microsatellite (i.e. short tandem repeat) or single nucleotide polymorphism (SNP) markers are becoming an affordable alternative with equivalent high accuracy for individual assignment and mixed stock analysis (Bekkevold et al., 2015b; Bradbury et al., 2016; Jeffery, Wringe, et al., 2018), which is why its development has attracted great interest.

Atlantic herring is an abundant and highly migratory pelagic fish that plays an important role in the marine ecosystem as a forage species. It also sustains a profitable fishery across the North Atlantic Ocean (FAO, 2019). Herring reaches first maturation at the age of 3-4 years and can live up to 20+ years (Benoît et al., 2018). Broadcast spawning occurs once a year at predictable and discrete locations in coastal and offshore waters near shore. Tagging data indicates herring exhibits strong spawning site fidelity, once a ground has been used before (Stobo, 1987; Wheeler & Winters, 1984b). Differential trends in abundance and growth rate suggest herring may also exhibit "natal homing" (i.e. return to the place where they were hatched), although this has not been proven yet (Stephenson et al., 2009). After hatching, larvae remain grouped in areas with particular oceanographic characteristics (Iles & Sinclair, 1982). Juveniles and adults undertake annual migrations among spawning, overwintering, and feeding areas, forming large schools often composed by diverse populations that separate during the breeding season (Waters & Clark, 2005; Wheeler & Winters, 1984b). In the Northwest (NW) Atlantic, the species is widely distributed from Cape Hatteras to northern Labrador (Scott & Scott, 1988), but its reproduction is restricted from Cape Cod to northern Newfoundland, in U.S. and Canada, respectively (Iles & Sinclair, 1982). Spawning takes place from April to October but mainly between April and May (spring-spawning) and from August to October (fall-spawning). For management purposes, fish are currently classified as spring or fall spawners mainly based on their gonadal maturity stage at the month of capture. Individuals spawning before July 1st are considered spring spawners and, after this date they are assumed fall spawners (LeBlanc, Poirier, MacDougall, Bourque, & Roy, 2008; I. H. McQuinn, 1987). The prevalence of either spawning type appears to follow a southwest-northeast gradient, where fall-spawning prevails in the south (including the Scotian Shelf, the Bay of Fundy, and the Gulf of Maine), while spring- and fall-spawning coexist in the north (including the Gulf of St. Lawrence and Newfoundland) (Melvin et al., 2009). Fishing data indicates that spring-spawning used to be prevalent in the northernmost part of the reproductive range (Winters & Wheeler, 1987), it was common in the brackish waters in Bras D'Or lake until its collapse due to overfishing in the 90's (Denny, Clark, Power, & Stephenson, 1998), and it was sporadically present in mid-coastal Nova Scotia (Bradford & Iles, 1992; Power et al., 2007). The shift in the prevalence of reproductive strategies observed in recent years appears to be linked with a sea warming trend, which has led to propose that spawn timing diversity may be an adaptive strategy of the species to increase reproductive success under changing environments (Melvin et al., 2009). The high biocomplexity (sensu Ruzzante et al., 2006) of NW Atlantic herring has historically made the description of its population structure difficult. More recent genomic studies revealed fine-scale population structuring in adaptive loci, despite low differentiation at neutral loci (Fuentes-Pardo et al., 2019; S. Lamichhaney et al., 2017). Specifically, contrasting allele frequencies in adaptive loci discriminate spawning aggregations by spawning season and latitudinal origin.

Here, we evaluated the efficacy of highly informative reduced SNP panels for mixed stock assessment. Based on the genomic data available for herring, we developed two SNP panels diagnostic of spawning season (SPW-panel) and latitudinal origin (LAT-panel), and obtained their genotypes for herring individuals from several spawning grounds throughout the reproductive range of the species in the NW Atlantic. The efficacy of the panels for individual assignment and mixed stock analysis was examined using samples from inshore and offshore mixed aggregations in the region. The two overarching questions were: i) What is the relative contribution of spring- and fall-spawning and of northern and southern individuals to offshore aggregations of presumed mixed origin and to spawning aggregations? ii) Is this contribution temporally stable? This study constitutes the most comprehensive spatial genetic survey of spawning grounds and offshore aggregations in the reproductive range of the species in western Atlantic to date. The accurate assessment of the composition of mixed stocks, achieved here using genetic tools, is critical for the ongoing management of herring fisheries.

5.3 Materials and Methods

5.3.1 Sample collection and DNA extraction

A total of 1010 adult herring were collected between 2005 and 2017 from 25 locations ranging from Newfoundland, Canada, to the Gulf of Maine, U.S. (Fig. 5.1 and Table 1). Twenty-two of these samples corresponded to inshore spawning aggregations (i.e. sites with known recurrent spawning activity over the years) and three were offshore aggregations (i.e. sites where individuals from different populations are known to gather outside of the breeding season). Individual gonadal maturity was recorded when possible (Fig. 5.1) (Maturity code equivalences, 1: Juvenile, 2: Early maturation, 3: Mid maturation, 4: Late maturation, 5: Spawning capable, 6: Spawning, 7: Spent, 8: Resting). To assess stability of allele frequencies over time, we obtained additional temporal replicates one, two to nine years apart from five locations, for a total of 30 samples available for the genetic study. For individual assignment analysis, collections from spawning sites were considered "baselines" and collections from offshore locations and from the Bras D'Or lake area were analyzed as "mixture" samples (see Table 5.1). Muscle or fin tissue samples were taken from each individual and stored in 95% ethanol at -20 ºC. DNA was isolated from the tissues following a standard phenol chloroform protocol. DNA integrity was evaluated with 0.8% agarose gel electrophoresis using 0.5x TBE buffer and a 1Kb molecular weight ladder. DNA concentration in ng/µL was measured using the Quant-iT PicoGreen dsDNA assay (Thermo Fisher Scientific) and a Cytation4 plate reader (BioTek Instruments, Inc.).

5.3.2 Development of diagnostic SNP panels

In our previous study based on whole genome resequencing of pooled DNA of individuals (Pool-seq; Fuentes-Pardo et al., 2019), we identified in NW Atlantic herring thousands of putatively adaptive outlier SNPs exhibiting contrasting allele frequency differences with respect to two main axes of population divergence: seasonal reproduction and a latitudinal climate-related cline (14 724 and 6 595, respectively). We thus aimed to find a reduced number of genetic markers that allow for the correct assignment of individuals to their spawning season (hereafter called SPW-panel and SPW-dataset), and geographic region of origin (hereafter called LAT-panel and LAT-dataset). For this, we separately examined the informativeness of outlier SNPs underlying each pattern of population divergence using random forest (RF) (Breiman, 2001), a powerful machine learning algorithm widely used for ranking features (genetic markers in this case) based on their importance for classification (Jeffery, Wringe, et al., 2018; Sylvester et al., 2017). RF builds a collection of decision trees ("forest") from randomly chosen subsets of samples ("in the bag") and explanatory variables (or predictors) to predict the outcome of a response variable. RF models can be used for classification (response variable set *as.factor*), when the response variable is categorical, or for regression, when the response variable is continuous. In classification RF models, the incorrect assignment of out-of-bag (OOB) samples constitutes an estimate of the classification power of the model (OOB error rate, OOB-ER); and the mean decrease in accuracy (MDA) of the model caused by the exclusion (or permutation) of a predictor is a measure of the relative importance of the predictor for correct assignment. Model accuracy is thus expected to decrease when an important variable is permuted, and variables with the highest MDA score are considered the most important for classification. A more detailed explanation of RF can be found elsewhere (Breiman, 2001; Brieuc, Waters, Drinan, & Naish, 2018; Goldstein et al., 2010; Schrider & Kern, 2018). In our case, we used SNP allele frequencies as explanatory variables, and the classes (or categories) of each SNP panel as response variables. The classes in the SPW-panel were spring and fall, and in the LAT-panel were north (southern Labrador, Newfoundland, and the Gulf of St. Lawrence), center (the Bay of Fundy and the Scotian Shelf), and south (Gulf of Maine).

For the RF runs, we simulated 50 individual genotypes per population based on the population-level allele frequencies obtained with Pool-seq for the outlier SNPs underlying each pattern of divergence. This simulation was performed using the function *sample.geno* implemented in the R package *pcadapt* v3.0.4 (Luu et al., 2017). This function samples genotypes using binomial random draws with the provided population allele frequencies. To avoid over-confidence of assignment accuracy (upward grading bias) (Anderson, 2010), the simulated genotype dataset was split into "training" and "testing" (hold-out) datasets (proportion 60:40). The training dataset was solely used for building the RF model and the testing dataset was used for cross-validation of the assignment power of a reduced number of SNP markers. We used the implementation of the RF algorithm in the R package *randomForest* (Liaw & Wiener, 2002; Andy Liaw & Wiener, 2018). In this package it is necessary to estimate three parameter values to build a RF model: (1) *mtry*, or the number of predictors randomly selected in each node of a tree; (2) *ntree*, or the total number of trees to build in a model; and (3) *sampsize*, or the number of samples per class included for building a classification model. The *sampsize* parameter assures a "balanced" representation of classes (i.e. same number of observations per class) during tree building, which avoids the RF model to be biased towards the majority class. To select optimal parameters for our dataset, we examined a combination of values following the recommendations of Brieuc et al. (2018) with some modifications. First, OOB-ER for different combinations of *mtry* and *ntree* values were compared. Values of *mtry* comprised default [sqrt(*p*)], twice default [2*sqrt(*p*)], half default [0.5*sqrt(*p*)], 0.1(*p*), 0.2(*p*), *p*/3, and *p*, where *p* is the number of loci. Values of *ntree* ranged from 100 to 1000, by increments of 100. The optimal *mtry* corresponded to the value where OOB-ER reached a plateau. Second, to ensure repeatability of RF models and convergence of importance values of the predictors, three independent classification RF runs were performed using the optimal *mtry* and a large *ntree* value (500 000 and 1 000 000). A correlation coefficient (*Cor*) was calculated for the importance values of each pair of RF replicates. The optimal *ntree* was the value with the highest *Cor*. Finally, the parameter *sampsize* was set to 2/3 of the class with the lower sample size (Brieuc et al., 2018).

The most informative SNPs for classification in each SNP panel were identified following the "elbow" method (Goldstein et al., 2010). In this method, the cut-off of importance values is determined based on a scatter plot of importance values of SNPs; predictors before the point where the differences between importance values level-off ("elbow") are considered important. To determine the minimum number of informative loci required in each panel for accurate assignment (>70%) of individuals to their class, we performed individual assignment simulations using the leave-one-out (LOO) method implemented in the R package *rubias* (Anderson, Waples, & Kalinowski, 2008; Moran & Anderson, 2018) (*self_assign* function). We tested the top 10, 20, 40, 70, and 100 loci, and calculated metrics such as accuracy (the proportion of correctly assigned individuals to their class), efficiency (the proportion of the total number of individuals in a class that were correctly assigned), and power (power = efficiency x accuracy) (Vähä & Primmer, 2006) using custom R scripts. Finally, we applied additional filters to the reduced panel of top loci: (1) loci with >2 flanking SNPs within ±150bp were excluded, as high SNP abundance within a short DNA fragment suggests the presence of a repetitive sequence that could lead to a spurious SNP and difficulties for primer design; (2) if two or more informative variants were located within the same scaffold and were <1 Kb apart, only the variant with the highest importance value was kept in order to reduce redundancy in the final panel.

5.3.3 SNP genotyping and quality filtering of raw data

Individual genotypes of the loci included in the final SNP panel were obtained with the MassARRAY System (Agena Bioscience) through the service provided by *Neogen Corporation* (Lincoln, U.S.). About 6% of the total individuals analyzed (60/1010) were genotyped twice to estimate the genotyping error rate following the method described in Pompanon, Bonin, Bellemain, & Taberlet (2005). The multilocus genotype error rate was equal to 0.16%. Individuals and loci with more than 20% missing data and loci with a minor allele frequency (MAF) below 1% were removed from the dataset using PLINK

(Purcell et al., 2007). Loci that did not show expected allele frequency patterns were also excluded.

5.3.4 Population structure

To quantify genetic divergence among populations based on each SNP panel, we computed pairwise fixation index F_{ST} with ARLEQUIN v3.5 (Excoffier & Lischer, 2010). To control for multiple comparisons, the correspondent *P*-values were adjusted for false discovery rate (FDR) using the R package *p.adjust* (R Core Development Team, 2019). An alpha significance value of 0.05 was used. To estimate the proportion of the total genetic variation that is explained by genetic differences between groups (F_{ST}), among subpopulations within groups (F_{SC}), and among individuals within subpopulations (F_{CT}), we performed a hierarchical analysis of molecular variance (AMOVA) for each SNP panel using ARLEQUIN v3.5 (Excoffier & Lischer, 2010). In the SPW-panel, subgroups were spring and fall spawning collections, and in the LAT-panel, subgroups corresponded to collections from the north and south regions. Population structure was examined using ADMIXTURE (Alexander, Novembre, & Lange, 2009), a program that uses a maximum likelihood method to estimate individual ancestry values from SNP genotype data that are equivalent to the Q-value from the program STRUCTURE (Pritchard, Stephens, & Donnelly, 2000).

5.3.5 Hybrid detection between reproductive and latitudinal adaptive components

To investigate the presence of hybrids between the different genetic lineages identified with the SPW- and LAT-panels, we obtained a hybrid index (Hindex) (Buerkle, 2005), or admixture proportion estimate, for each individual and SNP panel using the maximum-likelihood calculation implemented in GENODIVE (Meirmans & Van Tienderen, 2004). The calculation assumes that each locus presents independent information. For non-independent loci, as it is in our case, the calculation will result in excess of confidence, but the point estimate of the hybrid index will be correct (Alex Buerkle comm. pers.).

In addition, to assess phenotype-genotype correspondence of the SPW-panel, we compared individual gonadal maturities with hybrid indices in spring- and fall-spawning aggregations where >60% of individuals were mature (i.e. capable or actively spawning, stages 5 and 6, respectively, see Table 5.1) at the time of capture. For comparison, we also performed this analysis in aggregations composed by a mixture of individuals in various gonadal maturity stages.

5.3.6 Assessment of the predictive power of SNP panels

The accurate determination of genetic group membership of individuals from mixture samples relies on the correct definition of baselines (or "reporting units"), as they are the basis for the development of classification models that are used to perform individual assignment (Chen et al., 2018). Hence, for each SNP panel separately, we examined the self-assignment accuracy of spawning grounds to two proposed baselines using simulations performed with the R package *assignPOP* (Chen et al., 2018). We did not use traditional programs for individual assignment and mixed stock analysis like *rubias* (Moran & Anderson, 2018) or *ONCOR* (Kalinowski, Manlove, & Taper, 2007), because they assume independency between loci and ours are in LD. We used *assignPOP* instead, because it includes a PCA-based data dimensionality reduction step when performing assignment tests. The new variables (PCs) used for building training models should summarize the overall variance of the input data. In this regard, the loci in LD should not influence or are less likely to influence assignment results (Alex Kuan-Yu Chen pers. comm.). The proposed baselines for SPW-dataset were, option 1: all inshore spring- and fall-collected samples, option 2: as option 1 but excluding SIL12S, SMB15S and PLB16S (sites with a mix of spring and fall spawners). The proposed baselines for LAT-dataset were, option 1: all inshore northern and southern sites with respect to the genetic break point identified for herring in northwest Nova Scotia at ~45°N (Fuentes-Pardo et al., 2019), option 2: all inshore northern and the southernmost sites (ME415F, M314F) (Fig. S5.1). Data grouping and renaming was done with the R package *genepopedit* (Stanley, Jeffery, Wringe, DiBacco, & Bradbury, 2017) and file format conversions were achieved with PGDspider (Lischer & Excoffier, 2012).

Self-assignment simulations of individuals from baselines consisted of four main steps. Firstly, individuals from each baseline group were randomly divided into training and testing datasets of different sizes using the Monte-Carlo cross-validation method (Xu

& Liang, 2001) implemented in the function *assign.MC*. The training dataset sizes evaluated corresponded to the number of individuals from each baseline representing the 50%, 70%, and 90% of the baseline with smaller sample size. The remaining individuals in each case were assigned to the testing dataset. This resampling and dataset subdivision procedure therefore, assured the obtention of unbiased training datasets of equal size (i.e. "balanced"), which avoids upwardly biased estimates of accuracy commonly seen in "unbalanced" datasets (Anderson, 2010). Secondly, for each training set, three subsets of high-F_{ST} SNP loci (or training features) were tested, top 25%, 50% and 100%. The dimensionality of the different training datasets was reduced using PCA. Thirdly, the resulting dataset from PCA was then used for building a classification model with the default machine learning algorithm in *assingPOP*, which is Support Vector Machines (SMV) from the R package *e1071* (Meyer et al., 2015). And fourthly, assignment of individuals in the testing dataset to a baseline was then performed based on the classification model built with the training dataset. The whole process was repeated 1000 times. Assignment accuracy estimates therefore correspond to the proportion of correctly assigned individuals to their known baseline group for 1000 iterations.

5.3.7 Individual assignment of mixture samples

Assignment probability to a baseline of individuals from mixture samples was calculated using the function *assign.X* from the R package *assignPOP* (Chen et al., 2018). For each SNP panel, individual assignment was performed based on the SMV classification model built with the proposed baselines showing the overall highest assignment accuracy (>70%).

5.4 Results

5.4.1 SNP panel development and genotyping

We used RF for ranking outlier SNPs discovered from Pool-seq data, which distinguish NW Atlantic herring populations by their spawning season and broad geographic region of origin respect to a latitudinal climate-related cline. Pilot runs performed for parameter optimization of the RF algorithm indicated that any of the *mtry* values tested results in a low OOB-ER when more than 300 trees are built (*ntree* >300) (Fig. S5.2A, B). We thus used the default *mtry* for subsequent runs. A comparison of MDA values between replicate runs indicated that a *ntree* of 500,000 and 1,000,000 ascertains a very high convergence of importance values (MDA) (e.g. for *ntree* = 1,000,0000, SPW-panel: *Cor* = 0.9959 ± 0.00006, Fig. S5.3A; LAT-panel: *Cor* = 0.9982 ± 0.0001, Fig. S5.4A). A scatterplot of importance values of SNPs in the SPW- (Fig. S5.3B) and LAT-datasets (Fig. S5.4B) revealed that about 500 loci were highly informative. The top 10, 20, 40, 70, and 100 SNPs from each dataset were then used for testing their individual assignment power to predetermined groups (SPW-panel: spring or fall, LAT-panel: north, center, and south) through simulations using the leave-one-out (LOO) method implemented in the R package *rubias* (Anderson et al., 2008; Moran & Anderson, 2018). For the SPW dataset, all the panel sizes examined produced an accuracy, efficiency, and overall power of 100%, indicating that as few as 10 SNPs are enough to determine the spawning season of an individual with an accuracy of 100% (Fig. S5.3C). For the LAT-dataset, 10 SNPs also suffice to determine the broad geographic origin of individuals, for an overall power, accuracy and efficiency > 85%; however, a minimum of 70 SNPs are required to reach an accuracy of 100% (Fig. S5.4C). Given the possibility that genotyping of some of the top SNPs could be suboptimal (e.g. challenging primer design, poor PCR amplification, etc.), we decided to develop primers for the top 40 SNPs from the SPW-dataset and for the top 80 SNPs from the LAT-dataset that fulfilled the requirements described in Materials and Methods.

A total of 1010 individuals were successfully genotyped in 110 SNPs (SPW-panel: 36, LAT-panel: 74). Of those, 993 individuals and 77 SNP loci (SPW-panel: 25, LAT-panel: 52) passed quality filters (missing rate < 20% and MAF > 1%) and constituted the dataset used for further analyses (Table S5.1). The individual genotype data confirms the observations made form Pool-seq data (Fuentes-Pardo et al., 2019) of close to fixation of opposite alleles between spring and fall spawners and northern and the southernmost sites (Fig. S5.5, Fig. S5.6).

5.4.2 Population structure

For SPW-panel, pairwise F_{ST} between pure spring and fall spawners (as determined by gonadal maturity status at the time of capture, Fig. 5.4 as reference, and excluding mixture samples) varied between 0.242 and 0.910. Some F_{ST} estimates were non-significant at $\alpha = 0.05$ after False Discovery Rate (FDR) correction for multiple testing (Table S5.2). For LAT-panel, pairwise F_{ST} between the northern and southern spawners (excluding mixture samples) varied between 0.000 and 0.703. The majority of F_{ST} estimates were significant at $\alpha = 0.05$ using a False Discovery Rate (FDR) correction for multiple testing (Table S5.3). For SPW-panel, AMOVA showed that 73.4% of the total genetic variance was explained by spawning season, 4.6% by differences between aggregations within the same spawning type, and 22.1% among individuals within aggregations (Table S5.4). All fixation indices were significant using 10,100 permutations. Global F_{ST} between spawning types was high ($F_{\mathrm{ST}} = 0.78$, *P-value* < 0.0001), F_{SC} between spawning aggregations of the same spawning type was lower ($F_{\mathrm{SC}} = 0.17$, *P-value* < 0.0001), and F_{CT} between individuals within spawning aggregations was also high ($F_{\mathrm{CT}} = 0.73$, *P-value* < 0.0001). For LAT-panel, AMOVA showed that 47.2% of the total genetic variance was explained by the north-south genetic break along the latitudinal cline, 4.3% was explained by differences between aggregations within the same latitudinal region, and 48.6% by variation within spawning grounds (Table S5.5). All fixation indices were also significant with 10,100 permutations. Global F_{ST} between the north-south regions along the latitudinal cline was high ($F_{\mathrm{ST}} = 0.51$, *P-value* < 0.0001), F_{SC} between spawning aggregations of the same region was lower ($F_{\mathrm{SC}} = 0.08$, *P-value* < 0.0001), and F_{CT} between individuals within spawning aggregations was also high ($F_{\mathrm{CT}} = 0.47$, *P-value* < 0.0001).

In the SPW-dataset, the analysis of individual ancestry performed with ADMIXTURE revealed a localized distribution of two lineages, one corresponding to spring spawners and the other to fall spawners (Fig. 5.2A, B). The spring-spawning lineage was present only in the Gulf of St. Lawrence and southwest Newfoundland and, in a smaller proportion, in an offshore sample in Musquodoboit, mid-coastal Nova Scotia. In contrast, the fall-spawn lineage was widely distributed across the study area.

Individuals with intermediate admixture coefficients were present across sites but more commonly in spring spawners than in fall spawners. Among spring-spawning sites, SIL12S (northwest Gulf of St. Lawrence) and SMB15S (southwest Newfoundland) were aggregations comprising a mix of spring, fall, and admixed individuals (Fig. 5.2B). Interestingly, all individuals from SIL12S were in ripening and spawning condition whereas individuals in SMB15S had different gonadal maturity stages (Fig. 5.1). This observation suggests different population dynamics may be occurring at these sites. Temporal replicates of spawning aggregations (FTB14S-16S, NTS05S-14S, PEI05F-14F) and of one offshore mixed aggregation (OMU14F-15F) indicated allele frequency stability over the short time span covered (1, 2 to 9 years). Exception to this observation was BDO16S-17S, where significant allele frequency variation was observed between samples collected one year apart. Given that individuals in this sample were in different maturity condition and their genotype assigned them to different spawning seasons (Fig. 5.1, 2B), we infer different populations meet at this location (see also Kerr, Fuentes-Pardo, Kho, McDermid, & Ruzzante, 2018). The two offshore aggregations from southwest Newfoundland showed varying composition of spring and fall lineages, indicating diverse populations may reunite at these sites.

ADMIXTURE results of the LAT-dataset disclosed a latitudinal distribution of two distinct genetic groups, the northern and the southern lineages (Fig. 5.3A, B). Individuals with intermediate admixture coefficients were observed across sites but predominantly in the southern region, being more abundant on the Scotian Shelf (MUS14F, GEB14F) and the Gulf of Maine (ME415F, ME314F), followed by the Bay of Fundy (SCB15F) and the mixed aggregations from Bras D'Or lake (BDO16S-17S) (Fig. 5.3B). Interestingly, SCB15F had a greater proportion of northern genotypes than the other southern sites, and a few individuals with the "southern" genotypes were present in the north (e.g. FTB15F, TRB14F) and vice versa (e.g. MUS14F, GEB14F, ME314F). All the temporal replicates, including spawning (FTB14S-16S, NTS05S-14S, PEI05F-14F), inshore (BDO16S-17S) and offshore (OMU14F-15F) aggregations, showed remarkable allele frequency stability between samples collected 1, 2 to 9 years apart. It is important to note though, that no temporal replicates were analyzed for the spawning aggregations from the southern region, except for one offshore aggregation from mid-coastal Nova Scotia (OMU14F-

15F); thus, short term temporal stability of allele frequencies was only confirmed for the northern region. The two offshore aggregations from southwest Newfoundland showed different genetic composition; some individuals in OWN15S exhibited the southern genotypes while all the individuals in OEN15S had the northern genotypes. This observation reinforces the idea that different populations come together at these offshore locations.

5.4.3 Hybrid detection between reproductive and between latitudinal adaptive components

In the SPW-dataset, individual Hindex varied among sites of the same spawning type (spring or fall), ranging from 0% to 10% (0%: pure spring spawners, 100%: pure fall spawners). This variation in hybrid indices suggests the existence of first generation (Hindex 0.5) and recombinant hybrids (Hindex <0.5 and >0.5). In general, hybrids (i.e. offspring between parents of the same species but belonging to different populations, defined by 0.1 < Hindex <0.9) constituted a varying proportion of the total number of individuals collected from a given spawning aggregation (0.0-60.0%) and were more prevalent among spring (13.3-60.0%) than fall spawners (0.0-26.7%) (Fig. S5.7A, Fig. 5.2C).

The comparison of gonadal maturities and Hindex values in spring- (Fig. 5.4A) and fall-spawning (Fig. 5.4B) aggregations where most individuals were capable of, or actively spawning (gonads in stage 5-6), confirmed phenotype-genotype agreement in pure spring (Hindex 0-10%) and pure fall (Hindex 90-100%) spawners. This comparison also showed that individuals with Hindex <68.2% often spawn in spring while individuals with Hindex >53.8% commonly spawn in fall, suggesting that individuals with Hindex ~54-69% can either spawn in spring or fall. Moreover, contrary to expectations, this analysis revealed that in some cases capable/actively spawning adult individuals could be misclassified to a reproductive season, when solely gonadal maturity and the month of capture are considered. These are the two main criteria currently used for individual spawning season designation in the NW Atlantic (i.e. spawning individuals between January 1st and before July 1st are considered spring-spawners and after this date are designated fall spawners) (LeBlanc et al., 2008; McQuinn, 1987). For example, in

SIL12S (Fig. 5.4A) most individuals were spawning (stage 5-6) and a few spent (stage 7) in June. If the current criterion of spawning season designation were used, SIL12S would have been classified as a spring-spawning sample. In contrast, genetic data revealed that pure spring and fall spawners as well as hybrids were together at this site. In PLB16S some individuals were in late maturation (stage 4) and others in spawning condition (stage 5-6). From this we could have inferred individuals in stage 4 are likely fall spawners, and the ones in stage 5-6 are spring spawners. The genetic data indicated that some individuals in stage 4 were pure fall spawners, others pure spring spawners and hybrids. In NDB15F, there were also individuals in stages 4 and 5-6. The genetic data showed all individuals were fall spawners. The same comparison for sites comprising individuals at various maturity stages (Fig. 5.4C) confirmed the increased difficulty to assign individuals to a spawning season only based on gonadal maturation and date of collection, especially for non-actively spawning adult individuals (stages 3-4 and 7-8). For instance, in FTB14S there were individuals spawning and in late maturation (stage 4). As this sample was collected in May, individuals would have been classified as spring spawners. Genetic data indicated one of the maturing individuals was a pure fall spawner and some of the maturing and spawning individuals were hybrids. In SMB15S, there was a mix of spawning, mid-maturation (stage 3) and spent (stage 7) individuals when the sample was collected in June. With the conventional criteria, this sample would have been classified as a mix of spring and fall spawners. Genetic data indicated that some of the spawning individuals were pure spring, pure fall or hybrids, and that the mid maturation and spent individuals were likely fall spawners or hybrids.

In the LAT-dataset, Hindex values varied among sites following a latitudinal gradient, ranging from 0 to 100% in some cases (0%: pure south, 100%: pure north); for example, in TRB14F, MUS14F, GEB14F, ME314F, ME415F (Fig. 5.3C). In general, hybrids (i.e. offspring between parents of the same species but belonging to different populations, defined by $0.1 <$ Hindex <0.9) were observed across most sites but were predominant in the south, between BDO16S-17S and ME415F, where they appear to represent 21.4-55.6% (0.0-20.0% in the north) (Fig. S5.7B) of spawning aggregations. Interestingly, most hybrid individuals across sites exhibited a Hindex around 50%, suggesting they likely result from a relatively recent hybridization event or polymorphism at these loci is advantageous.

5.4.4 Assessment of the predictive power of SNP panels

Simulations performed with the R package *assignPOP* (Chen et al., 2018) indicated that in the SPW-dataset, the two proposed baselines provided a high assignment accuracy (>85%) across spawning classes for all training dataset sizes (50%, 70%, 90% of the smallest class) and train loci proportions tested (top 25%, 50%, 100% high-F_{ST} SNPs) (Fig. 5.5A, Fig. S5.8A). Assignment accuracy was high and similar between spawning classes, just slightly lower and with more variation for the spring class, likely due to the presence of a higher number of heterozygote individuals in this class. The proposed baseline that excluded mixed locations (option 2) (Fig. 5.5A, right) however, exhibited the best performance of the two, providing an assignment accuracy >95% for both spawning classes (Fall: Median 0.96, 25% quartile 0.96, 75% quartile 0.97; Spring: Median 0.98, 25% quartile 0.97, 75% quartile 1.00). In the LAT-dataset, assignment accuracy was generally high but was lower and more variable in the southern than in the northern class. The exclusion of mixed samples from proposed baselines also resulted in a substantial improvement of assignment accuracy, particularly for the southern class, for all training dataset sizes and train loci proportions evaluated (Fig. 5.5B) (North: Median 0.92, 25% quartile 0.92, 75% quartile 0.93; Maine: Median 0.86, 25% quartile 0.83, 75% quartile 0.90). For example, assignment accuracy of the south class in proposed baseline option 1 was ~65% (Fig. S5.8B), and it increased to >85% when only including the southernmost sites in proposed baseline option 2 (Fig. 5.5B). Overall, these results indicate that as few as 6 and 13 highly informative SNPs from the SPW- and LAT-panel, respectively, ascertain a high assignment accuracy >85%, confirming previous results of self-assignment baseline simulations based on Pool-seq data.

5.4.5 Individual assignment of mixture samples

We assigned individuals from one inshore (BDO16S-BDO17S) and three offshore (OEN15S, OWN15S, OMU14F-OMU15F) mixture samples, to a spawning season and geographic region using the classification models built with the best performing proposed baselines for each SNP panel (option 2 in both cases). Individual assignment tests disclosed the dynamic nature of these mixed aggregations.

The two spring-collected samples from Bras D'Or lake consisted of temporal replicates one year apart (BDO16S-17S) (Fig. 5.6A, B). In 2016, some pure spring and fall spawners as well as hybrids met at this site. Spring spawners and hybrids had the "northern" genotype, which suggests they likely came from the Gulf of St. Lawrence/Newfoundland/South Labrador (GNL) region. Some of the fall spawners had the southern and others the northern genotypes, indicating they possibly came from the Scotian shelf/Bay of Fundy/Gulf of Maine (SFM) and the GNL regions, respectively. In contrast, in 2017, all individuals captured were fall spawners and only one was likely a hybrid; most of the individuals came from GNL and two likely came from SFM. The offshore mixed aggregations sampled in spring 2015 at two sites in southern Newfoundland (OEN15S, OWN15S) exhibited different composition (Fig. 5.6C, D). In OEN15S, all individuals were pure fall spawners except one that was hybrid; all originated from GNL. In contrast, in OWN15S, pure spring- and fall-spawners and hybrids were present. The spring spawners and hybrids originated from GNL, while most of the fall spawners came from GNL and two came from SFM. The two fall-collected samples offshore Musquodoboit, mid-coastal Nova Scotia, comprised temporal replicates one year apart (OMU14F-15F) (Fig. 5.6E, F). The majority of individuals in these samples were fall spawners coming from GNL and SFM, and a few were pure spring spawners and hybrids from GNL. The composition of this aggregation appeared to be stable between these two temporal replicates. Overall, these results confirm the composition of inshore and offshore mixture aggregations in NW Atlantic herring is highly dynamic, implying mixing of different reproductive and geographic components at these sites, the proportions of which can be estimated with our methods.

5.5 Discussion

Here we developed two highly informative reduced SNP panels and demonstrate their utility for the genetic identification of spawning season (SPW-panel) and broad latitudinal origin (LAT-panel) of Atlantic herring, a highly migratory and abundant marine pelagic fish characterized by an intricate life history and elusive definition of population structure. With these genetic tools, we examined the spatial and temporal distribution of distinct reproductive and latitudinal genetic components in numerous spawning aggregations across the reproductive range of the species in the NW Atlantic.

5.5.1 SNP panel development and assignment accuracy power

The SNP panels consisted in a few dozens of the thousands outlier SNPs putatively under selection discovered from Pool-seq data in our previous study (Fuentes-Pardo et al., 2019). In that study, outlier SNPs revealed that population subdivision in NW Atlantic herring is mainly driven by temporal divergence due to seasonal reproduction, and by spatial divergence along a latitudinal climate-related cline defining a north-south genetic break. Opposite alleles close to fixation distinguished the different genetic groups found in each axis of genetic divergence (spring- and fall-spawning; north and southernmost regions, respectively). Individual genotype data generated here confirmed those observations. SNPs included in the final panels were selected following a genetic marker ranking procedure based on the importance values of loci for classification given by a random forest (RF) model. The RF model was built using the population allele frequencies obtained from Pool-seq data. Our results demonstrate that the top RF-ranked SNPs included in the panels are highly informative, which confirms the utility of machine learning algorithms for genetic marker selection for classification (Jeffery, Wringe, et al., 2018; Sylvester et al., 2017). Further self-assignment simulations based on SNP-panel genotypes of 933 individuals from inshore spawning herring aggregations, confirmed the high assignment accuracy of both panels, averaging 0.95% and 0.85% for the SPW-panel and LAT-panel, respectively. The slightly lower accuracy in the LAT-panel likely results from the high presence of intermediate allele frequencies in the southern region.

5.5.2 Spatial distribution of reproductive and latitudinal components

We observed a southwest-northeast gradient in the prevalence of spawning types across the reproductive range of the species in the NW Atlantic. This result is in agreement with previous observations based on fishery data (Melvin et al., 2009). Spring-spawning was only present in the northern region, including the Gulf of St. Lawrence and southwest

Newfoundland, while fall-spawning occurred across the whole range, from southern Labrador to the Gulf of Maine. The current restricted distribution of spring-spawning to the north suggests that this reproductive strategy, often associated with colder conditions (Melvin et al., 2009), has only found favorable environmental conditions in this region. Fishing reports, however, indicate that spring-spawning occurred in the past at sites where now is absent or less frequent. For example, spring-spawning was sparsely observed in southern locations such Spectacle Buoy, in mid-coastal Nova Scotia, and the Gulf of Maine (Bradford & Iles, 1992; Power et al., 2007), where it is no longer being reported from (DFO, 2017). In the last decades, an important reduction in recruitment and spatial distribution of the spring-spawning component has been noticed at both extremes of the reproductive range, with a coupled increase of the fall-spawning component (Bourne et al., 2018; McDermid et al., 2018; Melvin et al., 2009) and a sea warming trend (Melvin et al., 2009). Melvin et al. (2009) proposed a conceptual model explaining that the shift in the prevalence of spawning type may be an adaptive reproductive strategy in response to changing environments, where fall-spawning is favored by warming conditions while colder conditions are beneficial for spring-spawning. Although the evaluation of this hypothesis was beyond the scope of our study, the data generated here provides genetic evidence supporting the adaptive nature of these reproductive strategies. In particular, the observation that opposite alleles are close to fixation between spawning types is indicative of differential reproductive success of contrasting genotypes. Fall-spawning seems to be favored by the contemporary environmental conditions throughout the range, and the coupled effect of spawning time and specific adaptations. Some of such adaptations distinguish northern and southern fall-spawners along the latitudinal climate-related cline described in our previous study (Fuentes-Pardo et al., 2019). Considering that larvae remain aggregated near spawning grounds (Iles & Sinclair, 1982), larvae of fall-spawned fish typically spend 5+ months before metamorphosis to juvenile in March/April (Messieh, 1975; Tibbo, Legare, Scatterwood, & Temple, 1958), and average sea-water temperature of winter months appears to be the best predictor of the latitudinal genetic cline (Fuentes-Pardo et al., 2019), we infer that adaptations to spawning in northern or southern regions likely respond to natural selection acting on differential survival of larval stages over the winter months. We also hypothesize that adaptations to spawn in different seasons result from natural selection operating on differential reproductive success of spawning timing in adults and survivorship of early life stages of their offspring (from egg to hatching).

We confirmed allele frequency differences between spring and fall spawners are temporally stable over a span of 1, 2 and 9 years; same as reported before in two locations in the Gulf of St. Lawrence (Kerr et al., 2019). In addition, we observed spring and fall spawners in different gonadal maturity stages at the same spawning ground, particularly in southwest and northeast Newfoundland. We hypothesize these locations may not be in equilibrium likely due to the recent environmental changes occurring in the region (Melvin et al., 2009). The asynchrony in reproductive strategies at these locations perhaps documents the transition of the prevalence of one spawning type to the other. The observation that overall patterns of genetic differentiation persists between spring and fall spawners despite mixing outside of the breeding season, provides strong evidence of spawning site fidelity. This also indicates that selective pressures favoring particular genotype combinations are relatively stable at ecological time scales. We consider however, that the cumulated genetic differences between spring and fall spawners are antique and most likely result from standing genetic variation present in the populations that colonized the NW Atlantic after the last glacial maximum (Lamichhaney et al., 2017). Analysis of mitochondrial DNA may provide some insights on this regard. The short-term temporal stability of allele frequencies along the latitudinal cline was largely confirmed for spawning grounds from the northern region and for one offshore sample from mid-coastal Nova Scotia in the south (OMU14F-15F); no other southern site had temporal replicates. Such short-term temporal stability suggests that natural selection may be strong and stable in an ecological time scale. We found a few individuals with pure southern genotypes in the north and vice versa, which suggests they might be recent migrants and there is connectivity between fall-spawning aggregations along the latitudinal cline. The high dispersal potential across the range is also supported by tagging data that showed, for example, individuals tagged in northwest Gulf of St. Lawrence were recovered in consecutive years in mid-coastal Nova Scotia (Kim Emond pers. comm.).

5.5.3 Genetic evidence of hybridization

Given the persistent genetic differences between reproductive and latitudinal components, a natural question to address is whether they can interbreed and how common and viable hybrid individuals are. The hybrid index (Hindex) (Buerkle, 2005) provided a metric to investigate individual level of hybridization based on the genotypes obtained with the SPW- and LAT-panels. Our results suggest that the different components described in herring can freely interbreed, but they appear to be more successful in specific situations. For example, hybrid individuals (i.e. offspring between parents of the same species but belonging to different populations, defined by 0.1 < Hindex <0.9) were present across spawning aggregations but, more frequently, among spring spawners and in aggregations from the southern region. The observation of first generation and recombinant hybrids suggests multigenerational hybridization has occurred between spring and fall spawners over an extended period of time. Hybrid individuals represented a varying proportion of a spawning aggregation (about 13.3-60.0% of spring- and 0.0-26.7% of fall-spawning aggregations), suggesting perhaps that there is a reproductive advantage of being pure-spring or pure-fall spawner in some locations but not in others. It is not clear though, why most hybrids appear to be spring spawners. Perhaps spring-spawning hybrids have more chances to complete metamorphosis to juveniles during the favorable environmental window of the year (April and October) (Sinclair & Tremblay, 1984a), than fall-spawning hybrids. Further experiments are required to test this hypothesis.

The comparison of gonadal maturity and hybrid indices confirmed the high predictive power of the SPW-panel to determine the most likely spawning season of an individual, as previously observed in a smaller scale (Kerr et al., 2019). This analysis also revealed that different levels of hybridization may favor one spawning type over the other. In fact, we found that admixed individuals with a Hindex between ~54-69% can potentially spawn either in spring or fall. Further functional experiments are required to confirm this hypothesis, but in any case, this result points towards a genetic-based mechanism that could explain empirical observations based on otolith microstructure indicating that some spring-hatched individuals can spawn in the fall, and vice versa (Graham, 1962). We hypothesize some of the genetic variants analyzed here may provide flexibility of day-length and environmental cues assessment that trigger maturation in hybrid individuals. Hybridization between largely temporally isolated reproductive strategies most likely occurs between sympatric late spring spawners and early fall spawners. Indeed, such overlapping of spawning strategies has been well documented in Newfoundland (Winters & Wheeler, 1996; Winters, Wheeler, & Dalley, 1986; Winters, Wheeler, & Stansbury, 1993). Winters & Wheeler (1996) propose that January sea temperature may play a role in the onset of spring maturation, which sometimes vary as much as 4-5 weeks between years in this region.

Hybrids represented about 21.4-55.6% and 0.0-20.0 % of spawning aggregations in the southern and northern regions, respectively. Our results indicate that most hybrids between southern and northern fall-spawning components likely correspond to the first generation of hybridization. An alternative explanation to this pattern is that polymorphism may be advantageous in the south, especially for individuals from mid-coastal Nova Scotia and the Gulf of Maine. As discussed in our pervious study (Fuentes-Pardo et al., 2019), interannual variation in sea temperature or associated factors could potentially explain this pattern.

5.5.4 Analysis of mixture samples

Individual assignment of mixture inshore and offshore samples confirmed the dynamic nature of aggregations outside of the breeding season. For example, the composition of spring and fall spawners at Bras D'Or lake (BDO16S-17S) varied from one year to the following. Two possible explanations to this observation are that this location is being colonized by individuals from the Gulf of St. Lawrence and the Scotian Shelf, or that the relative abundance of spawning type has changed after the collapse in 90's of the spring component due to overfishing (Kerr et al., 2019). Offshore samples from mid-Nova Scotia (OMU14F-15F) comprised a few spring spawners from the north, where fall-spawning prevails, and fall spawners from the north and south. Similarly, the two offshore samples from southern Newfoundland had different compositions. All individuals in OEN15S were fall spawners from the north, whereas in OWN15S, some individuals were spring spawners from the north and others were fall spawners from the north and south. In summary, these results indicate the utility of the developed SNP panels for monitoring of stock composition outside of the breeding season.

5.5.5 Limitations

Despite our sampling effort, some locations were still underrepresented. For example, the southern region lacked temporal replicates and no spring-spawning samples were collected from northeast Newfoundland. In addition, the SNP panels described were designed to trace the adaptive genetic variation distinguishing the reproductive and latitudinal components revealed by Pool-seq data. Since such data did not reveal finer geographic differentiation, for example, among spawning aggregations within the north and south regions, individual assignment was restricted to a broad regional scale.

5.5.6 Implications in fisheries management and final remarks

This work demonstrates the utility of the reduced SNP panels developed here for the diagnosis of relevant biological components in the NW Atlantic herring. This means that it is now possible to estimate the relative composition of samples of mixed origin in close to real time for an affordable cost. Hence, we expect this genetic tool will facilitate mixed stock assessments even outside of the breeding season, which are important for the implementation of effective conservation and fisheries plans. Our results show that the joint analysis of gonadal maturation and SNP genotypes provides augmented accuracy of spawning season assignment than the currently implemented method, mostly based on the comparison of month of capture and inspection of gonads (LeBlanc et al., 2008; I. H. McQuinn, 1987). Moreover, our data indicates it is pertinent to review the current date of July 1st for the designation of spring and fall spawners, as it does not necessarily reflect the complex composition of some spawning grounds, such as in the northwest of the Gulf of St. Lawrence (SIL12S). Our results support the hypothesis that spawning at different seasons and geographic regions (i.e., biocomplexity, Ruzzante et al., 2006) may be an adaptive strategy of the species to increase the chances of successful reproduction and offspring survival under unpredictable changes in the environment (Melvin et al., 2009), namely a “portfolio effect” (Hilborn et al., 2003; Schindler et al., 2010). This work highlights the complexity of the herring mating system. It also warns that with the loss of adaptive genetic diversity some fish stocks could be restricted to a single reproductive strategy, implying increased vulnerability, reduced resilience and recovery potential to fishing pressure, if not stock depletion. Recent studies discuss the negative effects of fishing during the spawning period (van Overzee & Rijnsdorp, 2015). Therefore, it would be important to evaluate current fishing practices mostly targeting spawning grounds. Our results demonstrate the critical importance of characterizing and maintaining intraspecific genetic diversity for a sustainable use of harvested species to secure the ecological processes and economic activity that depend on them for the future.

5.6 Acknowledgements

We thank DFO staff, Lisa Pinkham from the Maine Department of Marine Resources, U.S., Comeau’s Seafoods Ltd, Cape Breeze Seafoods Ltd, fisherman Crystal and Donald Kent, Gordon McKay, and many others that provided herring samples. Special thanks to Quentin Kerr, James Kho, Connor Booker, and Gregory McCracken for collaboration in sampling and DNA extraction. SNP genotyping was done by Neogene Corporation (Lincoln, U.S.). Special thanks to Emma Sylvester, Sarah Lehnert, and Nick Jeffery for valuable data analysis advice. A.P.F.P. and D.E.R. thank the Killam Trust. A.P.F.P. thanks to the Vanier Canada Graduate Scholarship, the President’s Award of Dalhousie University, the Nova Scotia Graduate Scholarship, the NSERC Discovery of D.E.R., and the Lett Fund and NSERC for graduate studies funding. This study was funded by NSERC Discovery and Strategic grants to D.E.R.

5.7 Author contributions

D.E.R. and A.P.F.P. designed and conceived the study; A.P.F.P performed tissue collection, lab work and data analysis; C.B., R.S., H.B., and C.V. contributed tissue samples; A.P.F.P. drafted the manuscript and all authors contributed to writing and editing of the final version. All authors approved the manuscript before submission. Abbreviations of names as described in the statement of co-authorship (page 7).

5.8 Tables

Table 5.1 Characteristics of samples included in this study. Offshore samples are denoted with an asterisk as part of the population ID name. Temporal replicates are indicated by paired letters from "a" to "e". Fish with gonads in maturity stages 5 and 6 (ripening and actively spawning, respectively) were considered mature individuals. Abbreviations, IA: individual assignment, N: sample size.

Population ID	Location	Longitude	Latitude	N	Collection date (DD/MM/YY)	Proportion of mature individuals	Sample type for IA	Temporal replicate pairs
ME415F	Maine, area 514	-70.593	42.524	30	19-10-2015	0.20	Baseline	
ME314F	Maine, area 513	-69.640	43.607	27	25-08-2014	0.00	Baseline	
GEB14F	German Banks	-66.333	43.450	27	28-08-2014	1.00	Baseline	
MUS14F	Musquodoboit	-63.100	44.634	29	27-09-2014	1.00	Baseline	
OMU14F*	Offshore Musquodoboit	-62.300	44.417	49	04-10-2014	0.02	Mixture	a
OMU15F*	Offshore Musquodoboit	-62.133	44.483	48	28-10-2015	0.00	Mixture	a
SCB15F	Scots Bay	-64.917	45.167	30	24-08-2015	1.00	Baseline	
BDO16S	Bras D'Or lake	-60.849	45.929	28	20-04-2016	0.25	Mixture	b
BDO17S	Bras D'Or lake	-60.905	45.854	30	05-2017	NA	Mixture	b
NTS05S	Northumberland Strait	-64.527	46.397	30	28-04-2005, 06-05-2005	1.00	Baseline	c
NTS14S	Northumberland Strait	-64.122	46.303	30	01-05-2014	0.97	Baseline	c
NTS14F	Northumberland Strait	-63.088	45.807	31	16-09-2014	1.00	Baseline	
PEI05F	Prince Edward Island	-63.926	47.029	31	31-08-2005, 06-09-2005	0.94	Baseline	d
PEI14F	Prince Edward Island	-63.960	47.037	30	25-08-2014	1.00	Baseline	d
FTB14S	Fortune Bay	-55.633	47.283	29	03-05-2014, 26-05-2014	0.28	Baseline	e
FTB16S	Fortune Bay	-55.373	47.509	29	16-05-2016, 18-05-2016	0.03	Baseline	e
FTB15F	Fortune Bay	-54.934	47.575	30	12-10-2015	0.00	Baseline	
SMB15S	St. Mary's Bay	-53.646	46.917	30	14-06-2015, 11-04-2015	0.27	Baseline	
PLB16S	Placentia Bay	-54.077	47.734	29	08-06-2016, 07-05-2016	0.66	Baseline	
PLB15F	Placentia Bay	-54.007	47.407	30	14-12-2015	0.00	Baseline	
OEN15S*	Offshore Newfoundland	-56.907	46.997	50	17-04-2015	0.18	Mixture	
OWN15S*	Offshore Newfoundland	-58.102	47.210	50	22-04-2015	0.00	Mixture	
TRB14F	Trinity Bay	-53.473	47.842	30	28-09-2014	0.33	Baseline	
IBC14S	Inner Baie Des Chaleurs	-65.877	48.023	30	08-05-2014	0.60	Baseline	
OBC14F	Outer Baie Des Chaleurs	-64.435	47.894	30	20-08-2014	1.00	Baseline	
SPH12S	Stephenville	-57.940	49.732	30	30-05-2012	0.97	Baseline	
NDB15F	Notre Dame Bay	-55.469	49.550	30	26-10-2015	0.70	Baseline	
SIL12S	Seven Islands	-66.327	50.090	27	06-06-2012	0.81	Baseline	
BLS14F	Blanc Sablon	-57.314	51.379	30	13-08-2014	1.00	Baseline	
LAB15F	Labrador	-55.499	52.252	30	22-08-2015, 24-08-2014	0.40	Baseline	

5.9 Figures

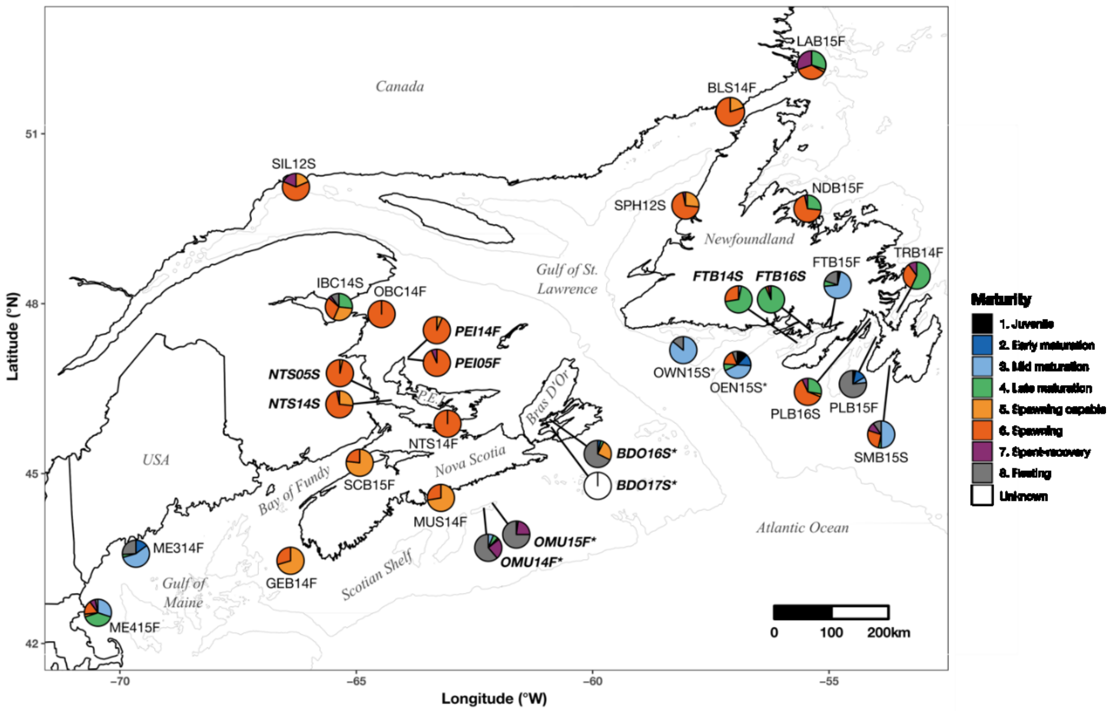


Figure 5.1 Map depicting sampling sites across the NW Atlantic. Pie charts represent the proportion of individuals in a given gonadal maturity stage at the time of collection. Collection date and site information are described in Table 5.1. Sample IDs in bold script letters indicate pairs of temporal replicates.

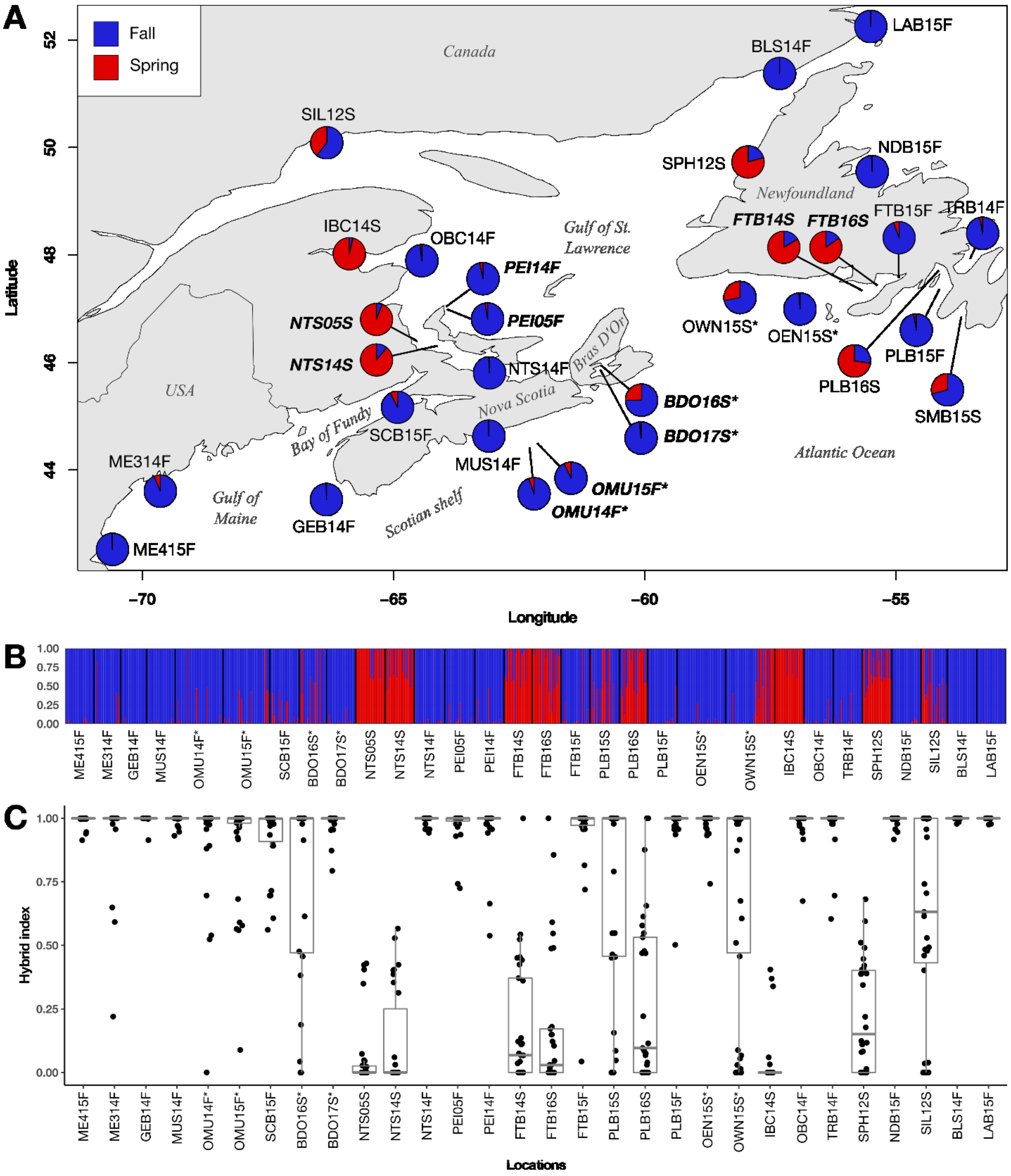


Figure 5.2 Population structure based on the SPW-panel. (A) Map showing the proportion of admixture coefficients per location, blue corresponds to the fall lineage and red to the spring lineage. **(B)** Bar plot depicting individual ancestry values (Q-value) obtained with ADMIXTURE (Alexander et al., 2009). Colors represent the same as in A. Each column represents the ancestry value of each individual. **(C)** Hybrid index (Hindex) per individual per sampling site. Hindex = 0 identify pure spring spawners, while Hindex = 1 identify pure fall spawners. The bars in gray show the median and the lower (25%) and upper (75%) quartiles.

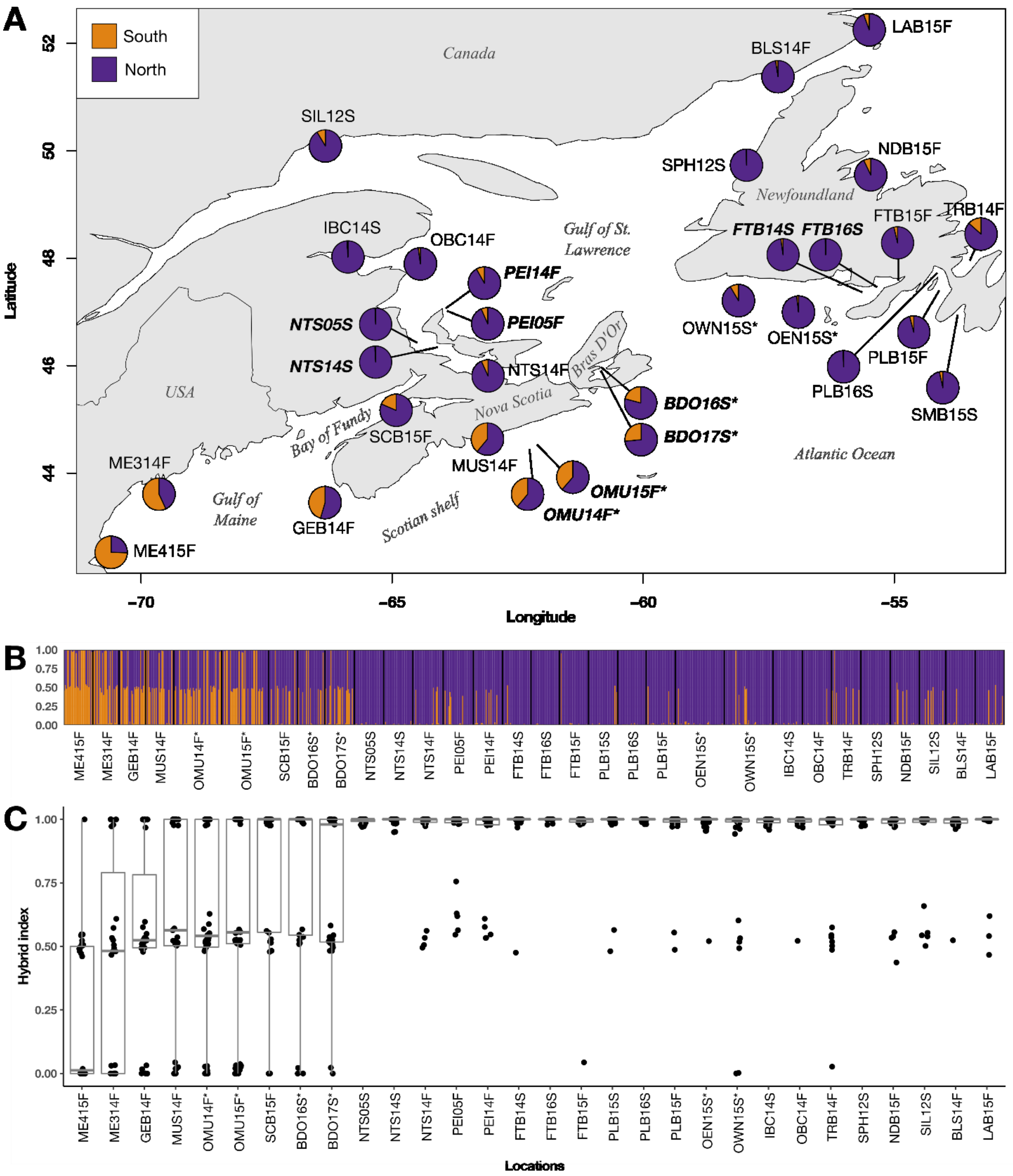


Figure 5.3 Population structure based on the LAT-panel. Map showing the proportion of admixture coefficients per location, purple corresponds to the northern lineage and orange to the southermost lineage. **(B)** Bar plot depicting individual ancestry values (Q-value) obtained with ADMIXTURE (Alexander et al., 2009). Colors represent the same as in A. Each column represents the ancestry value of each individual. **(C)** Hybrid index (Hindex) per individual per sampling site. Hindex = 0 identify pure south spawners, whereas Hindex = 1 identify pure north spawners. The bars in gray show the median and the lower (25%) and upper (75%) quartiles.

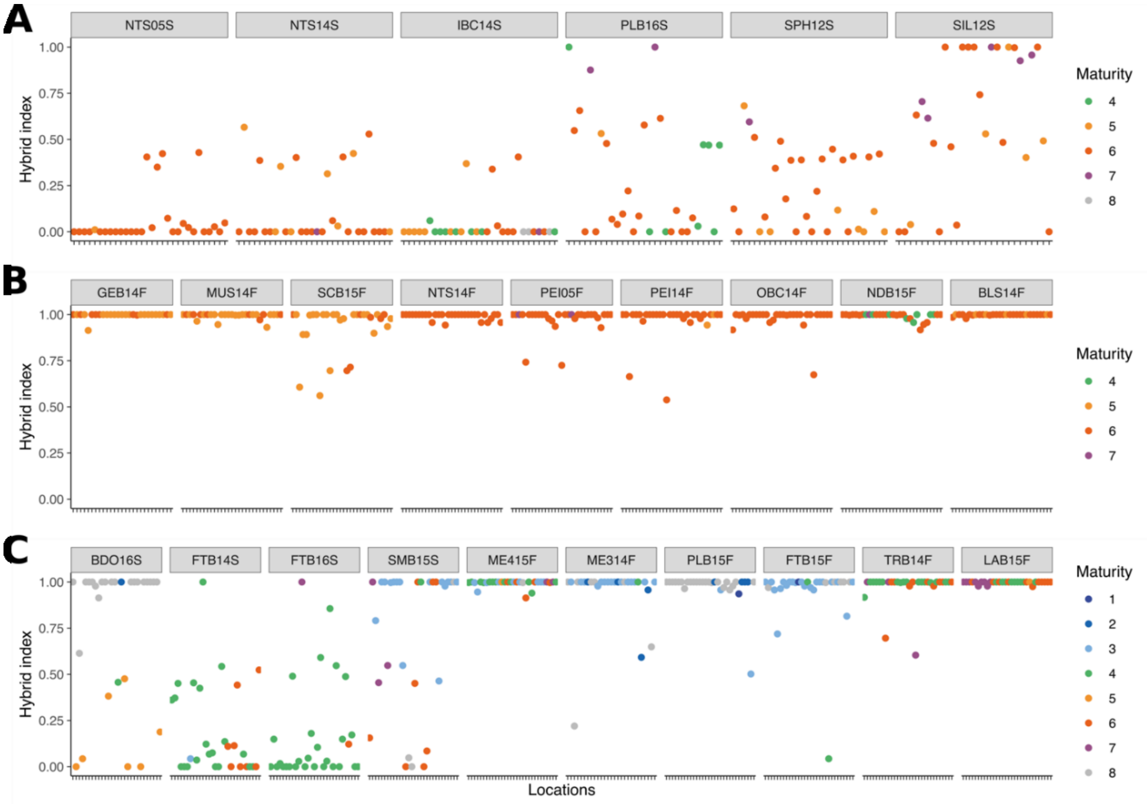


Figure 5.4 Comparison of gonadal maturity at the time of collection and Hindex values representing individual assignment to a spawning type based on the SPW-panel. (**A**) Pure spring spawning samples, (**B**) Pure fall spawning samples, (**C**) Samples with individuals in various maturity conditions. Colors of the dots represent gonadal maturity stage. 1 = Juvenile, 2 = Early maturation, 3 = Mid maturation, 4 = Late maturation, 5 = Spawning capable, 6 = Spawning, 7 = Spent-recovery, 8 = Resting. An index of 0% corresponds to a pure spring spawner, and of 100% to a pure fall spawner.

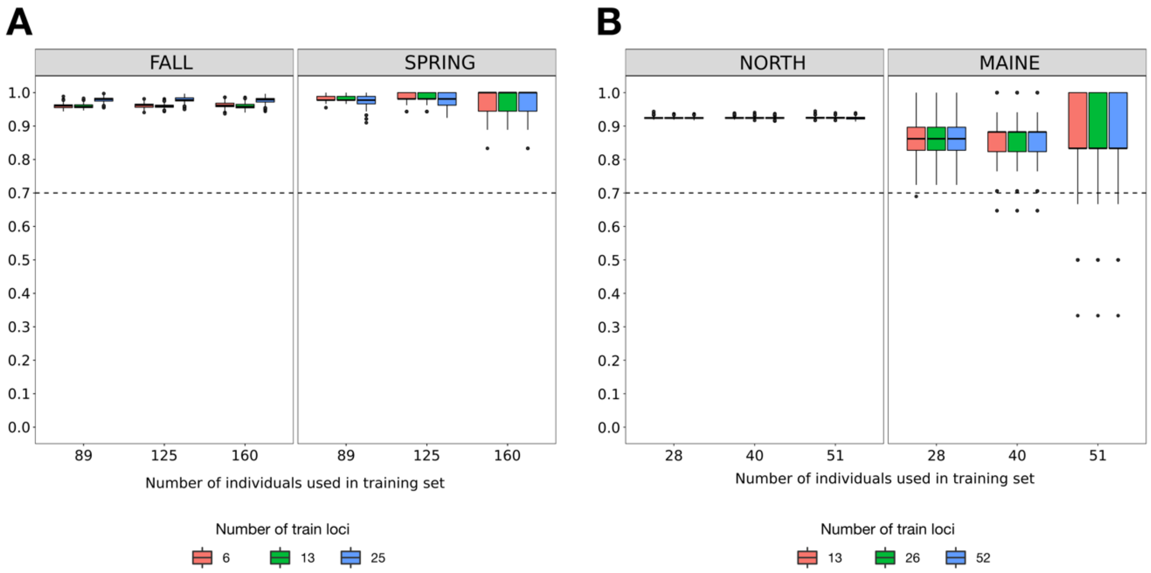


Figure 5.5 Self-assignment accuracy of spawning aggregations corresponding to the best performing baseline groups (option 2). Number of individuals used in training set were equivalent to 0.5, 0.7, 0.9 of the class with smaller sample size. Number of train loci used were equal to the top 0.25, 0.50 and 1.0 high-F_{ST} loci. **(A)** SPW-panel, **(B)** LAT-panel.

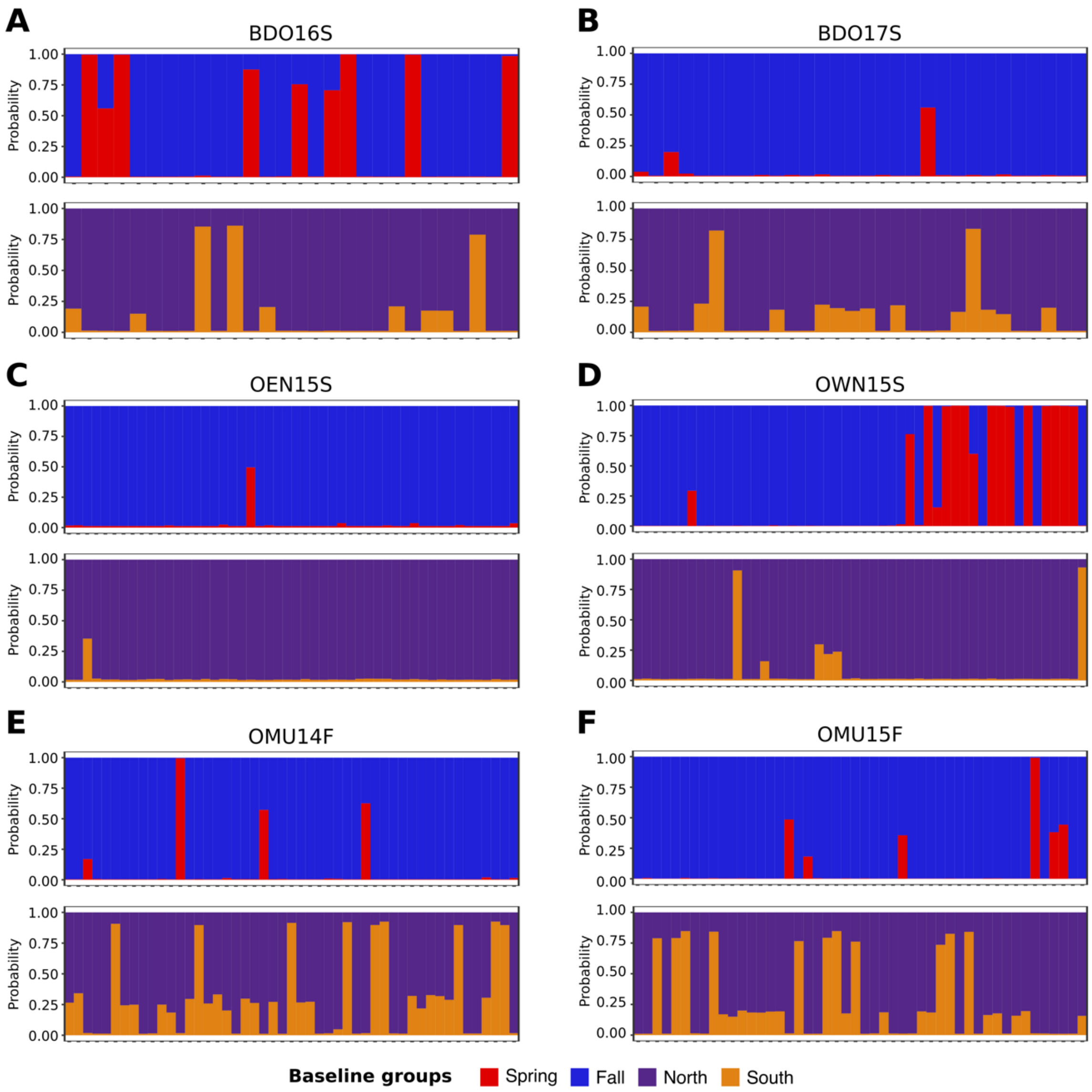


Figure 5.6 Barplots depicting assignment probability of individuals from mixed samples to baseline groups. (A) SPW-panel, **(B)** LAT-panel. Red represents spring-spawning, blue fall-spawning, purple spawning in the northern region, and orange, spawning in the southern region.

5.10 Supplementary Information

Supplementary Tables

Table S5.1 SNP loci that passed quality filters and constitute the SPW- and LAT-panels. **(electronic supplementary material).**

Table S5.2 Pairwise F_{ST} and P-values for the SPW-panel. **(electronic supplementary material).**

Table S5.3 Pairwise F_{ST} and P-values for the LAT-panel. **(electronic supplementary material).**

Table S5.4 Global AMOVA between spring- and fall-spawning individuals genotyped with the SPW-panel (10 100 permutations).

Source of variation	d.f.	Sum of squares	Variance components	Percentage of variation	Fixation indices	Significance (*P*-value)
Among groups	1	3650.61	5.86 Va	73.39	$F_{ST} = 0.78$	***
Among populations within groups	21	489.41	0.36 Vb	4.55	$F_{SC} = 0.17$	***
Within populations	1341	2364.01	1.76 Vc	22.06	$F_{CT} = 0.73$	***
Total	1363	6504.03	7.99			

Table S5.5 Global AMOVA between northern- and southern-spawning individuals genotyped with the LAT-panel (10 100 permutations).

Source of variation	d.f.	Sum of squares	Variance components	Percentage of variation	Fixation indices	Significance (*P*-value)
Among groups	1	1764.54	3.81 Va	47.16	$F_{ST} = 0.51$	***
Among populations within groups	22	536.90	0.35 Vb	4.28	$F_{SC} = 0.08$	***
Within populations	1394	5471.26	3.93 Vc	48.56	$F_{CT} = 0.47$	***
Total	1417	7772.70	8.08			

Supplementary Figures

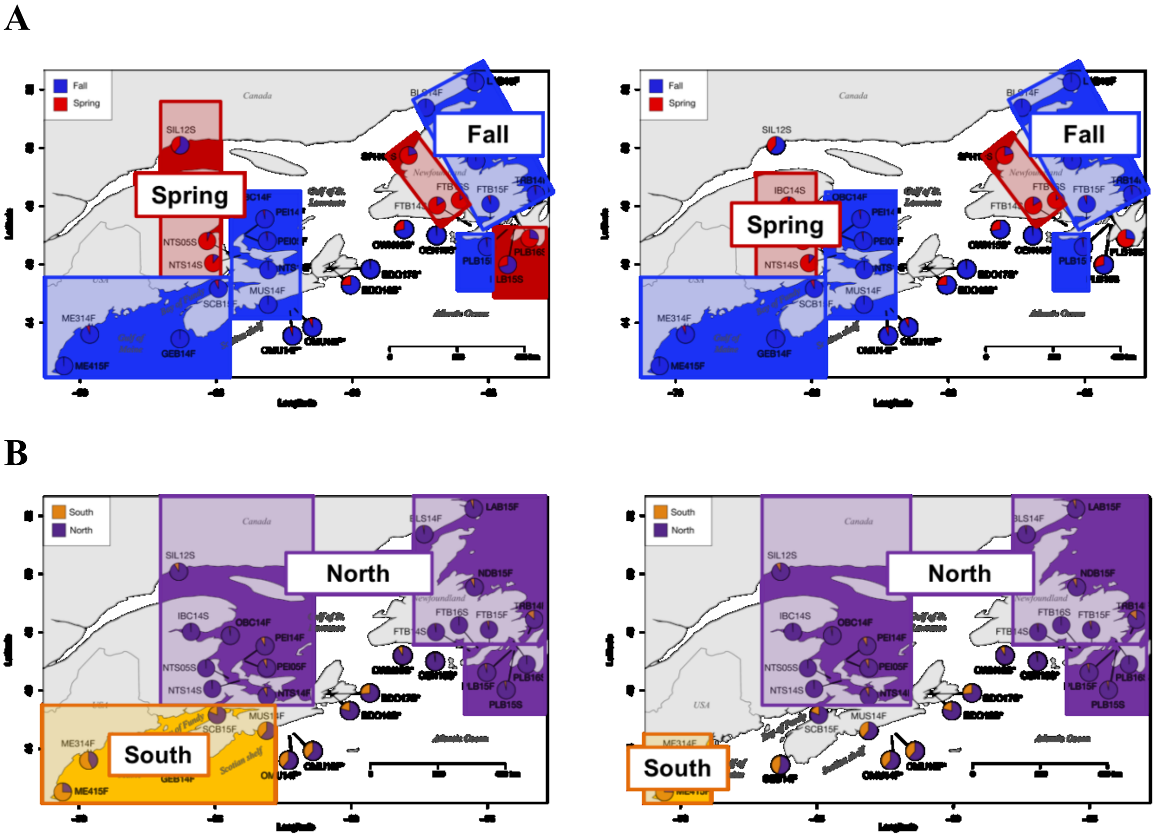


Figure S5.1 Proposed baselines used for self-assignment accuracy assessment of baseline groups. (**A**) SPW-panel, (left) baseline option 1, (right) baseline option 2; (**B**) LAT-panel, (left) baseline option 1, (right) baseline option 2. Explanation in the Materials and Methods section.

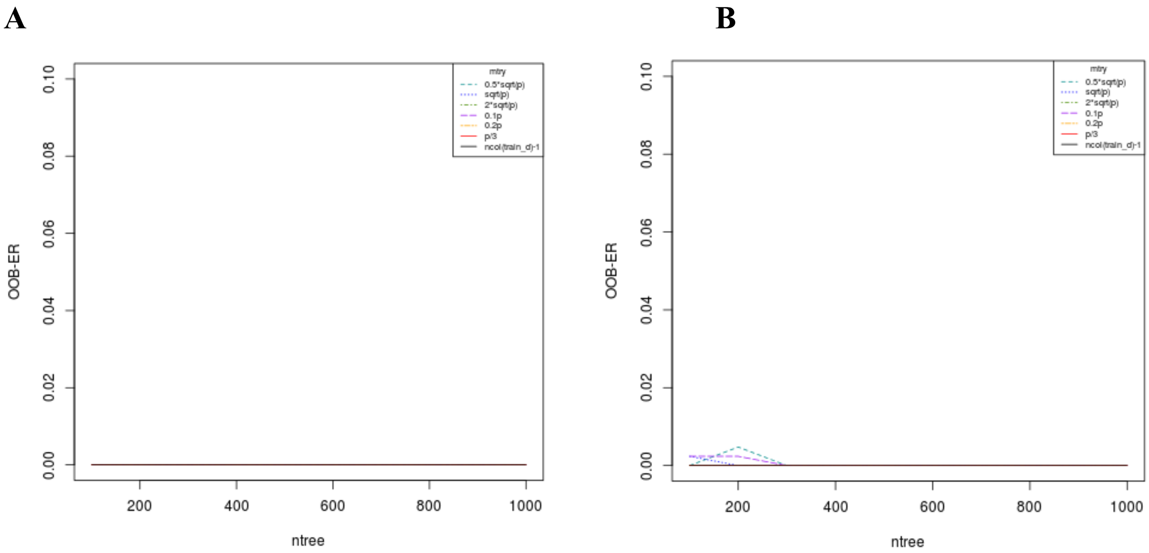


Figure S5.2 Testing of a combination of parameters for random forest runs. **(A)** For SPW-dataset, **(B)** For LAT-dataset.

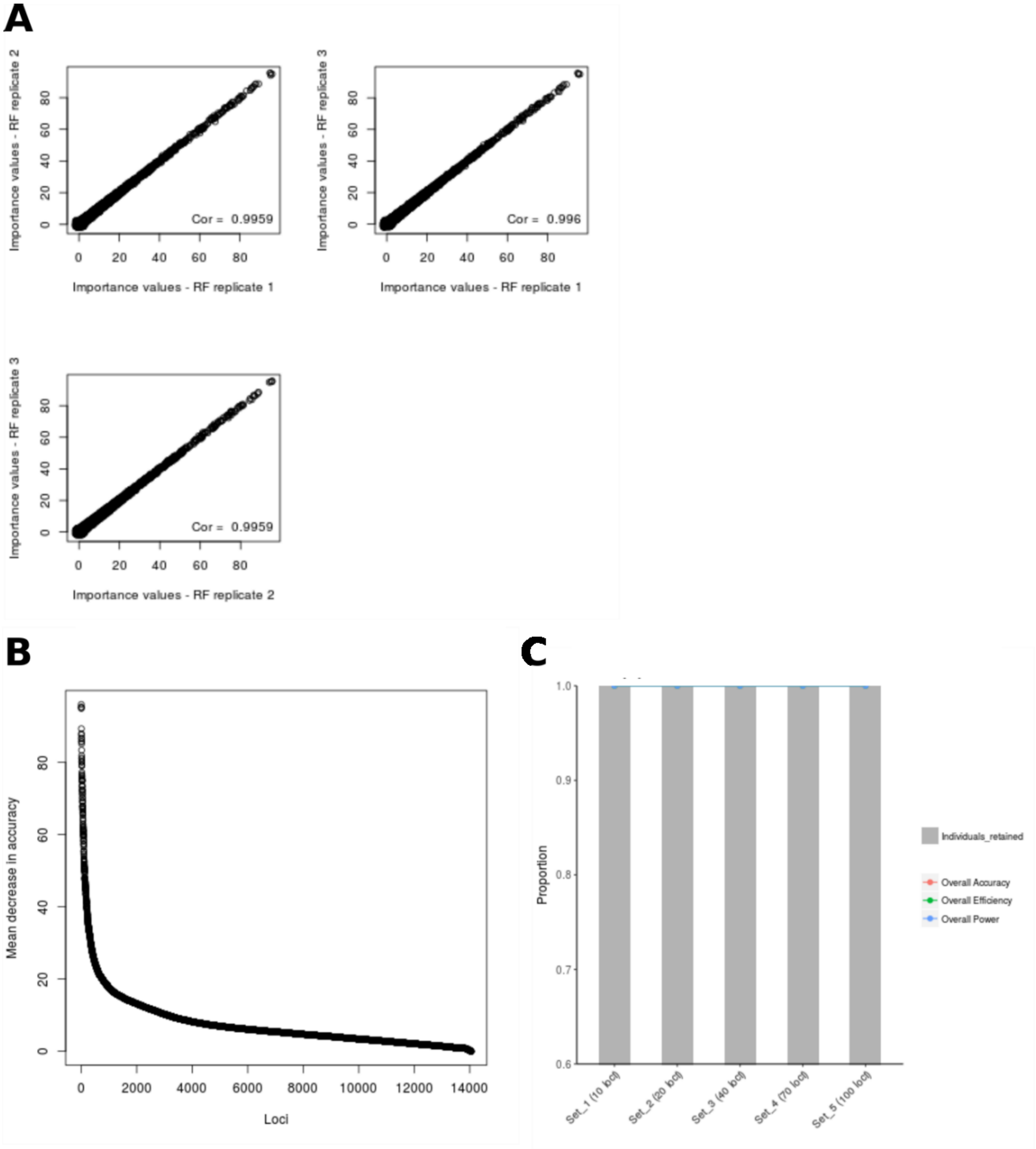


Figure S5.3 Random forest classification results and individual assignment for SPW-panel development. (**A**) Pairwise comparison of importance values (Mean Decrease in Accuracy – MDA) obtained in 3 random forest runs, (**B**) Scatterplot of importance values per loci, 14 052 SNPs, (**C**) Self-assignment metrics.

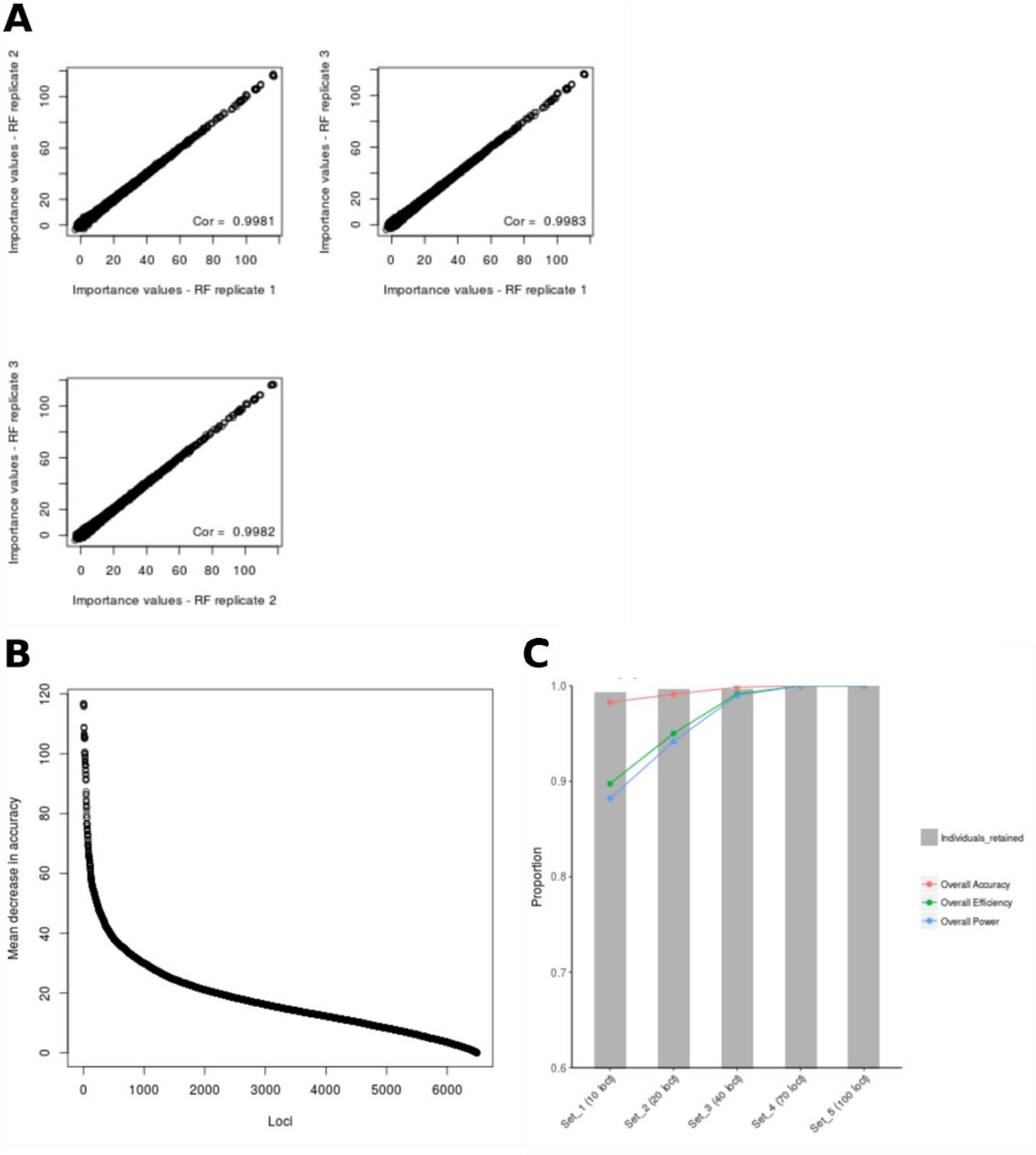


Figure S5.4 Random forest classification results and individual assignment for LAT-panel development. (**A**) Pairwise comparison of importance values (Mean Decrease in Accuracy – MDA) obtained in 3 random forest runs, (**B**) Scatterplot of importance values per loci, 6 493 SNPs, (**C**) Self-assignment metrics.

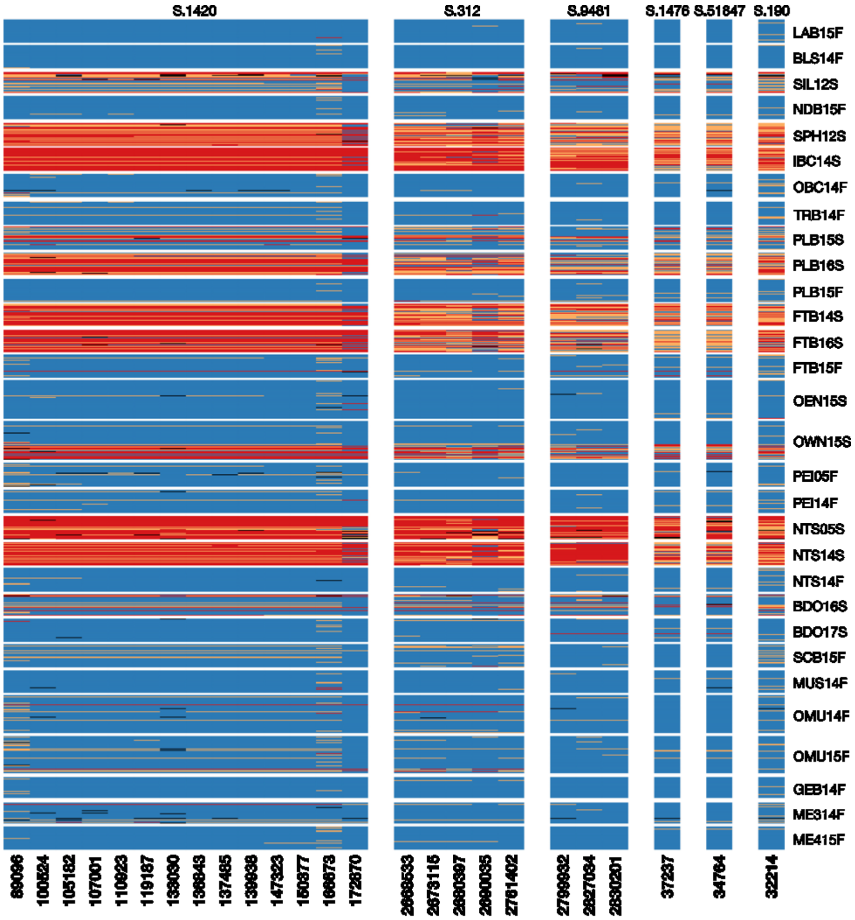


Figure S5.5 Heatmap representing the individual genotypes of the SNP loci in the SPW-panel. Each row corresponds to an individual which are clustered in a block representing a collection group. Each column is a SNP locus and a group of SNPs within the same scaffold are shown as a vertical group. Colors represent each genotype, blue for homozygote for the major allele, yellow is for heterozygotes, and red is for homozygotes for the alternate allele.

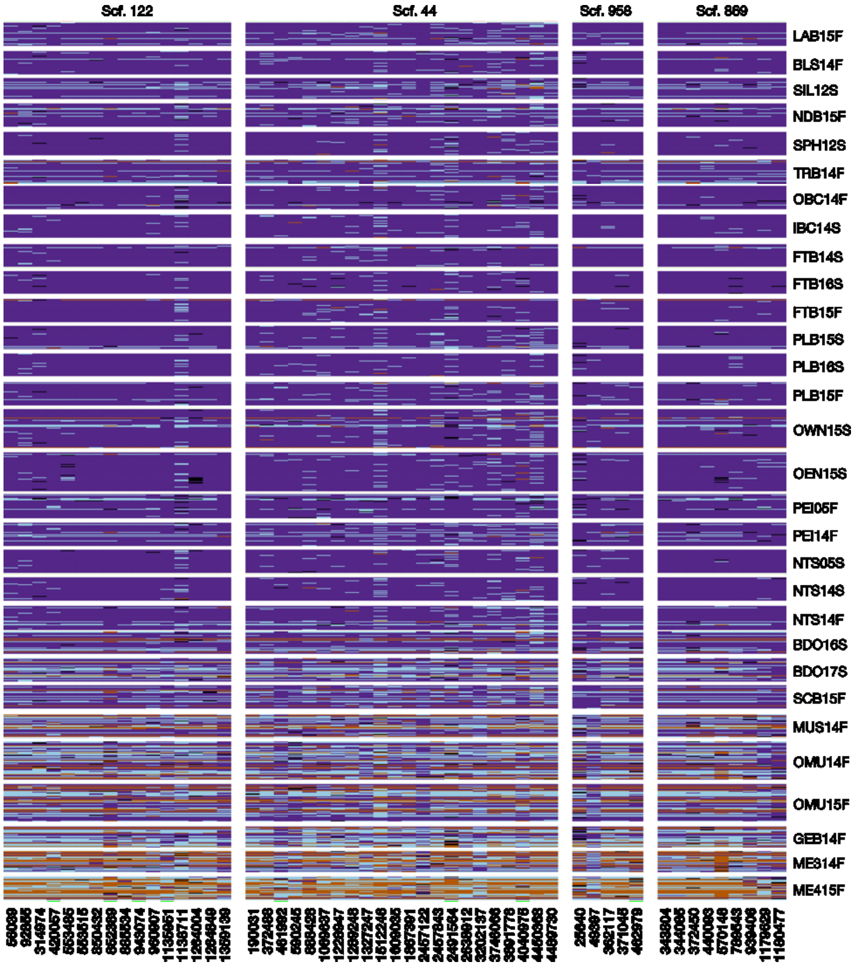


Figure S5.6 Heatmap representing the individual genotypes of the SNP loci in the LAT-panel. Each row corresponds to an individual which are clustered in a block representing a collection group. Each column is a SNP locus and a group of SNPs within the same scaffold are shown as a vertical group. Colors represent each genotype, purple for homozygote for the major allele, light blue is for heterozygotes, and orange is for homozygotes for the alternate allele.

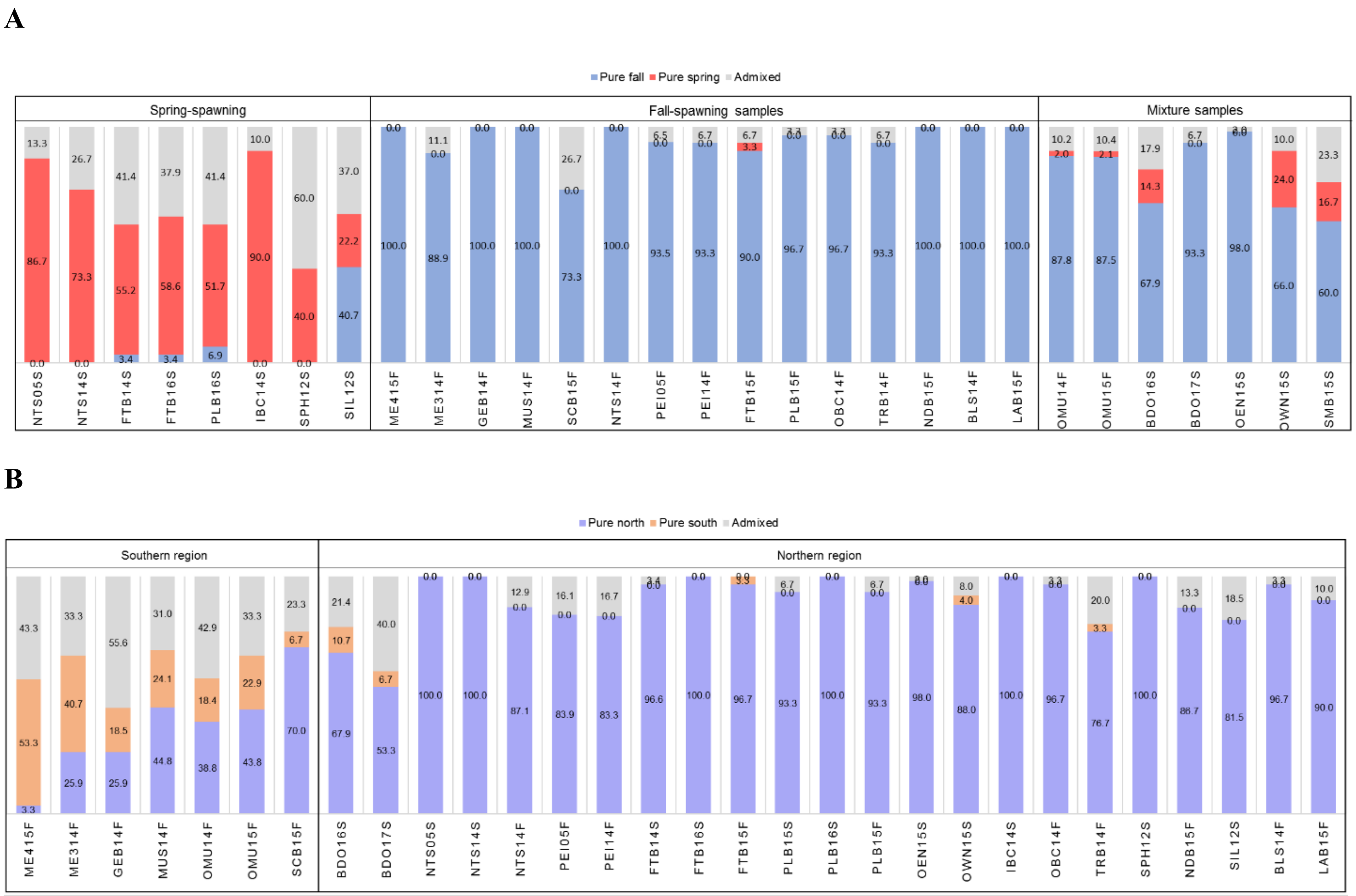


Figure S5.7 Barplot indicating the relative proportion of components represented in 30 NW Atlantic herring aggregations. (**A**) Respect to the SPW-panel, pure spring-spawning individuals are denoted in red, pure fall-spawning in blue, and admixed individuals are shown in gray. (**B**) Respect to the LAT-panel, pure northern individuals are denoted in purple, pure southern in orange, and admixed individuals are shown in gray.

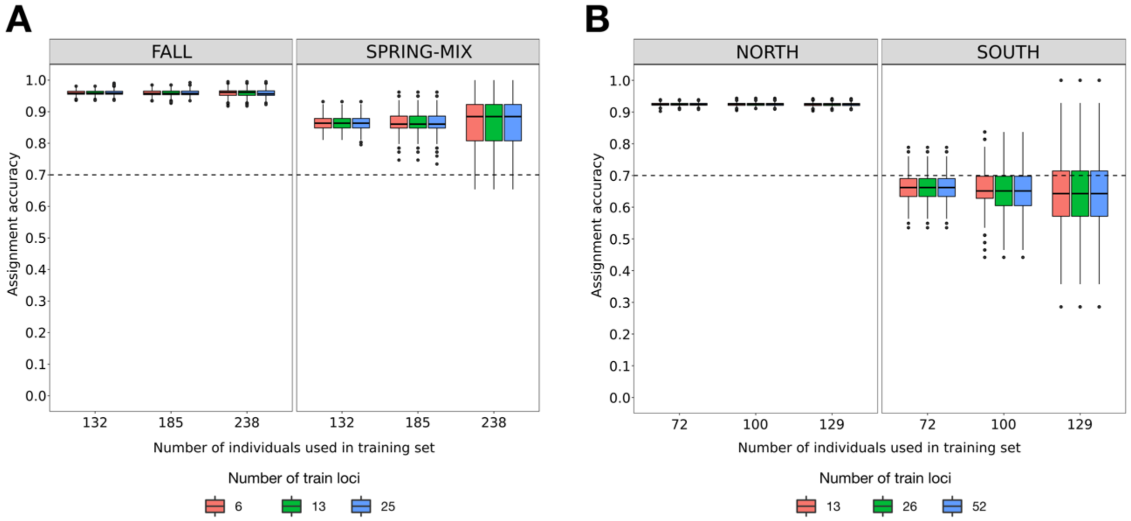


Figure S5.8 Assignment accuracy of the best performing baseline groups (option 1). Number of individuals used in training set corresponded to the 0.5, 0.7, 0.9 of the class with smaller sample size. Number of train loci corresponded to the top high-F_{ST} loci. **(A)** SPW-panel, **(B)** LAT-panel.

5.11 References

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Allendorf, F. W. (2016). Genetics and the conservation of natural populations: Allozymes to genomes. *Molecular Ecology*, *38*(1), 42–49. https://doi.org/10.1111/mec.13948

Anderson, E. C. (2010). Assessing the power of informative subsets of loci for population assignment: standard methods are upwardly biased. *Molecular Ecology Resources*, *10*(4), 701–710. https://doi.org/10.1111/j.1755-0998.2010.02846.x

Anderson, E. C., Waples, R. S., & Kalinowski, S. T. (2008). An improved method for predicting the accuracy of genetic stock identification. *Canadian Journal of Fisheries and Aquatic Sciences*, *65*(7), 1475–1486. https://doi.org/10.1139/F08-049

Bekkevold, D., Helyar, S. J., Limborg, M. T., Nielsen, E. E., Hemmer-Hansen, J., Clausen, L. A. W., & Carvalho, G. R. (2015). Gene-associated markers can assign origin in a weakly structured fish, Atlantic herring. *ICES Journal of Marine Science: Journal Du Conseil*, *72*(6), 1790–1801. https://doi.org/10.1093/icesjms/fsu247

Benoît, H., Swain, D., Hutchings, J., Knox, D., Doniol-Valcroze, T., & Bourne, C. (2018). Evidence for reproductive senescence in a broadly distributed harvested marine fish. *Marine Ecology Progress Series*, *592*, 207–224. https://doi.org/10.3354/meps12532

Bernatchez, L., Wellenreuther, M., Araneda, C., Ashton, D. T., Barth, J. M. I., Beacham, T. D., … Withler, R. E. (2017). Harnessing the Power of Genomics to Secure the Future of Seafood. *Trends in Ecology & Evolution*, *32*(9), 665–680. https://doi.org/10.1016/j.tree.2017.06.010

Bourne, C., Mowbray, F., Squires, B., & Koen-Alonso, M. (2018). 2017 Assessment of Newfoundland east and south coast Atlantic herring (Clupea harengus) stock complexes. *DFO Can. Sci. Advis. Sec. Res. Doc. 2018/026*, v + 45 p.

Bradbury, I. R., Hamilton, L. C., Sheehan, T. F., Chaput, G., Robertson, M. J., Dempson, J. B., … Bernatchez, L. (2016). Genetic mixed-stock analysis disentangles spatial and temporal variation in composition of the West Greenland Atlantic Salmon fishery. *ICES Journal of Marine Science*, *73*(9), 2311–2321. https://doi.org/10.1093/icesjms/fsw072

Bradford, R. G., & Iles, T. D. (1992). Unique biological characteristics of spring-spawning herring (Clupea harengus L.) in Minas Basin, Nova Scotia, a tidally dynamic environment. *Canadian Journal of Zoology*, *70*, 641–648.

Breiman, L. (2001). Random Forest. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Brieuc, M. S. O., Waters, C. D., Drinan, D. P., & Naish, K. A. (2018). A practical introduction to Random Forest for genetic association studies in ecology and evolution. *Molecular Ecology Resources*, *18*(4), 755–766. https://doi.org/10.1111/1755-0998.12773

Buerkle, A. C. (2005). Maximum-likelihood estimation of a hybrid index based on molecular markers. *Molecular Ecology Notes*, *5*(3), 684–687. https://doi.org/10.1111/j.1471-8286.2005.01011.x

Chen, K.-Y., Marschall, E. A., Sovic, M. G., Fries, A. C., Gibbs, H. L., & Ludsin, S. A. (2018). assignPOP: An r package for population assignment using genetic, non-genetic, or integrated data in a machine-learning framework. *Methods in Ecology and Evolution*, *9*(2), 439–446. https://doi.org/10.1111/2041-210X.12897

Denny, S., Clark, K. J., Power, M. J., & Stephenson, R. L. (1998). The status of the herring in the Bras d'Or Lakes in 1996–1997. *Canadian Stock Assessment Secretariat Research Document*, *80*, 1–32.

DFO. (2017). Stock Status Update of 4VWX Herring. In *DFO Can. Sci. Advis. Sec. Sci. Resp. 2017/037*.

Excoffier, L., & Lischer, H. E. L. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, *10*(3), 564–567. https://doi.org/10.1111/j.1755-0998.2010.02847.x

FAO. (2019). Species Fact Sheets: Clupea harengus (Linnaeus, 1758). Retrieved from http://www.fao.org/fishery/species/2886/en

Frank, K. T., & Brickman, D. (2000). Allee effects and compensatory population dynamics within a stock complex. *Canadian Journal of Fisheries and Aquatic Sciences*, *57*, 513–517.

Freamo, H., O'reilly, P., Berg, P. R., Lien, S., & Boulding, E. G. (2011). Outlier SNPs show more genetic structure between two Bay of Fundy metapopulations of Atlantic salmon than do neutral SNPs. *Molecular Ecology Resources*, *11*, 254–267. https://doi.org/10.1111/j.1755-0998.2010.02952.x

Fuentes-Pardo, A. P., Bourne, C., Singh, R., Emond, K., Pinkham, L., McDermid, J. L., … Ruzzante, D. E. (2019). Adaptation to seasonal reproduction and thermal minima-related factors drives fine-scale divergence despite gene flow in Atlantic herring populations. *BioRxiv*. https://doi.org/https://doi.org/10.1101/578484

Goldstein, B. A., Hubbard, A. E., Cutler, A., & Barcellos, L. F. (2010). An application of Random Forests to a genome-wide association dataset: Methodological considerations & new findings. *BMC Genetics*, *11*(1), 49. https://doi.org/10.1186/1471-2156-11-49

Graham, T. R. (1962). A relationship between growth, hatching and spawning season in Canadian Atlantic herring (Clupea harengus L .). *J. Fish. Res. Bd. Canada*, *19*(5), 985–987.

Hilborn, R., Quinn, T. P., Schindler, D. E., & Rogers, D. E. (2003). Biocomplexity and fisheries sustainability. *Proceedings of the National Academy of Sciences*, *100*(11), 6564–6568. https://doi.org/10.1073/pnas.1037274100

Iles, T. D., & Sinclair, M. (1982). Atlantic Herring: Stock Discreteness and Abundance. *Science*, *215*(4533), 627–633. https://doi.org/10.1126/science.215.4533.627

Jeffery, N. W., Wringe, B. F., McBride, M. C., Hamilton, L. C., Stanley, R. R. E., Bernatchez, L., … Bradbury, I. R. (2018). Range-wide regional assignment of Atlantic salmon (Salmo salar) using genome wide single-nucleotide polymorphisms. *Fisheries Research*, *206*, 163–175. https://doi.org/10.1016/j.fishres.2018.05.017

Kalinowski, S. T., Manlove, K. R., & Taper, M. L. (2007). *ONCOR: software for genetic stock identification*.

Kerr, Q., Fuentes-Pardo, A. P., Kho, J., McDermid, J. L., & Ruzzante, D. E. (2019). Temporal stability and assignment power of adaptively divergent genomic regions between herring (Clupea harengus) seasonal spawning aggregations. *Ecology and Evolution*, (9), 500–510. https://doi.org/10.1002/ece3.4768

Laikre, L., Lundmark, C., Jansson, E., Wennerström, L., Edman, M., & Sandström, A. (2016). Lack of recognition of genetic biodiversity: International policy and its implementation in Baltic Sea marine protected areas. *Ambio*, *45*(6), 661–680. https://doi.org/10.1007/s13280-016-0776-7

Lamichhaney, S., Barrio, A. M., Rafati, N., Sundstrom, G., Rubin, C.-J., Gilbert, E. R., … Andersson, L. (2012). Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. *Proceedings of the National Academy of Sciences*, *109*(47), 19345–19350. https://doi.org/10.1073/pnas.1216128109

Lamichhaney, S., Fuentes-Pardo, A. P., Rafati, N., Ryman, N., McCracken, G. R., Bourne, C., … Andersson, L. (2017). Parallel adaptive evolution of geographically distant herring populations on both sides of the North Atlantic Ocean. *Proceedings of the National Academy of Sciences*, *114*(17), E3452–E3461. https://doi.org/10.1073/pnas.1617728114

LeBlanc, C. H., Poirier, A. G., MacDougall, C., Bourque, C., & Roy, J. (2008). Assessment of the NAFO Division 4T southern Gulf of St. Lawrence herring stocks in 2007. In *Canadian Science Advisory Secretariat Research Document*.

Liaw, A, & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22.

Liaw, Andy, & Wiener, M. (2018). Breiman and Cutler's Random Forests for Classification and Regression.

Lischer, H. E. L., & Excoffier, L. (2012). PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, *28*(2), 298–299. https://doi.org/10.1093/bioinformatics/btr642

Luu, K., Bazin, E., & Blum, M. G. B. (2017). pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources*, *17*(1), 67–77. https://doi.org/10.1111/1755-0998.12592

Martinez Barrio, A., Lamichhaney, S., Fan, G., Rafati, N., Pettersson, M., Zhang, H., … Andersson, L. (2016). The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. *ELife*, *5*(MAY2016), 1–32. https://doi.org/10.7554/eLife.12081

McDermid, J. L., Swain, D. P., Turcotte, F., Robichaud, S. A., & Surette, T. (2018).

Assessment of the NAFO Division 4T southern Gulf of St. Lawrence Atlantic herring (Clupea harengus) in 2016 and 2017. *DFO Can. Sci. Advis. Sec. Res. Doc. 2018/052*, xiv + 122 p.

McQuinn, I. H. (1987). New maturity cycle charts for herring stocks along the west coast of Newfoundland (NAFO Division 4R) and the north shore of Quebec (NAFO Division 4S). In *Canadian Atlantic Fisheries Scientific Advisory Committee Research Document*.

Meirmans, P. G., & Van Tienderen, P. H. (2004). genotype and genodive: two programs for the analysis of genetic diversity of asexual organisms. *Molecular Ecology Notes*, *4*(4), 792–794. https://doi.org/10.1111/j.1471-8286.2004.00770.x

Melvin, G. D., Stephenson, R. L., & Power, M. J. (2009). Oscillating reproductive strategies of herring in the western Atlantic in response to changing environmental conditions. *ICES Journal of Marine Science*, *66*(8), 1784–1792. https://doi.org/10.1093/icesjms/fsp173

Messieh, S. N. (1975). Maturation and spawning of Atlantic herring (Clupea harengus harengus) in the southern Gulf of St Lawrence. *Journal of the Fisheries Research Board of Canada*, *32*, 66–68.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C. -C., & Lin, C. -C. (2015). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071).

Moran, B. M., & Anderson, E. C. (2018). Bayesian inference from the conditional genetic stock identification model. *Canadian Journal of Fisheries and Aquatic Sciences*, 1–10. https://doi.org/10.1139/cjfas-2018-0016

Nayfa, M. G., & Zenger, K. R. (2016). Unravelling the effects of gene flow and selection in highly connected populations of the silver-lip pearl oyster (Pinctada maxima). *Marine Genomics*, *28*, 99–106. https://doi.org/10.1016/j.margen.2016.02.005

Nosil, P., & Feder, J. L. (2012). Genomic divergence during speciation: causes and consequences. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1587), 332–342. https://doi.org/10.1098/rstb.2011.0263

Pompanon, F., Bonin, A., Bellemain, E., & Taberlet, P. (2005). Genotyping errors: causes, consequences and solutions. *Nature Reviews Genetics*, *6*(11), 847–859. https://doi.org/10.1038/nrg1707

Power, M. J., Clark, K. J., Fife, J. F., Knox, D., Melvin, G. D., & Stephenson, R. L. (2007). 2007 evaluation of 4VWX herring. In *Canadian Science Advisory Secretariat Research Document, 2007/040*.

Pritchard, J., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

R Core Development Team. (2019). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Reiss, H., Hoarau, G., Dickey-Collas, M., & Wolff, W. J. (2009). Genetic population structure of marine fish: mismatch between biological and fisheries management units. *Fish and Fisheries*, *10*(4), 361–395. https://doi.org/10.1111/j.1467-2979.2008.00324.x

Ruzzante, D. E., Mariani, S., Bekkevold, D., André, C., Mosegaard, H., Clausen, L. A. W., … Carvalho, G. R. (2006). Biocomplexity in a highly migratory pelagic marine fish, Atlantic herring. *Proceedings of the Royal Society B: Biological Sciences*, *273*(1593), 1459–1464. https://doi.org/10.1098/rspb.2005.3463

Ruzzante, D. E., Taggart, C. T., Lang, S., Cook, D., Applications, E., & Aug, N. (2000). Mixed-stock analysis of Atlantic cod near the Gulf of St. Lawrence based on microsatellite DNA. *Ecological Applications*, *10*(4), 1090–1109.

Satterthwaite, W. H., & Carlson, S. M. (2015). Weakening portfolio effect strength in a hatchery-supplemented Chinook salmon population complex. *Canadian Journal of Fisheries and Aquatic Sciences*, *72*(12), 1860–1875. https://doi.org/10.1139/cjfas-2015-0169

Schindler, D. E., Hilborn, R., Chasco, B., Boatright, C. P., Quinn, T. P., Rogers, L. A., & Webster, M. S. (2010). Population diversity and the portfolio effect in an exploited species. *Nature*, *465*(7298), 609–612. https://doi.org/10.1038/nature09060

Schrider, D. R., & Kern, A. D. (2018). Supervised Machine Learning for Population Genetics: A New Paradigm. *Trends in Genetics*, *34*(4), 301–312. https://doi.org/10.1016/j.tig.2017.12.005

Scott, W. B., & Scott, M. G. (1988). *Atlantic fishes of Canada. Canadian Bulletin of Fisheries and Aquatic Sciences, bulletin 219*. Toronto, CA: University of Toronto Press.

Shafer, A. B. A., Wolf, J. B. W., Alves, P. C., Bergström, L., Bruford, M. W., Brännström, I., … Zieliński, P. (2015). Genomics and the challenging translation into conservation practice. *Trends in Ecology and Evolution*, *30*(2), 78–87. https://doi.org/10.1016/j.tree.2014.11.009

Sinclair, M., & Tremblay, M. J. (1984). Timing of Spawning of Atlantic Herring (Clupea harengus harengus) Populations and the Match–Mismatch Theory. *Canadian Journal of Fisheries and Aquatic Sciences*, *41*(7), 1055–1065. https://doi.org/10.1139/f84-123

Stanley, R. R. E., Jeffery, N. W., Wringe, B. F., DiBacco, C., & Bradbury, I. R. (2017). <scp>genepopedit</scp> : a simple and flexible tool for manipulating multilocus molecular data in R. *Molecular Ecology Resources*, *17*(1), 12–18. https://doi.org/10.1111/1755-0998.12569

Stephenson, R. L., Melvin, G. D., & Power, M. J. (2009). Population integrity and connectivity in Northwest Atlantic herring: a review of assumptions and evidence. *ICES Journal of Marine Science*, *66*(8), 1733–1739. https://doi.org/10.1093/icesjms/fsp189

Stobo, W. T. (1987). Atlantic herring (Clupea harengus) movement along the Scotian Shelf and management considerations. *Proceedings of the Conference on Forage Fishes of the Southeastern Bering Sea, Anchorage, Alaska, 4–5 November 1986, Pp. 75–85. US Department of the Interior, Minerals Management Services, Alaska OCS Region, MMS Report, 87-0017. 122 Pp.*

Sylvester, E. V. A., Bentzen, P., Bradbury, I. R., Clément, M., Pearce, J., Horne, J., & Beiko, R. G. (2017). Applications of random forest feature selection for fine-scale genetic population assignment. *Evolutionary Applications*, (September 2016), 1–13. https://doi.org/10.1111/eva.12524

Tibbo, S. N., Legare, J. E. H., Scatterwood, L. W., & Temple, R. F. (1958). On the occurrence and distribution of larval herring (Clupea harengus L.) in the Bay of Fundy and the Gulf of Maine. *Journal of the Fisheries Research Board of Canada*, *15*, 1451–1469.

Vähä, J.-P., & Primmer, C. R. (2006). Efficiency of model-based Bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci. *Molecular Ecology*, *15*, 63–72. https://doi.org/10.1111/j.1365-294X.2005.02773.x

van Overzee, H. M. J., & Rijnsdorp, A. D. (2015). Effects of fishing during the spawning period: implications for sustainable management. *Reviews in Fish Biology and Fisheries*, *25*(1), 65–83. https://doi.org/10.1007/s11160-014-9370-x

Waters, C. L., & Clark, K. J. (2005). 2005 summary of the weir herring tagging project with an update of the HSC/PRC/DFO herring tagging program. In *Canadian Science Advisory Secretariat Research Document, 2005/025*.

Wheeler, J. P., & Winters, G. H. (1984). Migrations and stock relationships of east and southeast Newfoundland herring (Clupea harengus) as shown by tagging studies. *Journal of Northwest Atlantic Fishery Science*, *5*, 121–129.

Winters, G. H., & Wheeler, J. P. (1987). Recruitment dynamics of spring-spawning herring in the Northwest Atlantic. *Canadian Journal of Fisheries and Aquatic Sciences*, *44*, 882–900.

Winters, G. H., & Wheeler, J. P. (1996). Environmental and phenotypic factors affecting the reproductive cycle of Atlantic herring. *ICES Journal of Marine Science*, *53*, 73–88. https://doi.org/10.1006/jmsc.1996.0007

Winters, G. H., Wheeler, J. P., & Dalley, E. L. (1986). Survival of a herring stock subjected to a catastrophic event and fluctuating environmental conditions. *Journal Du Conseil International Pour l'Exploration de La Mer*, *43*, v.

Winters, G. H., Wheeler, J. P., & Stansbury, D. (1993). Variability in the reproductive output of spring-spawning herring in the north-west atlantic. *ICES Journal of Marine Science*, Vol. 50, pp. 15–25. https://doi.org/10.1006/jmsc.1993.1003

Xu, Q. -S., & Liang, Y. -Z. (2001). Monte–Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems*, *56*, 1–11.

CHAPTER 6. CONCLUSION

6.1 Summary

The preservation of diverse reproductive strategies and of locally adapted populations is crucial for species' persistence and resilience to environmental change and to human pressures (Hilborn et al., 2003; Ruzzante et al., 2006; Schindler et al., 2010). Yet, in marine organisms, typically assumed to be genetically homogeneous or minimally subdivided, it is not well understood at what spatial scale intraspecific diversity is distributed across the species' range; which evolutionary processes and environmental variables play a role in the origin and maintenance of population divergence and local adaptation in the presence of gene flow; or which parts of the genome underpin adaptive phenotypes nor what potential function they have. These questions have been limitedly addressed, in part, because of the still high cost of high-throughput sequencing and the minimal genomic resources existing for most non-model organisms. The lack of knowledge on intraspecific diversity and on the appropriate genetic tools that facilitate its assessment, can be detrimental from a conservation and fisheries perspective, as this lack of knowledge increases their vulnerability to depletion.

In this thesis, I used Atlantic herring (*C. harengus*) as a model species to address these questions, given its large population size implying minute effect of genetic drift, high dispersal capability, and extensive geographic distribution, characteristics that make this species ideal for the study of ecological adaptation. The overarching goals of my thesis were to try to (*i*) identify the patterns, genomic regions, evolutionary processes and environmental variables involved in the origin and maintenance of population divergence and local adaptation in the northwest Atlantic, and to (*ii*) develop a diagnostic genetic tool that can inform conservation and fisheries management. An expanded comprehension of the patterns, genetic basis, and mechanisms underpinning population divergence in the sea has several evolutionary and practical implications. From an evolutionary point of view, this knowledge will expand our understanding on how biodiversity arises in the sea, and how adaptation occurs despite gene flow. From a conservation and management perspective, this knowledge will help in the implementation of effective conservation plans and sustainable fisheries. A better understanding of the spatial scale of population subdivision will improve fish stock delineation, so it reflects the biological complexity of a species. Moreover, genetic tools that facilitate the assessment of the relative abundance of population components will help resolve the composition of mixed stocks, and as such, they can help in the prioritization for the protection of vulnerable population components. Ultimately, improved conservation and management practices ideally informed by genetics, will contribute to the preservation of species, and of the ecosystem functions, livelihood and commercial activities derived from them.

6.2 Main conclusions

In chapter 2, I conducted a critical evaluation of the literature on whole-genome resequencing (WGR) approaches and their uses in population genomics to identify advantages and limitations of the current techniques, discuss potential solutions, and explore areas of conservation biology and management to which the analysis of whole genomes could make important contributions. I found that the main benefits of using WGR are the possibility of screening neutral and functional genetic variation at the highest genomic resolution possible and in diverse variant types, from single mutations to structural modifications. The augmented marker density therefore, increases the power to detect genomic signatures of selection necessary for the identification of local adaptation and of the genetic basis of phenotypic traits. The comparison of the two commonly used WGR methods in population genomics, low coverage whole-genome resequencing of numerous individuals (i.e., lcWGR), and high coverage whole-genome resequencing of pooled DNA of numerous individuals (i.e., Pool-seq), indicated that the implementation of barcoded individuals in lcWGR is preferred, as opposed to using a single barcode for a pool of individuals as in Pool-seq, particularly in species with unresolved population structure. The reason being that mixed samples and potential migrants can be identified when individual sequence reads are distinguishable. Yet, both techniques provide equivalent high genomic resolution and have their own limitations. The lack of a reference genome, the relatively high cost of high-throughput sequencing, the need for bioinformatic expertise and for large computer resources, however, are some of the main factors limiting the application of WGR in the study of non-model species. The question then becomes: When is the investment in WGR justified (either lcWGR or Pool-seq)? It entirely depends on the main goal of the research and on the availability of genomic resources for the species of interest. Assuming funding and a reference genome are not restrictive, WGR would be the most powerful technique to detect signatures of selection, to resolve population structure mediated by natural selection, to identify the genetic basis of phenotypic traits and diseases, and to obtain a detailed reconstruction of species' historical demography. Studies focused on resolving species' population structure derived from genetic drift or on the estimation of effective population size using traditional methods, are some of the examples in which having high marker density does not necessarily provide an added benefit or even, could be problematic (Waples et al., 2016). Consequently, the justification for using WGR in my thesis resided in the need of screening adaptive and neutral variation in a marine fish species characterized by high gene flow, large effective population size, minute effect of genetic drift, and low population structure at neutral markers, for which the tracing of signatures of selection and local adaptation was sought. In conclusion, WGR data can help inform conservation and management plans by providing a refined assessment of intraspecific genetic diversity.

In Chapter 3, I analyzed Pool-seq data of 6 NW and 19 NE Atlantic herring populations to evaluate if genetic differences exist between spring and fall spawners in the NW Atlantic, as previously found in NE Atlantic populations (Martinez Barrio et al., 2016), and to estimate to what extent such differences are shared between populations at each side of the ocean. I discovered that indeed, significant genetic divergence exists at putatively adaptive loci between spring and fall spawners in the NW Atlantic, despite low genetic differentiation at neutral loci. In addition, I found that, to a large extent, the genetic factors underlying such genetic divergence are shared with populations in the NE Atlantic Ocean and the Baltic Sea. This result provides evidence for parallel evolution of adaptations to seasonal reproduction that most likely result from standing genetic variation predating the last glacial maximum. Many of the shared loci are near genes with a known role in reproduction, such as the thyroid-stimulating hormone receptor (*TSHR*), the SOX11 transcription factor (*SOX11*), calmodulin (*CALM*), and the estrogen receptor

2 (*ESR2A*). Two SNPs near the *TSHR* gene exhibited the strongest association with seasonal reproduction in both, the NW and NE Atlantic populations. Overall, these results indicate that the large population sizes, wide geographic distribution, and available genomic resources make Atlantic herring an ideal species for the study of ecological adaptation, given the negligible participation of genetic drift in shaping patterns of population divergence, which opens up an opportunity for the identification of the genetic basis of local adaptation in a widely distributed and highly migratory marine fish. Notably, in addition to the shared loci, numerous loci strongly associated with seasonal reproduction were uniquely found in NW Atlantic populations, which suggests local adaptation and justify further research.

Consequently, in Chapter 4, sampling coverage in the NW Atlantic was increased, thus I analyzed Pool-seq data of a total of 14 spawning aggregations distributed across the reproductive range of the species in the region. The goals of this chapter were to examine fine-scale patterns of genomic differentiation, to identify their genomic basis, and to determinate their possible association with environmental conditions in the sea. As in the previous chapter, I found fine-scale population structure at putatively adaptive loci, notwithstanding low genetic differentiation at neutral loci. In addition, I disentangled an intricate pattern of population subdivision mainly driven by two factors, seasonal reproduction (discriminating between spring and fall spawners) and a latitudinal cline (distinguishing northern and southern spawning aggregations). Each pattern of divergence was underlined by thousands of outlier SNPs putatively under selection that were characterized by exhibiting opposite alleles close to fixation between the contrasting genetic groups (e.g. spring vs. fall, and north vs. south). These genetic variants were distributed in particular parts of the genome spanning a specific set of genes, forming so-called "genomic regions or islands of divergence" (i.e. sections of the genome that underlie reproductive isolation or adaptation and appear to be resistant to gene flow (Nosil et al., 2009)). Winter sea-surface temperature appears to be the best environmental predictor of the latitudinal differentiation. Overall, the results of this chapter revealed that seasonal reproduction and latitudinal spawning location are features under disruptive selection leading to local adaptation in herring.

Taking advantage of the whole genome data generated in previous chapters and using state-of-the-art machine learning algorithms, in Chapter 5, I developed two highly informative and cost-effective SNP panels that allow for the genetic identification of spawning season (SPW-panel) and latitudinal origin (LAT-panel) of northwest Atlantic herring. Based on the individual genotypes obtained with these panels, I examined spatial and temporal variation of SNP allele frequencies in 993 individuals from 30 sites, including spawning aggregations and inshore and offshore mixed aggregations, distributed thorough the reproductive range of the species in the western Atlantic.

The high self-assignment accuracy (>85%) obtained for both panels through cross-validation simulations based on machine learning algorithms, together with the confirmation of phenotype-genotype agreement (i.e. gonadal maturity status at the date of capture match genetic-based spawning season assignment), demonstrate the high predictive power of these SNP panels. The genetic data obtained in this chapter confirmed a southwest-northeast gradient in the prevalence of spawning types. This data also provided genetic evidence supporting the hypothesis that the diverse reproductive strategies in herring have an adaptive basis for increasing the chances of adult reproductive success and offspring survival under stochastic changes in the environment. The analysis of temporal replicates indicated stability in allele frequency differences between spawning types and in fall spawners from the northern region for the time period covered (1 to 9 years). Short-term temporal persistence of the genetic differences between reproductive strategies despite their mixing outside of the spawning season, suggests spawning time and site fidelity as well as stability of selective pressures shaping the current patterns of divergence in an ecological time scale. I found that hybridization between reproductive and latitudinal components is not restricted. Hybrids between spawning seasons constituted a varying proportion of spawning aggregations (13.3-60.0% in spring- and 0.0-26.7% in fall-spawning grounds, respectively). Hybrids between latitudinal regions were more predominant towards the south (21.4-55.6%) than in the north (0.0-20.0 %). The varying proportion of hybrids found across the reproductive and latitudinal components suggests that perhaps there is a reproductive advantage of being pure-spring or pure-fall spawner in some locations but not in others, that polymorphism at loci related to the latitudinal cline increases reproductive success in the southern region, and that variable levels of connectivity between regions may exist. Moreover, intermediate to high admixture levels were commonly observed in individuals that spawned in either season, suggesting the possibility that some individuals hatched in one season are capable of spawning in the other season. Further functional studies are necessary to corroborate this observation. The analysis of mixture samples demonstrated the utility of the SNP panels developed here for the assessment of mixed stock composition. For example, it was possible to determine that groups of fish spawning in different seasons and latitudinal regions congregate in both, the offshore aggregations in southwest Newfoundland and in mid-Nova Scotia. Overall, the results of this chapter corroborate the highly complex mating and migratory system of herring as well as the mixed composition of aggregations outside of the spawning season. These results also highlight the significance of preserving biologically relevant genetic diversity for long-term species and fisheries persistence, and demonstrate the utility of the genetic tools developed here for the close to real time assessment of the composition of mixed stocks at any life stage.

In conclusion, with this thesis I demonstrate the utility of studying adaptive and neutral genetic variation at the whole genome level to increase our understanding of patterns and processes involved in the arising of population divergence and local adaptation in a highly migratory and abundant marine fish.

6.3 Implications

The findings of this thesis have numerous implications and potential applications in NW Atlantic herring fisheries management and, by extension, to other marine species. First, the SNP panels made available here can facilitate the assessment of the composition of mixed stock fisheries, spawning grounds, offshore aggregations, and larval retention areas. These genetic tools can diagnose reproductive season and latitudinal origin with high accuracy at any life stage from a small well-preserved tissue sample, for an affordable cost and a small fraction of the time that would take to process a large number of samples using traditional methods. Genetically informed fisheries management thus, will help implement more sustainable fishing practices that, for instance, adjust fishing intensity according to the abundance of relevant biological components in a given stock. Secondly, the discovery of diverse reproductive and latitudinal components throughout the reproductive range of the species in the NW Atlantic requires revision of current management units to reflect the biological complexity of the species. Thirdly, the mismatch between the genetic data and the current method used for spawning season assignment (based on the comparison of month of capture and gonadal maturity stage) (McQuinn, 1987; LeBlanc et al., 2008) indicates the latter as well as the July 1st cut off used for spring- and fall- spawning designation of individuals should be revised. Fourthly, the genetic data here obtained are not in complete agreement with the currently proposed herring population models, suggesting they could be revised or a holistic model could be proposed in light of the new findings. Fifthly, as fishing targeting spawning grounds can have negative effects on species recovery and persistence (van Overzee & Rijnsdorp, 2015), it will be important to revise current practices and perhaps consider the implementation of rotational closures. Lastly, the findings of this thesis highlight the critical importance of characterizing and maintaining intraspecific genetic diversity of a commercially harvested species, as it constitutes the required knowledge for the implementation of sustainable fishing practices that assure the long-term persistence of the species and of the economic activity that it sustains.

6.4 Limitations

Despite the valuable findings described in this thesis, there are some limitations that need to be addressed. For example, population divergence was defined by genetic variation at loci exhibiting large allele frequency differences, meaning that variation characterized by more subtle changes in allele frequencies (commonly observed in polygenic traits) was excluded. The nature of the Pool-seq method did not provide the confidence to explore such genetic variation, given the uncertainty in the equal representation of individual DNA in a pool. Also, DNA pooling implies that individuals collected at the same time and location belong to the same genetic pool or breeding group. Certainly, this could not be assured with the methods available at the time of collection. The co-occurrence of spring and fall spawners in a pool was attempted to be controlled by selecting individuals with mature gonads at the time of capture (which would be considered as "spawners of the season”). Even if spawners of the same season were assured, it would not be possible to determine whether they come from different regions, as observed in fall spawners from the north and south. Thus, either individual targeted sequencing or genotyping, or the use of individual barcodes in whole genome sequencing are preferred. Regardless, the Pool-seq data allowed the discovery of general patterns of genetic divergence between reproductive strategies and geographic regions in herring. Despite the significant increase in sampling coverage achieved in this thesis, some locations remain understudied. For example, no temporal replicates were analyzed for the southern region, nor were spring-spawning aggregations from northeast Newfoundland examined. Finally, the SNP panels developed in this thesis are only diagnostic of the adaptive genetic variation distinguishing the reproductive and latitudinal components revealed by Pool-seq data. Thus, further studies using different or a combination of sets of markers are necessary to detect genetic differences in a smaller spatial scale, if possible.

6.5 Future research

The findings achieved in this thesis set the foundation for further research that will help expand our knowledge on the biology and population dynamics of Atlantic herring. For instance, the analysis of additional samples from the southern region will allow to examine whether the genetic differences observed among spawning aggregations in the south are temporally stable. In fact, the analysis of time series samples will be ideal, as it would be possible to monitor temporal changes in mixed stock composition as well as fluctuations of the latitudinal genetic cline and associated oceanographic variables. Another interesting venue will be the genetic study of stocks at different life stages, for example in eggs, larvae, and adults. This kind of study will help elucidate at which period of the herring life cycle natural selection mainly acts on. Moreover, functional experiments are required to establish a direct link between the genetic variants identified here and specific phenotypic traits and to confirm that hybrids can spawn either in spring or fall and that minimum temperature regimes are involved in the shaping of the latitudinal genetic cline. Finally, the genetic tools developed here, can certainly facilitate fisheries stock assessment. Thus, the implementation of a high-throughput genetic diagnosis method will significantly reduce sequencing costs and will provide close to real time assessment of stock composition, facilitating the work of fisheries managers.

6.6 References

Hilborn, R., Quinn, T. P., Schindler, D. E., & Rogers, D. E. (2003). Biocomplexity and fisheries sustainability. *Proceedings of the National Academy of Sciences*, *100*(11), 6564–6568. https://doi.org/10.1073/pnas.1037274100

Martinez Barrio, A., Lamichhaney, S., Fan, G., Rafati, N., Pettersson, M., Zhang, H., … Andersson, L. (2016). The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. *ELife*, *5*(MAY2016), 1–32. https://doi.org/10.7554/eLife.12081

Nosil, P., Funk, D. J., & Ortiz-Barrientos, D. (2009). Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, *18*(3), 375–402. https://doi.org/10.1111/j.1365-294X.2008.03946.x

Ruzzante, D. E., Mariani, S., Bekkevold, D., André, C., Mosegaard, H., Clausen, L. A. W., … Carvalho, G. R. (2006). Biocomplexity in a highly migratory pelagic marine fish, Atlantic herring. *Proceedings of the Royal Society B: Biological Sciences*, *273*(1593), 1459–1464. https://doi.org/10.1098/rspb.2005.3463

Schindler, D. E., Hilborn, R., Chasco, B., Boatright, C. P., Quinn, T. P., Rogers, L. A., & Webster, M. S. (2010). Population diversity and the portfolio effect in an exploited species. *Nature*, *465*(7298), 609–612. https://doi.org/10.1038/nature09060

van Overzee, H. M. J., & Rijnsdorp, A. D. (2015). Effects of fishing during the spawning period: implications for sustainable management. *Reviews in Fish Biology and Fisheries*, *25*(1), 65–83. https://doi.org/10.1007/s11160-014-9370-x

Waples, R. K., Larson, W. A., & Waples, R. S. (2016). Estimating contemporary effective population size in non-model species using linkage disequilibrium across thousands of loci. *Heredity*, *117*(4), 233–240. https://doi.org/10.1038/hdy.2016.60

BIBLIOGRAPHY

Adey, A., Burton, J. N., Kitzman, J. O., Hiatt, J. B., Lewis, A. P., Martin, B. K., … Shendure, J. (2013). The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature*, *500*(7461), 207–211. https://doi.org/10.1038/nature12064

Agarwala, V., Flannick, J., Sunyaev, S., & Altshuler, D. (2013). Evaluating empirical bounds on complex disease genetic architecture. *Nature Genetics*, *45*(12), 1418–1427. https://doi.org/10.1038/ng.2804

Aird, D., Ross, M. G., Chen, W., Danielsson, M., Fennell, T., Russ, C., … Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, *12*(2), R18. https://doi.org/10.1186/gb-2011-12-2-r18

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews. Genetics*, *12*(5), 363–376. https://doi.org/10.1038/nrg2958

Allendorf, F. W. (2016). Genetics and the conservation of natural populations: Allozymes to genomes. *Molecular Ecology*, *38*(1), 42–49. https://doi.org/10.1111/mec.13948

Allendorf, F. W., Hohenlohe, P. A., & Luikart, G. (2010). Genomics and the future of conservation genetics. *Nature Reviews Genetics*, *11*(10), 697–709. https://doi.org/10.1038/nrg2844

Allendorf, F. W., Luikart, G., & Aitken, S. N. (2013). *Conservation and the Genetics of Populations* (2nd ed.). Wiley-Blackwell.

Anderson, E. C. (2010). Assessing the power of informative subsets of loci for population assignment: standard methods are upwardly biased. *Molecular Ecology Resources*, *10*(4), 701–710. https://doi.org/10.1111/j.1755-0998.2010.02846.x

Anderson, E. C., Skaug, H. J., & Barshis, D. J. (2014). Next-generation sequencing for molecular ecology: a caveat regarding pooled samples. *Molecular Ecology*, *23*(3), 502–512. https://doi.org/10.1111/mec.12609

Anderson, E. C., Waples, R. S., & Kalinowski, S. T. (2008). An improved method for predicting the accuracy of genetic stock identification. *Canadian Journal of Fisheries and Aquatic Sciences*, *65*(7), 1475–1486. https://doi.org/10.1139/F08-049

Andersson, L., Ryman, N., Rosenberg, R., & Ståhl, G. (1981). Genetic variability in Atlantic herring (Clupea harengus harengus): description of protein loci and population data. *Hereditas*, *95*(1), 69–78. https://doi.org/10.1111/j.1601-5223.1981.tb01330.x

Andersson, L. S., Larhammar, M., Memic, F., Wootz, H., Schwochow, D., Rubin, C.-J., … Kullander, K. (2012). Mutations in DMRT3 affect locomotion in horses and spinal circuit function in mice. *Nature*, *488*(7413), 642–646. https://doi.org/10.1038/nature11399

André, C., Larsson, L. C., Laikre, L., Bekkevold, D., Brigham, J., Carvalho, G. R., … Ryman, N. (2011). Detecting population structure in a high gene-flow species, Atlantic herring (Clupea harengus): direct, simultaneous evaluation of neutral vs putatively selected loci. *Heredity*, *106*(2), 270–280. https://doi.org/10.1038/hdy.2010.71

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, *17*(2), 81–92. https://doi.org/10.1038/nrg.2015.28

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

Andrews, Simon. (2010). FastQC: a quality control tool for high throughput sequence data. Retrieved May 1, 2018, from http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Angeloni, F., Wagemaker, N., Vergeer, P., & Ouborg, J. (2012). Genomic toolboxes for conservation biologists. *Evolutionary Applications*, *5*(2), 130–143. https://doi.org/10.1111/j.1752-4571.2011.00217.x

Arnold, B., Corbett-Detig, R. B., Hartl, D., & Bomblies, K. (2013). RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*, *22*(11), 3179–3190. https://doi.org/10.1111/mec.12276

Auer, P. L., & Lettre, G. (2015). Rare variant association studies: considerations, challenges and opportunities. *Genome Medicine*, *7*(1), 16. https://doi.org/10.1186/s13073-015-0138-2

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Avise, J. C. (2010). Perspective: Conservation genetics enters the genomics era. *Conservation Genetics*, *11*(2), 665–669. https://doi.org/10.1007/s10592-009-0006-y

Aykanat, T., Lindqvist, M., Pritchard, V. L., & Primmer, C. R. (2016). From population genomics to conservation and management: a workflow for targeted analysis of markers identified using genome-wide approaches in Atlantic salmon Salmo salar. *Journal of Fish Biology*, *89*(6), 2658–2679. https://doi.org/10.1111/jfb.13149

Ayllon, F., Kjærner-Semb, E., Furmanek, T., Wennevik, V., Solberg, M. F., Dahle, G., … Wargelius, A. (2015). The vgll3 Locus Controls Age at Maturity in Wild and Domesticated Atlantic Salmon (Salmo salar L.) Males. *PLoS Genetics*, *11*(11), 1–15. https://doi.org/10.1371/journal.pgen.1005628

Baird, N., Etter, P., Atwood, T., Currey, M., Shiver, A., Lewis, Z., … Johnson, E. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS One*, *3*(10), e3376. https://doi.org/10.1371/journal.pone.0003376

Baker, M. (2012). De novo genome assembly: what every biologist should know. *Nature Methods*, *9*(4), 333–337. https://doi.org/10.1038/nmeth.1935

Barrett, R. D. H., & Hoekstra, H. E. (2011). Molecular spandrels: tests of adaptation at the genetic level. *Nature Review Genetics*, *12*(11), 767–780. https://doi.org/10.1038/nrg3015

Barson, N. J., Aykanat, T., Hindar, K., Baranski, M., Bolstad, G. H., Fiske, P., … Primmer, C. R. (2015). Sex-dependent dominance at a single locus maintains variation in age at maturity in salmon. *Nature*, *528*(7582), 405–408. https://doi.org/10.1038/nature16062

Baum, B. R. (1989). PHYLIP: Phylogeny Inference Package. Version 3.2 . Joel Felsenstein. *The Quarterly Review of Biology*, *64*(4), 539–541. https://doi.org/10.1086/416571

Baxter, I. G. (1959). Fecundities of winte-spring and summer-autumn herring spawners. *J. Cons. Int. Explor. Mer.*, *25*, 73–80.

Baym, M., Kryazhimskiy, S., Lieberman, T. D., Chung, H., Desai, M. M., & Kishony, R. K. (2015). Inexpensive multiplexed library preparation for megabase-sized genomes. *PLoS ONE*, *10*(5), 1–15. https://doi.org/10.1371/journal.pone.0128036

Beissinger, T. M., Rosa, G. J., Kaeppler, S. M., Gianola, D., & de Leon, N. (2015). Defining window-boundaries for genomic analyses using smoothing spline techniques. *Genetics Selection Evolution*, *47*(1), 30. https://doi.org/10.1186/s12711-015-0105-9

Bekkevold, D., Helyar, S. J., Limborg, M. T., Nielsen, E. E., Hemmer-Hansen, J., Clausen, L. A. W., & Carvalho, G. R. (2015a). Gene-associated markers can assign origin in a weakly structured fish, Atlantic herring. *ICES Journal of Marine Science: Journal Du Conseil*, *72*(6), 1790–1801. https://doi.org/10.1093/icesjms/fsu247

Bekkevold, D., Helyar, S. J., Limborg, M. T., Nielsen, E. E., Hemmer-Hansen, J., Clausen, L. A. W., & Carvalho, G. R. (2015b). Gene-associated markers can assign origin in a weakly structured fish, Atlantic herring. *ICES Journal of Marine Science: Journal Du Conseil*, *72*(6), 1790–1801. https://doi.org/10.1093/icesjms/fsu247

Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q. B., Antipenko, A., … Abel, L. (2015). Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proceedings of the National Academy of Sciences*, *112*(17), 5473–5478. https://doi.org/10.1073/pnas.1418631112

Benestan, L., Gosselin, T., Perrier, C., Sainte-Marie, B., Rochette, R., & Bernatchez, L. (2015). RAD genotyping reveals fine-scale genetic structuring and provides powerful population assignment in a widely distributed marine species, the American lobster (Homarus americanus). *Molecular Ecology*, *24*(13), 3299–3315. https://doi.org/10.1111/mec.13245

Benestan, L. M., Ferchaud, A.-L., Hohenlohe, P. A., Garner, B. A., Naylor, G. J. P., Baums, I. B., … Luikart, G. (2016). Conservation genomics of natural and managed populations: building a conceptual and practical framework. *Molecular Ecology*, *25*(13), 2967–2977. https://doi.org/10.1111/mec.13647

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B*, *57*(1), 289–300. https://doi.org/10.2307/2346101

Benoît, H., Swain, D., Hutchings, J., Knox, D., Doniol-Valcroze, T., & Bourne, C. (2018). Evidence for reproductive senescence in a broadly distributed harvested marine fish. *Marine Ecology Progress Series*, *592*, 207–224. https://doi.org/10.3354/meps12532

Berg, F., Almeland, O. W., Skadal, J., Slotte, A., Andersson, L., & Folkvord, A. (2018). Genetic factors have a major effect on growth, number of vertebrae and otolith shape in Atlantic herring (Clupea harengus). *PLOS ONE*, *13*(1), e0190995. https://doi.org/10.1371/journal.pone.0190995

Berg, J. J., & Coop, G. (2014). A Population Genetic Signal of Polygenic Adaptation. *PLoS Genetics*, *10*(8), e1004412. https://doi.org/10.1371/journal.pgen.1004412

Bergland, A. O., Behrman, E. L., O'Brien, K. R., Schmidt, P. S., & Petrov, D. A. (2014). Genomic Evidence of Rapid and Stable Adaptive Oscillations over Seasonal Time Scales in Drosophila. *PLoS Genetics*, *10*(11), e1004775. https://doi.org/10.1371/journal.pgen.1004775

Berlin, K., Koren, S., Chin, C.-S., Drake, J. P., Landolin, J. M., & Phillippy, A. M. (2015). Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature Biotechnology*, *33*(6), 623–630. https://doi.org/10.1038/nbt.3238

Bernatchez, L. (2016). On the maintenance of genetic variation and adaptation to environmental change: considerations from population genomics in fishes. *Journal of Fish Biology*, 1–38. https://doi.org/10.1111/jfb.13145

Bernatchez, Louis, Wellenreuther, M., Araneda, C., Ashton, D. T., Barth, J. M. I., Beacham, T.

D., … Withler, R. E. (2017). Harnessing the Power of Genomics to Secure the Future of Seafood. *Trends in Ecology & Evolution*, *32*(9), 665–680. https://doi.org/10.1016/j.tree.2017.06.010

Bi, K., Vanderpool, D., Singhal, S., Linderoth, T., Moritz, C., & Good, J. M. (2012). Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics*, *13*(1), 403. https://doi.org/10.1186/1471-2164-13-403

Bickhart, D. M., Rosen, B. D., Koren, S., Sayre, B. L., Hastie, A. R., Chan, S., … Smith, T. P. L. (2017). Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nature Genetics*, *53*(9), 1689–1699. https://doi.org/10.1038/ng.3802

Bierne, N., Welch, J., Loire, E., Bonhomme, F., & David, P. (2011). The coupling hypothesis: Why genome scans may fail to map local adaptation genes. *Molecular Ecology*, *20*(10), 2044–2072. https://doi.org/10.1111/j.1365-294X.2011.05080.x

Blanquart, F., Kaltz, O., Nuismer, S. L., & Gandon, S. (2013). A practical guide to measuring local adaptation. *Ecology Letters*, *16*(9), 1195–1205. https://doi.org/10.1111/ele.12150

Bleidorn, C. (2016). Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. *Systematics and Biodiversity*, *14*(1), 1–8. https://doi.org/10.1080/14772000.2015.1099575

Blischak, P. D., Wenzel, A. J., & Wolfe, A. D. (2014). Gene Prediction and Annotation in Penstemon (Plantaginaceae): A Workflow for Marker Development from Extremely Low-Coverage Genome Sequencing. *Applications in Plant Sciences*, *2*(12), 1400044. https://doi.org/10.3732/apps.1400044

Boitard, S., Rodríguez, W., Jay, F., Mona, S., & Austerlitz, F. (2016). Inferring Population Size History from Large Samples of Genome-Wide Molecular Data - An Approximate Bayesian Computation Approach. *PLOS Genetics*, *12*(3), e1005877. https://doi.org/10.1371/journal.pgen.1005877

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bondesson, M., Hao, R., Lin, C.-Y., Williams, C., & Gustafsson, J.-Å. (2015). Estrogen receptor signaling during vertebrate development. *Biochim Biophys Acta*, *1849*(2), 142–151. https://doi.org/http://dx.doi.org/10.1016/j.bbagrm.2014.06.005

Bossdorf, O., Richards, C. L., & Pigliucci, M. (2008). Epigenetics for ecologists. *Ecology Letters*, *11*(2), 106–115. https://doi.org/10.1111/j.1461-0248.2007.01130.x

Bourne, C., Mowbray, F., Squires, B., & Koen-Alonso, M. (2018). 2017 Assessment of Newfoundland east and south coast Atlantic herring (Clupea harengus) stock complexes. *DFO Can. Sci. Advis. Sec. Res. Doc. 2018/026*, v + 45 p.

Bradbury, I. R., Hamilton, L. C., Dempson, B., Robertson, M. J., Bourret, V., Bernatchez, L., & Verspoor, E. (2015). Transatlantic secondary contact in Atlantic Salmon, comparing microsatellites, a single nucleotide polymorphism array and restriction-site associated DNA sequencing for the resolution of complex spatial structure. *Molecular Ecology*, *24*(20), 5130–5144. https://doi.org/10.1111/mec.13395

Bradbury, I. R., Hamilton, L. C., Sheehan, T. F., Chaput, G., Robertson, M. J., Dempson, J. B., … Bernatchez, L. (2016). Genetic mixed-stock analysis disentangles spatial and temporal variation in composition of the West Greenland Atlantic Salmon fishery. *ICES Journal of Marine Science*, *73*(9), 2311–2321. https://doi.org/10.1093/icesjms/fsw072

Bradbury, I. R., Hubert, S., Higgins, B., Borza, T., Bowman, S., Paterson, I. G., … Bentzen, P. (2010). Parallel adaptive evolution of Atlantic cod on both sides of the Atlantic Ocean in response to temperature. *Proceedings of the Royal Society of London B: Biological Sciences*, *277*(1701), 3725–3734. https://doi.org/10.1098/rspb.2010.0985

Bradbury, I. R., Hubert, S., Higgins, B., Bowman, S., Borza, T., Paterson, I. G., … Bentzen, P. (2013). Genomic islands of divergence and their consequences for the resolution of spatial structure in an exploited marine fish. *Evolutionary Applications*, *6*(3), 450–461. https://doi.org/10.1111/eva.12026

Bradford, R. G., & Iles, T. D. (1992). Unique biological characteristics of spring-spawning herring (Clupea harengus L.) in Minas Basin, Nova Scotia, a tidally dynamic environment. *Canadian Journal of Zoology*, *70*, 641–648.

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., … Korf, I. F. (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, *2*(1), 10. https://doi.org/10.1186/2047-217X-2-10

Breiman, L. (2001). Random Forest. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Brickman, D., Hebert, D., & Wang, Z. (2018). Mechanism for the recent ocean warming events on the Scotian Shelf of eastern Canada. *Continental Shelf Research*, *156*, 11–22. https://doi.org/10.1016/j.csr.2018.01.001

Brieuc, M. S. O., Waters, C. D., Drinan, D. P., & Naish, K. A. (2018). A practical introduction to Random Forest for genetic association studies in ecology and evolution. *Molecular Ecology Resources*, *18*(4), 755–766. https://doi.org/10.1111/1755-0998.12773

Britten, G. L., Dowd, M., & Worm, B. (2016). Changing recruitment capacity in global fish stocks. *Proceedings of the National Academy of Sciences*, *113*(1), 134–139. https://doi.org/10.1073/pnas.1504709112

Broad Institute. (2014). Calling variants on cohorts of samples using the HaplotypeCaller in GVCF mode. Retrieved May 20, 2018, from https://software.broadinstitute.org/gatk/documentation/article.php?id=3893

Broad Institute. (2016). Understanding and adapting the generic hard-filtering recommendations. Retrieved May 20, 2018, from https://gatkforums.broadinstitute.org/gatk/discussion/6925/understanding-and-adapting-the-generic-hard-filtering-recommendations

Broad Institute. (2018). Picard tools. Retrieved May 20, 2018, from http://broadinstitute.github.io/picard/

Browning, S. R., & Browning, B. L. (2011). Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, *12*(10), 703–714. https://doi.org/10.1038/nrg3054

Browning, S. R., & Browning, B. L. (2016). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, *81*(5), 1084–1097. https://doi.org/10.1086/521987

Bucholtz, R. H., Tomkiewicz, J., & Dalskov, J. (2008). Manual to determine gonadal maturity of herring (Clupea harengus L.). In *DTU Aqua-report*. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Manual+to+determine+gonadal+maturity+of+herring+(Clupea+harengus+L.)#0

Buels, R., Yao, E., Diesh, C. M., Hayes, R. D., Munoz-Torres, M., Helt, G., … Holmes, I. H. (2016). JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biology*, *17*(1), 66. https://doi.org/10.1186/s13059-016-0924-1

Buerkle, A. C. (2005). Maximum-likelihood estimation of a hybrid index based on molecular markers. *Molecular Ecology Notes*, *5*(3), 684–687. https://doi.org/10.1111/j.1471-8286.2005.01011.x

Buerkle, A. C., & Gompert, Z. (2013). Population genomics based on low coverage sequencing: how low should we go? *Molecular Ecology*, *22*(11), 3028–3035. https://doi.org/10.1111/mec.12105

Bustamante, C. D., Wakeley, J., Sawyer, S., & Hartl, D. L. (2001). Directional selection and the site-frequency spectrum. *Genetics*, *159*(4), 1779–1788.

Cariou, M., Duret, L., & Charlat, S. (2016). How and how much does RAD-seq bias genetic diversity estimates? *BMC Evolutionary Biology*, *16*(1), 240. https://doi.org/10.1186/s12862-016-0791-0

Carneiro, M., Rubin, C.-J., Di Palma, F., Albert, F. W., Alfoldi, J., Barrio, A. M., … Andersson, L. (2014). Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. *Science*, *345*(6200), 1074–1079. https://doi.org/10.1126/science.1253714

Catchen, J. M., Hohenlohe, P. A., Bernatchez, L., Funk, W. C., Andrews, K. R., & Allendorf, F. W. (2017). Unbroken: RADseq remains a powerful tool for understanding the genetics of adaptation in natural populations. *Molecular Ecology Resources*, *38*(1), 42–49. https://doi.org/10.1111/1755-0998.12669

Chakraborty, M., Baldwin-Brown, J. G., Long, A. D., & Emerson, J. J. (2016). Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Research*, *44*(19), 1–12. https://doi.org/10.1093/nar/gkw654

Chan, C. X., & Ragan, M. A. (2013). Next-generation phylogenomics. *Biology Direct*, *8*(1), 3. https://doi.org/10.1186/1745-6150-8-3

Chen, K.-Y., Marschall, E. A., Sovic, M. G., Fries, A. C., Gibbs, H. L., & Ludsin, S. A. (2018). assignPOP: An r package for population assignment using genetic, non-genetic, or integrated data in a machine-learning framework. *Methods in Ecology and Evolution*, *9*(2), 439–446. https://doi.org/10.1111/2041-210X.12897

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., … Ruden, D. M. (2012a). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., … Ruden, D. M. (2012b). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2009). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, *38*(6), 1767–1771. https://doi.org/10.1093/nar/gkp1137

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., … Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*(1), 13. https://doi.org/10.1186/s13059-016-0881-8

Conover, D. O. (1998). Local adaptation in marine fishes -evidence and implications for stock enhancement. *Bulletin of Marine Science*, *62*(2), 477–493.

Crawford, J. E., Riehle, M. M., Markianos, K., Bischoff, E., Guelbeogo, W. M., Gneme, A., … Lazzaro, B. P. (2016). Evolution of GOUNDRY, a cryptic subgroup of Anopheles gambiae s.l., and its impact on susceptibility to Plasmodium infection. *Molecular Ecology*, *25*(7), 1494–1510. https://doi.org/10.1111/mec.13572

Cushing, D. H. (1967). The grouping of herring populations. *J. Mar. Biol. Ass., U.K*, *47*, 193–208.

Cutler, D. J., & Jensen, J. D. (2010). To pool, or not to pool? *Genetics*, *186*(1), 41–43. https://doi.org/10.1534/genetics.110.121012

da Fonseca, R. R., Albrechtsen, A., Themudo, G. E., Ramos-Madrigal, J., Sibbesen, J. A., Maretty, L., … Pereira, R. J. (2016). Next-generation biology: Sequencing and data analysis approaches for non-model organisms. *Marine Genomics*, *30*, 1–11. https://doi.org/10.1016/j.margen.2016.04.012

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

De Wit, P., Pespeni, M. H., Ladner, J. T., Barshis, D. J., & Palumbi, S. R. (2012). The simple fool' s guide to population genomics via RNA-Seq : an introduction to high-throughput sequencing data analysis. *Molecular Ecology Resources*, *12*, 1058–1067. https://doi.org/10.1111/1755-0998.12003

Delsuc, F., Brinkmann, H., & Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nature Reviews. Genetics*, *6*(5), 361–375. https://doi.org/10.1038/nrg1603

Dennenmoser, S., Vamosi, S. M., Nolte, A. W., & Rogers, S. M. (2017). Adaptive genomic divergence under high gene flow between freshwater and brackish-water ecotypes of prickly sculpin (Cottus asper) revealed by Pool-Seq. *Molecular Ecology*, *26*(1), 25–42. https://doi.org/10.1111/mec.13805

Denny, S., Clark, K. J., Power, M. J., & Stephenson, R. L. (1998). The status of the herring in the Bras d'Or Lakes in 1996–1997. *Canadian Stock Assessment Secretariat Research Document*, *80*, 1–32.

Department of Fisheries & Oceans Canada. (2012). *Assessment of Atlantic herring in the southern Gulf of St. Lawrence (NAFO Div. 4T). Available online at http://www.dfo-mpo.gc.ca/csas-sccs/Publications/SAR-AS/2012/2012_014-eng.pdf.*

Department of Fisheries and Oceans Canada. (2011). Canadian fisheries statistics 2008, Available online at http://www.dfo-mpo.gc.ca/stats/commercial/cfs/2008/CFS2008_e.pdf.

DePristo, M. a, Banks, E., Poplin, R., Garimella, K. V, Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498. https://doi.org/10.1038/ng.806

DFO. (1997). State of the Ocean: Northwest Atlantic. In *DFO Science Stock Status Report G0-01*. Retrieved from https://login.proxy.lib.duke.edu/login?url=https://search.proquest.com/docview/1668269133?accountid=10598%0Ahttp://pm6mt7vg3j.search.serialssolutions.com?ctx_ver=Z39.88-2004&ctx_enc=info:ofi/enc:UTF-8&rfr_id=info:sid/ProQ%3Aasfabiological&rft_val_fmt=info

DFO. (2017). Stock Status Update of 4VWX Herring. In *DFO Can. Sci. Advis. Sec. Sci. Resp. 2017/037*.

Dixon, P. (2003). VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science*, *14*(6), 927–930. https://doi.org/10.1111/j.1654-1103.2003.tb02228.x

Dohm, J. C., Lottaz, C., Borodina, T., & Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, *36*(16), e105–e105. https://doi.org/10.1093/nar/gkn425

Dray, S., & Dufour, A.-B. (2007). The ade4 Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software*, *22*(4). https://doi.org/10.18637/jss.v022.i04

Durbin, R. M., Altshuler, D. L., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., … Africa, W. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–1073. https://doi.org/10.1038/nature09534

Durrett, R. (2008). *Probability Models for DNA Sequence Evolution* (2nd ed.). In *Probability and Its Applications* (2nd ed.). https://doi.org/10.1007/978-0-387-78168-6

Earl, D., Bradnam, K., St. John, J., Darling, A., Lin, D., Fass, J., … Paten, B. (2011). Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Research*, *21*(12), 2224–2241. https://doi.org/10.1101/gr.126599.111

Ekblom, R., & Wolf, J. B. W. (2014). A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications*, *7*(9), 1026–1042. https://doi.org/10.1111/eva.12178

Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution*, *29*(1), 51–63. https://doi.org/10.1016/j.tree.2013.09.008

Engelhard, G. H., & Heino, M. (2004). Maturity changes in Norwegian spring-spawning herring before, during, and after a major population collapse. *Fisheries Research*, *66*(2–3), 299–310. https://doi.org/10.1016/S0165-7836(03)00195-4

Epstein, D. J. (2009). Cis-regulatory mutations in human disease. *Briefings in Functional Genomics and Proteomics*, *8*(4), 310–316. https://doi.org/10.1093/bfgp/elp021

Evans, J. D., Brown, S. J., Hackett, K. J. J., Robinson, G., Richards, S., Lawson, D., … Zhou, X. (2013). The i5K initiative: Advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *Journal of Heredity*, *104*(5), 595–600. https://doi.org/10.1093/jhered/est050

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Excoffier, L., Foll, M., & Petit, R. J. (2009). Genetic Consequences of Range Expansions. *Annual Review of Ecology, Evolution, and Systematics*, *40*(1), 481–501. https://doi.org/10.1146/annurev.ecolsys.39.110707.173414

Excoffier, L., & Lischer, H. E. L. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, *10*(3), 564–567. https://doi.org/10.1111/j.1755-0998.2010.02847.x

Fadista, J., Manning, A. K., Florez, J. C., & Groop, L. (2016). The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *European Journal of Human Genetics*, *24*(8), 1202–1205. https://doi.org/10.1038/ejhg.2015.269

FAO. (2019). Species Fact Sheets: Clupea harengus (Linnaeus, 1758). Retrieved from http://www.fao.org/fishery/species/2886/en

Feder, A. F., Petrov, D. A., & Bergland, A. O. (2012). LDx: Estimation of Linkage Disequilibrium from High-Throughput Pooled Resequencing Data. *PLoS ONE*, *7*(11), e48588. https://doi.org/10.1371/journal.pone.0048588

Feder, J. L., Egan, S. P., & Nosil, P. (2012). The genomics of speciation-with-gene-flow. *Trends in Genetics*, *28*(7), 342–350. https://doi.org/10.1016/j.tig.2012.03.009

Felsenstein, J. (1989). PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*, *5*, 164–166.

Ferretti, L., Ramos-Onsins, S. E., & Pérez-Enciso, M. (2013). Population genomics from pool sequencing. *Molecular Ecology*, *22*(22), 5561–5576. https://doi.org/10.1111/mec.12522

Fichefet, T., & Maqueda, M. A. M. (1997). Sensitivity of a global sea ice model to the treatment of ice thermodynamics and dynamics. *Journal of Geophysical Research: Oceans*, *102*(C6), 12609–12646. https://doi.org/10.1029/97JC00480

Field, Y., Boyle, E. A., Telis, N., Gao, Z., Gaulton, K. J., Golan, D., … Pritchard, J. K. (2016). Detection of human adaptation during the past 2000 years. *Science*, *354*(6313), 760–764. https://doi.org/10.1126/science.aag0776

Fischer, M. C., Rellstab, C., Leuzinger, M., Roumet, M., Gugerli, F., Shimizu, K. K., … Widmer, A. (2017). Estimating genomic diversity and population differentiation – an empirical comparison of microsatellite and SNP variation in Arabidopsis halleri. *BMC Genomics*, *18*(1), 69. https://doi.org/10.1186/s12864-016-3459-7

Fischer, M. C., Rellstab, C., Tedder, A., Zoller, S., Gugerli, F., Shimizu, K. K., … Widmer, A. (2013). Population genomic footprints of selection and associations with climate in natural populations of Arabidopsis halleri from the Alps. *Molecular Ecology*, *22*(22), 5594–5607. https://doi.org/10.1111/mec.12521

Fleming, D. S., Koltes, J. E., Fritz-Waters, E. R., Rothschild, M. F., Schmidt, C. J., Ashwell, C. M., … Lamont, S. J. (2016). Single nucleotide variant discovery of highly inbred Leghorn and Fayoumi chicken breeds using pooled whole genome resequencing data reveals insights into phenotype differences. *BMC Genomics*, *17*(1), 812. https://doi.org/10.1186/s12864-016-3147-7

Fonseca, N. A., Rung, J., Brazma, A., & Marioni, J. C. (2012). Tools for mapping high-throughput sequencing data. *Bioinformatics*, *28*(24), 3169–3177. https://doi.org/10.1093/bioinformatics/bts605

Fontanesi, L., Di Palma, F., Flicek, P., Smith, A. T., Thulin, C.-G., & Alves, P. C. (2016). LaGomiCs—Lagomorph Genomics Consortium: An International Collaborative Effort for Sequencing the Genomes of an Entire Mammalian Order. *Journal of Heredity*, *107*(4), 295–308. https://doi.org/10.1093/jhered/esw010

Foote, A. D., Vijay, N., Ávila-Arcos, M. C., Baird, R. W., Durban, J. W., Fumagalli, M., … Wolf, J. B. W. (2016). Genome-culture coevolution promotes rapid divergence of killer whale ecotypes. *Nature Communications*, *7*(May), 11693. https://doi.org/10.1038/ncomms11693

Forester, B. R., Lasky, J. R., Wagner, H. H., & Urban, D. L. (2018). Comparing methods for detecting multilocus adaptation with multivariate genotype–environment associations. *Molecular Ecology*, *27*(9), 2215–2233. https://doi.org/10.1111/mec.14584

Fracassetti, M., Griffin, P. C., & Willi, Y. (2015). Validation of pooled whole-genome resequencing in Arabidopsis lyrata. *PLoS ONE*, *10*(10), 1–15. https://doi.org/10.1371/journal.pone.0140462

Frank, K. T., & Brickman, D. (2000). Allee effects and compensatory population dynamics within a stock complex. *Canadian Journal of Fisheries and Aquatic Sciences*, *57*, 513–517.

Frankham, R. (2010). Challenges and opportunities of genetic approaches to biological conservation. *Biological Conservation*, *143*(9), 1919–1927. https://doi.org/10.1016/j.biocon.2010.05.011

Freamo, H., O'reilly, P., Berg, P. R., Lien, S., & Boulding, E. G. (2011). Outlier SNPs show more genetic structure between two Bay of Fundy metapopulations of Atlantic salmon than do neutral SNPs. *Molecular Ecology Resources*, *11*, 254–267. https://doi.org/10.1111/j.1755-0998.2010.02952.x

Fuentes-Pardo, A. P., Bourne, C., Singh, R., Emond, K., Pinkham, L., McDermid, J. L., … Ruzzante, D. E. (2019). Adaptation to seasonal reproduction and thermal minima-related factors drives fine-scale divergence despite gene flow in Atlantic herring populations. *BioRxiv*. https://doi.org/https://doi.org/10.1101/578484

Fumagalli, M. (2013). Assessing the effect of sequencing depth and sample size in population genetics inferences. *PloS One*, *8*(11), e79667. https://doi.org/10.1371/journal.pone.0079667

Fumagalli, M., Vieira, F. G., Korneliussen, T. S., Linderoth, T., Huerta-Sánchez, E., Albrechtsen, A., & Nielsen, R. (2013). Quantifying population genetic differentiation from next-generation sequencing data. *Genetics*, *195*(3), 979–992. https://doi.org/10.1534/genetics.113.154740

Fumagalli, M., Vieira, F. G., Linderoth, T., & Nielsen, R. (2014). NgsTools: Methods for population genetics analyses from next-generation sequencing data. *Bioinformatics*, *30*(10), 1486–1487. https://doi.org/10.1093/bioinformatics/btu041

Funk, W. C., McKay, J. K., Hohenlohe, P. a, & Allendorf, F. W. (2012). Harnessing genomics for delineating conservation units. *Trends in Ecology & Evolution*, *27*(9), 489–496. https://doi.org/10.1016/j.tree.2012.05.012

Fussi, B., Westergren, M., Aravanopoulos, F., Baier, R., Kavaliauskas, D., Finzgar, D., … Kraigher, H. (2016). Forest genetic monitoring: an overview of concepts and definitions. *Environmental Monitoring and Assessment*, *188*(8), 493. https://doi.org/10.1007/s10661-016-5489-7

Futschik, A., & Schlötterer, C. (2010). The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics*, *186*(1), 207–218. https://doi.org/10.1534/genetics.110.114397

Gaggiotti, O. E., Bekkevold, D., Jørgensen, H. B. H., Foll, M., Carvalho, G. R., Andre, C., & Ruzzante, D. E. (2009). Disentangling the effects of evolutionary, demographic, and environmental factors influencing genetic structure of natural populations: Atlantic herring as a case study. *Evolution*, *63*(11), 2939–2951. https://doi.org/10.1111/j.1558-5646.2009.00779.x

Gagnaire, P.-A., & Gaggiotti, O. E. (2016). Detecting polygenic selection in marine populations by combining population genomics and quantitative genetics approaches. *Current Zoology*, *62*(August), 1–14. https://doi.org/10.1093/cz/zow088

Garner, B. A., Hand, B. K., Amish, S. J., Bernatchez, L., Foster, J. T., Miller, K. M., … Luikart, G. (2016). Genomics in Conservation: Case Studies and Bridging the Gap between Data and Application. *Trends in Ecology & Evolution*, *31*(2), 81–83.

https://doi.org/10.1016/j.tree.2015.10.009

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *ArXiv Preprint ArXiv:1207.3907*, 9. https://doi.org/arXiv:1207.3907

Gautier, M., Foucaud, J., Gharbi, K., Cézard, T., Galan, M., Loiseau, A., … Estoup, A. (2013). Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. *Molecular Ecology*, *22*(14), 3766–3779. https://doi.org/10.1111/mec.12360

Gavrilets, S. (2003). Perspective: Models of speciation: what have we learned in 40 years? *Evolution*, *57*(10), 2197–2215. https://doi.org/10.1111/j.0014-3820.2003.tb00233.x

Geffen, A. J. (2009). Advances in herring biology: from simple to complex, coping with plasticity and adaptability. *ICES Journal of Marine Science*, *66*(8), 1688–1695. https://doi.org/10.1093/icesjms/fsp028

GIGA. (2014). The Global Invertebrate Genomics Alliance (GIGA): Developing Community Resources to Study Diverse Invertebrate Genomes. *Journal of Heredity*, *105*(1), 1–18. https://doi.org/10.1093/jhered/est084

Goldstein, B. A., Hubbard, A. E., Cutler, A., & Barcellos, L. F. (2010). An application of Random Forests to a genome-wide association dataset: Methodological considerations & new findings. *BMC Genetics*, *11*(1), 49. https://doi.org/10.1186/1471-2156-11-49

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333–351. https://doi.org/10.1038/nrg.2016.49

Gould, B. A., Chen, Y., & Lowry, D. B. (2017). Pooled ecotype sequencing reveals candidate genetic mechanisms for adaptive differentiation and reproductive isolation. *Molecular Ecology*, *26*(1), 163–177. https://doi.org/10.1111/mec.13881

Graham, T. R. (1962). A relationship between growth, hatching and spawning season in Canadian Atlantic herring (Clupea harengus L .). *J. Fish. Res. Bd. Canada*, *19*(5), 985–987.

Grigoriev, I. V., Nikitin, R., Haridas, S., Kuo, A., Ohm, R., Otillar, R., … Shabalov, I. (2014). MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Research*, *42*(D1), D699–D704. https://doi.org/10.1093/nar/gkt1183

Gröger, J. P., Kruse, G. H., & Rohlf, N. (2009). Slave to the rhythm: how large-scale climate cycles trigger herring (Clupea harengus) regeneration in the North Sea. *ICES Journal of Marine Science: Journal Du Conseil* . https://doi.org/10.1093/icesjms/fsp259

Grossen, C., Biebach, I., Angelone-Alasaad, S., Keller, L. F., & Croll, D. (2017). Population genomics analyses of European ibex species show lower diversity and higher inbreeding in reintroduced populations. *Evolutionary Applications*. https://doi.org/10.1111/eva.12490

Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, *32*(18), 2847–2849. https://doi.org/10.1093/bioinformatics/btw313

Guo, B., Li, Z., & Merilä, J. (2016). Population genomic evidence for adaptive differentiation in the Baltic Sea herring. *Molecular Ecology*, *25*(12), 2833–2852. https://doi.org/10.1111/mec.13657

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genetics*, *5*(10), e1000695. https://doi.org/10.1371/journal.pgen.1000695

Haasl, R. J., & Payseur, B. A. (2016). Fifteen years of genomewide scans for selection: Trends, lessons and unaddressed genetic sources of complication. *Molecular Ecology*, *25*(1), 5–23. https://doi.org/10.1111/mec.13339

Habicht, C., Munro, A., Dann, T., Eggers, D., Templin, W., Witteveen, M., … Volk, E. (2012). *Harvest and Harvest Rates of Sockeye Salmon Stocks in Fisheries of the Western Alaska Salmon Stock Identification Program (WASSIP), 2006– 2008*. Alaska, US.

Haig, S. M., Miller, M. P., Bellinger, R., Draheim, H. M., Mercer, D. M., & Mullins, T. D. (2016). The conservation genetics juggling act: integrating genetics and ecology, science and policy. *Evolutionary Applications*, *9*(1), 181–195. https://doi.org/10.1111/eva.12337

Han, E., Sinsheimer, J. S., & Novembre, J. (2015). Fast and accurate site frequency spectrum estimation from low coverage sequence data. *Bioinformatics*, *31*(5), 720–727. https://doi.org/10.1093/bioinformatics/btu725

Hanon, E. A., Lincoln, G. A., Fustin, J.-M., Dardente, H., Masson-Pévet, M., Morgan, P. J., & Hazlerigg, D. G. (2015). Ancestral TSH mechanism signals summer in a photoperiodic mammal. *Current Biology*, *18*(15), 1147–1152. https://doi.org/10.1016/j.cub.2008.06.076

Hansen, T. F. (2006). The Evolution of Genetic Architecture. *Annual Review of Ecology, Evolution, and Systematics*, *37*(1), 123–157. https://doi.org/10.1146/annurev.ecolsys.37.091305.110224

Hatem, A., Bozdağ, D., & Çatalyürek, Ü. V. (2013). Benchmarking short sequence mapping tools. *BMC Bioinformatics*, *14*(184), 1–25. Retrieved from http://www.biomedcentral.com/1471-2105/14/184%0ARESEARCH

Hauser, L., & Carvalho, G. R. (2008). Paradigm shifts in marine fisheries genetics: ugly hypotheses slain by beautiful facts. *Fish and Fisheries*, *9*(4), 333–362. https://doi.org/10.1111/j.1467-2979.2008.00299.x

Head, S. R., Komori, H. K., LaMere, S. A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D. R., & Ordoukhanian, P. (2014). Library construction for next-generation sequencing: Overviews and challenges. *BioTechniques*, *56*(2), 167–203. https://doi.org/10.2144/000114133

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003

Hedrick, P. W., Hellsten, U., & Grattapaglia, D. (2016). Examining the cause of high inbreeding depression: Analysis of whole-genome sequence data in 28 selfed progeny of Eucalyptus grandis. *New Phytologist*, *209*(2), 600–611. https://doi.org/10.1111/nph.13639

Hedrick, P. W., & Miller, P. S. (1992). Conservation Genetics: Techniques and Fundamentals. *Ecological Applications*, *2*(1), 30–46.

Hendry, A. P., & Day, T. (2005). Population structure attributable to reproductive time: isolation by time and adaptation by time. *Molecular Ecology*, *14*(4), 901–916. https://doi.org/10.1111/j.1365-294X.2005.02480.x

Hilborn, R., Quinn, T. P., Schindler, D. E., & Rogers, D. E. (2003). Biocomplexity and fisheries sustainability. *Proceedings of the National Academy of Sciences*, *100*(11), 6564–6568.

https://doi.org/10.1073/pnas.1037274100

Hivert, V. (2018). *Measuring genetic differentiation from Pool-seq data*.

Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., … Whitlock, M. C. (2016). Finding the Genomic Basis of Local Adaptation: Pitfalls, Practical Solutions, and Future Directions. *The American Naturalist*, *188*(4), 379–397. https://doi.org/10.1086/688018

Hoekstra, H. E., & Nachman, M. W. (2003). Different genes underlie adaptive melanism in different populations of rock pocket mice. *Molecular Ecology*, *12*(5), 1185–1194. https://doi.org/10.1046/j.1365-294X.2003.01788.x

Hoffmann, A. A., & Rieseberg, L. H. (2008). Revisiting the Impact of Inversions in Evolution: From Population Genetic Markers to Drivers of Adaptive Shifts and Speciation? *Annual Review of Ecology, Evolution, and Systematics*, *39*(2008), 21–42. https://doi.org/10.1146/annurev.ecolsys.39.110707.173532

Hohenlohe, P. a, Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. a, & Cresko, W. a. (2010). Population Genomics of Parallel Adaptation in Threespine Stickleback using Sequenced RAD Tags. *PLoS Genetics*, *6*(2), e1000862. https://doi.org/10.1371/journal.pgen.1000862

Holsinger, K. E., & Weir, B. S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nature Reviews. Genetics*, *10*(September), 639–650. https://doi.org/10.1038/nrg2611

Human Genome Sequencing Consortium, I. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*(7011), 931–945. https://doi.org/10.1038/nature03001

Iles, T. D., & Sinclair, M. (1982). Atlantic Herring: Stock Discreteness and Abundance. *Science*, *215*(4533), 627–633. https://doi.org/10.1126/science.215.4533.627

Ivy, J. A., Putnam, A. S., Navarro, A. Y., Gurr, J., & Ryder, O. A. (2016). Applying SNP-derived molecular coancestry estimates to captive breeding programs. *Journal of Heredity*, *107*(5), 403–412. https://doi.org/10.1093/jhered/esw029

Jarvis, E., Zhang, G., Li, C., Li, Q., Li, B., Larkin, D. M., … Froman, D. P. (2014). Comparative genomics reveals insights into avian genome evolution and adaptation. *Science*, *346*(6215), 1311–1320. https://doi.org/10.1126/science.1251385

Jean, Y. (1956). A Study of Spring and Fall Spawning Herring (Clupea Harengus L.) at Grande-Rivière, Bay of Chaleur, Québec. *Department of Fisheries Québec Constribution*, *49*, 76p.

Jeffery, N. W., Bradbury, I. R., Stanley, R. R. E., Wringe, B. F., Van Wyngaarden, M., Lowen, J. Ben, … DiBacco, C. (2018). Genomewide evidence of environmentally mediated secondary contact of European green crab (Carcinus maenas) lineages in eastern North America. *Evolutionary Applications*, *11*(6), 869–882. https://doi.org/10.1111/eva.12601

Jeffery, N. W., Wringe, B. F., McBride, M. C., Hamilton, L. C., Stanley, R. R. E., Bernatchez, L., … Bradbury, I. R. (2018). Range-wide regional assignment of Atlantic salmon (Salmo salar) using genome wide single-nucleotide polymorphisms. *Fisheries Research*, *206*, 163–175. https://doi.org/10.1016/j.fishres.2018.05.017

Jensen, J. D., Foll, M., & Bernatchez, L. (2016). The past, present and future of genomic scans for selection. *Molecular Ecology*, *25*(1), 1–4. https://doi.org/10.1111/mec.13493

Johannessen, A., Skaret, G., Langard, L., Slotte, A., Husebo, A., & Ferno, A. (2014). The dynamics of a metapopulation: changes in life-history traits in resident herring that co-occur with oceanic herring during spawning. *PloS One*, *9*(7), e102462.

https://doi.org/10.1371/journal.pone.0102462

Johnston, I. A., Vieira, V. L. A., & Temple, G. K. (2001). Functional consequences and population differences in the developmental plasticity of muscle to temperature in Atlantic herring Clupea harengus. *Marine Ecology Progress Series*, *213*, 285–300. https://doi.org/10.3354/meps213285

Jonas, A., Taus, T., Kosiol, C., Schlotterer, C., & Futschik, A. (2016). Estimating the Effective Population Size from Temporal Allele Frequency Changes in Experimental Evolution. *Genetics*, *204*(2), 723–735. https://doi.org/10.1534/genetics.116.191197

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., … Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, *484*(7392), 55–61. https://doi.org/10.1038/nature10944

Jones, M. R., & Good, J. M. (2016). Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, *25*(1), 185–202. https://doi.org/10.1111/mec.13304

Jorgensen, H. B. H., Hansen, M. M., Bekkevold, D., Ruzzante, D. E., & Loeschcke, V. (2005). Marine landscapes and population genetic structure of herring (Clupea harengus L.) in the Baltic Sea. *Molecular Ecology*, *14*(10), 3219–3234. https://doi.org/10.1111/j.1365-294X.2005.02658.x

Joron, M., Frezal, L., Jones, R. T., Chamberlain, N. L., Lee, S. F., Haag, C. R., … ffrench-Constant, R. H. (2011). Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature*, *477*(7363), 203–206. https://doi.org/10.1038/nature10341

Kaiser, T. S., Poehn, B., Szkiba, D., Preussner, M., Sedlazeck, F. J., Zrim, A., … Tessmar-Raible, K. (2016). The genomic basis of circadian and circalunar timing adaptations in a midge. *Nature*, *540*(7631), 69–73. https://doi.org/10.1038/nature20151

Kalinowski, S. T., Manlove, K. R., & Taper, M. L. (2007). *ONCOR: software for genetic stock identification*.

Kardos, M., Taylor, H. R., Ellegren, H., Luikart, G., & Allendorf, F. W. (2016). Genomics advances the study of inbreeding depression in the wild. *Evolutionary Applications*, *9*(10), 1205–1218. https://doi.org/10.1111/eva.12414

Kerr, Q., Fuentes-Pardo, A. P., Kho, J., McDermid, J. L., & Ruzzante, D. E. (2019). Temporal stability and assignment power of adaptively divergent genomic regions between herring (Clupea harengus) seasonal spawning aggregations. *Ecology and Evolution*, (9), 500–510. https://doi.org/10.1002/ece3.4768

Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLoS Computational Biology*, *8*(2), e1002375. https://doi.org/10.1371/journal.pcbi.1002375

Kim, H.-D., Choe, H. K., Chung, S., Kim, M., Seong, J. Y., Son, G. H., & Kim, K. (2011). Class-C SOX transcription factors control GnRH gene expression via the intronic transcriptional enhancer. *Molecular Endocrinology*, *25*(7), 1184–1196. https://doi.org/10.1210/me.2010-0332

Kim, S., Lohmueller, K., Albrechtsen, A., Li, Y., Korneliussen, T., Tian, G., … Nielsen, R. (2011). Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC Bioinformatics*, *12*(1), 231. https://doi.org/10.1186/1471-2105-12-231

King, T. L., Kalinowski, S. T., Schill, W. B., Spidle, A. P., & Lubinski, B. A. (2001). Population structure of Atlantic salmon (Salmo salar L.): a range-wide perspective from microsatellite DNA variation. *Molecular Ecology*, *10*(4), 807–821.

Kjærner-Semb, E., Ayllon, F., Furmanek, T., Wennevik, V., Dahle, G., Niemelä, E., … Edvardsen, R. B. (2016). Atlantic salmon populations reveal adaptive divergence of immune related genes - a duplicated genome under selection. *BMC Genomics*, *17*(1), 610. https://doi.org/10.1186/s12864-016-2867-z

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., … Wilson, R. K. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, *22*(3), 568–576. https://doi.org/10.1101/gr.129684.111

Koepfli, K., Paten, B., Genome 10K Community of Scientists, & O'Brien, S. J. (2015). The Genome 10K Project: a way forward. *Annual Review of Animal Biosciences*, *3*, 57–111. https://doi.org/10.1146/annurev-animal-090414-014900

Kofler, R., Langmuller, A. M., Nouhaud, P., Otte, K. A., & Schlotterer, C. (2016). Suitability of Different Mapping Algorithms for Genome-wide Polymorphism Scans with Pool-Seq Data. *Genes|Genomes|Genetics*, *6*(November), 1–20. https://doi.org/10.1534/g3.116.034488

Kofler, R., Orozco-terWengel, P., De Maio, N., Pandey, R. V., Nolte, V., Futschik, A., … Schlötterer, C. (2011). PoPoolation: A Toolbox for Population Genetic Analysis of Next Generation Sequencing Data from Pooled Individuals. *PLoS ONE*, *6*(1), e15925. https://doi.org/10.1371/journal.pone.0015925

Kofler, R., Pandey, R. V., & Schlötterer, C. (2011a). PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics (Oxford, England)*, *27*(24), 3435–3436. https://doi.org/10.1093/bioinformatics/btr589

Kofler, R., Pandey, R. V., & Schlötterer, C. (2011b). PoPoolation2: Identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics*, *27*(24), 3435–3436. https://doi.org/10.1093/bioinformatics/btr589

Kolaczkowski, B., Kern, A. D., Holloway, A. K., & Begun, D. J. (2011). Genomic Differentiation Between Temperate and Tropical Australian Populations of Drosophila melanogaster. *Genetics*, *187*(1), 245–260. https://doi.org/10.1534/genetics.110.123059

Korneliussen, T. S. T., Albrechtsen, A., Nielsen, R., Nielsen, R., Paul, J., Albrechtsen, A., … Ballinger, Dge. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, *15*(356), 1–13. https://doi.org/10.1186/s12859-014-0356-4

Küpper, C., Stocks, M., Risse, J. E., Remedios, N., Farrell, L. L., Mcrae, B., … Burke, T. (2015). A supergene determines highly divergent male reproductive morphs in the ruff. *Nature Publishing Group*, *48*(1), 79–83. https://doi.org/10.1038/ng.3443

Laehnemann, D., Borkhardt, A., & McHardy, A. C. (2016). Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction. *Briefings in Bioinformatics*, *17*(1), 154–179. https://doi.org/10.1093/bib/bbv029

Laikre, L., Lundmark, C., Jansson, E., Wennerström, L., Edman, M., & Sandström, A. (2016). Lack of recognition of genetic biodiversity: International policy and its implementation in Baltic Sea marine protected areas. *Ambio*, *45*(6), 661–680. https://doi.org/10.1007/s13280-016-0776-7

Lamichhaney, S., Fuentes-Pardo, A. P., Rafati, N., Ryman, N., McCracken, G. R., Bourne, C., … Andersson, L. (2017). Parallel adaptive evolution of geographically distant herring populations on both sides of the North Atlantic Ocean. *Proceedings of the National Academy of Sciences*, *114*(17), E3452–E3461. https://doi.org/10.1073/pnas.1617728114

Lamichhaney, S, Barrio, A. M., Rafati, N., Sundstrom, G., Rubin, C.-J., Gilbert, E. R., … Andersson, L. (2012). Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. *Proceedings of the National Academy of Sciences*, *109*(47), 19345–19350. https://doi.org/10.1073/pnas.1216128109

Lamichhaney, Sangeet, Berglund, J., Almén, M. S., Maqbool, K., Grabherr, M., Martinez-Barrio, A., … Andersson, L. (2015). Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature*, *518*(7539), 371–375. https://doi.org/10.1038/nature14181

Lamichhaney, Sangeet, Fan, G., Widemo, F., Gunnarsson, U., Thalmann, D. S., Hoeppner, M. P., … Andersson, L. (2015). Structural genomic changes underlie alternative reproductive strategies in the ruff (Philomachus pugnax). *Nature Genetics*, *48*(1), 84–88. https://doi.org/10.1038/ng.3430

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Larsson, L. C., Laikre, L., André, C., Dahlgren, T. G., & Ryman, N. (2010). Temporally stable genetic structure of heavily exploited Atlantic herring (Clupea harengus) in Swedish waters. *Heredity*, *104*(1), 40–51. https://doi.org/10.1038/hdy.2009.98

LeBlanc, C. H., Poirier, A. G., MacDougall, C., Bourque, C., & Roy, J. (2008). Assessment of the NAFO Division 4T southern Gulf of St. Lawrence herring stocks in 2007. In *Canadian Science Advisory Secretariat Research Document*.

Leblanc, C., Swain, D., MacDougall, C., & Bourque, C. (2010). Assessment of the NAFO Division 4T southern Gulf of St. Lawrence herring stocks in 2009. *DFO Canadian Science Advisory Secretariat*, (Research Document 2010/059), 143 p.

Lee, H., Gurtowski, J., Yoo, S., Nattestad, M., Marcus, S., Goodwin, S., … Schatz, M. (2016). Third-generation sequencing and the future of genomics. *BioRxiv*, (Table 1), 048603. https://doi.org/doi.org/10.1101/048603

Lee, S., Abecasis, G. R., Boehnke, M., & Lin, X. (2014). Rare-variant association analysis: Study designs and statistical tests. *American Journal of Human Genetics*, *95*(1), 5–23. https://doi.org/10.1016/j.ajhg.2014.06.009

Lehnert, S. J., DiBacco, C., Van Wyngaarden, M., Jeffery, N. W., Ben Lowen, J., Sylvester, E. V. A., … Bradbury, I. R. (2018). Fine-scale temperature-associated genetic structure between inshore and offshore populations of sea scallop (Placopecten magellanicus). *Heredity*, 1–12. https://doi.org/10.1038/s41437-018-0087-9

Lewontin, R. C. (2002). Directions in Evolutionary Biology. *Annual Review of Genetics*, *36*(1), 1–18. https://doi.org/10.1146/annurev.genet.36.052902.102704

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv Preprint ArXiv*, *00*(00), 3. https://doi.org/arXiv:1303.3997 [q-bio.GN]

Li, H., & Durbin, R. (2009a). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., & Durbin, R. (2009b). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, *26*(5), 589–595. https://doi.org/10.1093/bioinformatics/btp698

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, H., Ruan, J., & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, *18*(11), 1851–1858. https://doi.org/10.1101/gr.078212.108

Li, H., & Wren, J. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, *30*(20), 2843–2851. https://doi.org/10.1093/bioinformatics/btu356

Li, R., Li, Y., Fang, X., Yang, H., Wang, J., Kristiansen, K., & Wang, J. (2009). SNP detection for massively parallel whole-genome resequencing. *Genome Research*, *19*(6), 1124–1132. https://doi.org/10.1101/gr.088013.108

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., … Wang, J. (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, *20*(2), 265–272. https://doi.org/10.1101/gr.097261.109

Liaw, A, & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22.

Liaw, Andy, & Wiener, M. (2018). Breiman and Cutler's Random Forests for Classification and Regression.

Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Matthew, P., Leong, J. S., … Vik, J. O. (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature*, *533*(6020), 200–205. https://doi.org/10.1038/nature17164

Limborg, M. T., Helyar, S., De Bruyn, M., Taylor, M. I., Nielsen, E. E., Ogden, R., … Bekkevold, D. (2012). Environmental selection on transcriptome-derived SNPs in a high gene flow marine fish, the Atlantic herring (Clupea harengus). *Molecular Ecology*, *21*(15), 3686–3703. https://doi.org/10.1111/j.1365-294X.2012.05639.x

Lischer, H. E. L., & Excoffier, L. (2012). PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, *28*(2), 298–299. https://doi.org/10.1093/bioinformatics/btr642

Lopes, R. J., Johnson, J. D., Toomey, M. B., Ferreira, M. S., Araujo, P. M., Melo-Ferreira, J., … Carneiro, M. (2016). Genetic Basis for Red Coloration in Birds. *Current Biology*, *26*(11), 1427–1434. https://doi.org/10.1016/j.cub.2016.03.076

Lotterhos, K. E., & Whitlock, M. C. (2015). The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology*, *24*(5), 1031–1046. https://doi.org/10.1111/mec.13100

Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., & Storfer, A. (2017a). Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Molecular Ecology Resources*, *17*(2), 142–152. https://doi.org/10.1111/1755-0998.12635

Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., & Storfer, A. (2017b). Responsible RAD: Striving for best practices in population genomic studies of adaptation. *Molecular Ecology Resources*, *38*(1), 42–49. https://doi.org/10.1111/1755-

0998.12677

Lunter, G., & Goodson, M. (2011). Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, *21*(6), 936–939. https://doi.org/10.1101/gr.111120.110

Luo, Y., Widmer, A., & Karrenberg, S. (2015). The roles of genetic drift and natural selection in quantitative trait divergence along an altitudinal gradient in Arabidopsis thaliana. *Heredity*, *114*, 220–228.

Luu, K., Bazin, E., & Blum, M. G. B. (2017). pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources*, *17*(1), 67–77. https://doi.org/10.1111/1755-0998.12592

Mace, G. M. (2004). The role of taxonomy in species conservation. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *359*(1444), 711–719. https://doi.org/10.1098/rstb.2003.1454

Madec, G., Delecluse, P., Imbard, M., & Levy, C. (1998). *OPA8.1 Ocean general Circulation Model reference manual*. France.

Malmstrøm, M., Matschiner, M., Tørresen, O. K., Jakobsen, K. S., & Jentoft, S. (2017). Whole genome sequencing data and de novo draft assemblies for 66 teleost species. *Scientific Data*, *4*, 160132. https://doi.org/10.1038/sdata.2016.132

Manthey, J. D., Campillo, L. C., Burns, K. J., & Moyle, R. G. (2016). Comparison of Target-Capture and Restriction-Site Associated DNA Sequencing for Phylogenomics: A Test in Cardinalid Tanagers (Aves, Genus: Piranga). *Systematic Biology*, *65*(4), 640–650. https://doi.org/10.1093/sysbio/syw005

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10–12. https://doi.org/http://dx.doi.org/10.14806/ej.17.1.200

Martin, S. H., & Jiggins, C. D. (2013). Genomic Studies of Adaptation in Natural Populations. In *eLS*. https://doi.org/10.1002/9780470015902.a0024613

Martinez Barrio, A., Lamichhaney, S., Fan, G., Rafati, N., Pettersson, M., Zhang, H., … Andersson, L. (2016). The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. *ELife*, *5*(MAY2016), 1–32. https://doi.org/10.7554/eLife.12081

Martinsohn, J. T., & Ogden, R. (2009). FishPopTrace—Developing SNP-based population genetic assignment methods to investigate illegal fishing. *Forensic Science International: Genetics Supplement Series*, *2*(1), 294–296. https://doi.org/10.1016/j.fsigss.2009.08.108

McDermid, J. L., Swain, D. P., Turcotte, F., Robichaud, S. A., & Surette, T. (2018). Assessment of the NAFO Division 4T southern Gulf of St. Lawrence Atlantic herring (Clupea harengus) in 2016 and 2017. *DFO Can. Sci. Advis. Sec. Res. Doc. 2018/052*, xiv + 122 p.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

McKinney, G. J., Larson, W. A., Seeb, L. W., & Seeb, J. E. (2017). RADseq provides unprecedented insights into molecular ecology and evolutionary genetics: comment on Breaking RAD by Lowry et al . (2016). *Molecular Ecology Resources*, *17*(3), 356–361. https://doi.org/10.1111/1755-0998.12649

McMahon, B. J., Teeling, E. C., & Höglund, J. (2014). How and why should we implement genomics into conservation? *Evolutionary Applications*, *7*(9), 999–1007. https://doi.org/10.1111/eva.12193

McPherson, A., O'Reilly, P. T., & Taggart, C. T. (2004). Genetic Differentiation, Temporal Stability, and the Absence of Isolation by Distance among Atlantic Herring Populations. *Transactions of the American Fisheries Society*, *133*(2), 434–446. https://doi.org/10.1577/02-106

McPherson, A., Stephenson, R. L., O'Reilly, P. T., Jones, M. W., & Taggart, C. T. (2001). Genetic diversity of coastal Northwest Atlantic herring populations: implications for management. *Journal of Fish Biology*, *59*(SUPPL. A), 356–370. https://doi.org/10.1006/jfbi.2001.1769

McQuinn, I. H. (1987). New maturity cycle charts for herring stocks along the west coast of Newfoundland (NAFO Division 4R) and the north shore of Quebec (NAFO Division 4S). In *Canadian Atlantic Fisheries Scientific Advisory Committee Research Document*.

McQuinn, Ian H. (1997). Metapopulations and the Atlantic herring. *Reviews in Fish Biology and Fisheries*, *7*(3), 297–329. https://doi.org/10.1023/A:1018491828875

Meirmans, P. G., & Van Tienderen, P. H. (2004). genotype and genodive: two programs for the analysis of genetic diversity of asexual organisms. *Molecular Ecology Notes*, *4*(4), 792–794. https://doi.org/10.1111/j.1471-8286.2004.00770.x

Melamed, P., Savulescu, D., Lim, S., Wijeweera, A., Luo, Z., Luo, M., & Pnueli, L. (2012). Gonadotrophin-Releasing Hormone signalling downstream of Calmodulin. *Journal of Neuroendocrinology*, *24*(12), 1463–1475. https://doi.org/10.1111/j.1365-2826.2012.02359.x

Melton, C., Reuter, J. A., Spacek, D. V, & Snyder, M. (2015). Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nature Genetics*, *47*(7), 710–716. https://doi.org/10.1038/ng.3332

Melvin, G. D., Stephenson, R. L., & Power, M. J. (2009). Oscillating reproductive strategies of herring in the western Atlantic in response to changing environmental conditions. *ICES Journal of Marine Science*, *66*(8), 1784–1792. https://doi.org/10.1093/icesjms/fsp173

Messer, P. W., & Petrov, D. A. (2013). Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology and Evolution*, *28*(11), 659–669. https://doi.org/10.1016/j.tree.2013.08.003

Messieh, S. N. (1975). Maturation and spawning of Atlantic herring (Clupea harengus harengus) in the southern Gulf of St Lawrence. *Journal of the Fisheries Research Board of Canada*, *32*, 66–68.

Messieh, S. N. (1988). Spawning of Atlantic Herring in the Gulf of St. Lawrence. *American Fisheries Society Symposium*, *5*, 31–48.

Messieh, S. N., Anthony, V., & Sinclair, M. (1985). Fecundities of Atlantic herring Clupea harengus L. populations in the Northwest Atlantic. *ICES C.M. 1985/H:8.*, 22 pp.

Metzger, B. P. H., Duveau, F., Yuan, D. C., Tryban, S., Yang, B., & Wittkopp, P. J. (2016). Contrasting Frequencies and Effects of cis - and trans -Regulatory Mutations Affecting Gene Expression. *Molecular Biology and Evolution*, *33*(5), 1131–1146. https://doi.org/10.1093/molbev/msw011

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C. -C., & Lin, C. -C. (2015). e1071: Misc Functions of the Department of Statistics, Probability Theory Group

(Formerly: E1071).

Miller, M. R., Brunelli, J. P., Wheeler, P. A., Liu, S., Rexroad, C. E., Palti, Y., … Thorgaard, G. H. (2012). A conserved haplotype controls parallel adaptation in geographically distant salmonid populations. *Molecular Ecology*, *21*(2), 237–249. https://doi.org/10.1111/j.1365-294X.2011.05305.x

Moran, B. M., & Anderson, E. C. (2018). Bayesian inference from the conditional genetic stock identification model. *Canadian Journal of Fisheries and Aquatic Sciences*, 1–10. https://doi.org/10.1139/cjfas-2018-0016

Moutsianas, L., Agarwala, V., Fuchsberger, C., Flannick, J., Rivas, M. A., Gaulton, K. J., … McCarthy, M. I. (2015). The Power of Gene-Based Rare Variant Methods to Detect Disease-Associated Variation and Test Hypotheses About Complex Disease. *PLoS Genetics*, *11*(4), 1–24. https://doi.org/10.1371/journal.pgen.1005165

Mukherjee, S., Stamatis, D., Bertsch, J., Ovchinnikova, G., Verezemska, O., Isbandi, M., … Reddy, T. B. K. (2017). Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. *Nucleic Acids Research*, *45*(D1), D446–D456. https://doi.org/10.1093/nar/gkw992

Muñoz, I., Henriques, D., Johnston, J. S., Ch?vez-Galarza, J., Kryger, P., & Pinto, M. A. (2015). Reduced SNP Panels for Genetic Identification and Introgression Analysis in the Dark Honey Bee (Apis mellifera mellifera). *PLOS ONE*, *10*(4), e0124365. https://doi.org/10.1371/journal.pone.0124365

Myburg, A. A., Grattapaglia, D., Tuskan, G. A., Hellsten, U., Hayes, R. D., Grimwood, J., … Schmutz, J. (2014). The genome of Eucalyptus grandis. *Nature*, *510*(7505), 356–362. https://doi.org/10.1038/nature13308

Nadachowska-Brzyska, K., Burri, R., Smeds, L., & Ellegren, H. (2016). PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white Ficedula flycatchers. *Molecular Ecology*, *25*(5), 1058–1072. https://doi.org/10.1111/mec.13540

Nagasaki, M., Yasuda, J., Katsuoka, F., Nariai, N., Kojima, K., Kawai, Y., … Yamamoto, M. (2015). Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nature Communications*, *6*, 8018. https://doi.org/10.1038/ncomms9018

Nakao, N., Ono, H., Yamamura, T., Anraku, T., Takagi, T., Higashi, K., … Yoshimura, T. (2008). Thyrotrophin in the pars tuberalis triggers photoperiodic response. *Nature*, *452*(7185), 317–322.

Nayfa, M. G., & Zenger, K. R. (2016). Unravelling the effects of gene flow and selection in highly connected populations of the silver-lip pearl oyster (Pinctada maxima). *Marine Genomics*, *28*, 99–106. https://doi.org/10.1016/j.margen.2016.02.005

Neale, D. B., Wegrzyn, J. L., Stevens, K. a, Zimin, A. V, Puiu, D., Crepeau, M. W., … Langley, C. H. (2014). Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biology*, *15*(3), R59. https://doi.org/10.1186/gb-2014-15-3-r59

Nei, M. (2007). The new mutation theory of phenotypic evolution. *Proceedings of the National Academy of Sciences*, *104*(30), 12235–12242. https://doi.org/10.1073/pnas.0703349104

Nei, M. (1987). *Molecular Evolutionary Genetics*. New York: Columbia Univ Press.

Nei, Masatoshi. (1972). Genetic distance between populations. *The American Naturalist*, *1062*(949), 283–292. https://doi.org/10.1086/285153

Nevado, B., Ramos-Onsins, S. E., & Perez-Enciso, M. (2014). Resequencing studies of nonmodel organisms using closely related reference genomes: Optimal experimental designs and bioinformatics approaches for population genomics. *Molecular Ecology*, *23*(7), 1764–1779. https://doi.org/10.1111/mec.12693

Nielsen, R. (2009). Adaptionism - 30 years after gould and lewontin. *Evolution*, *63*(10), 2487–2490. https://doi.org/10.1111/j.1558-5646.2009.00799.x

Nielsen, R., Akey, J. M., Jakobsson, M., Pritchard, J. K., Tishkoff, S., & Willerslev, E. (2017). Tracing the peopling of the world through genomics. *Nature*, *541*(7637), 302–310. https://doi.org/10.1038/nature21347

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., & Wang, J. (2012). SNP Calling, Genotype Calling, and Sample Allele Frequency Estimation from New-Generation Sequencing Data. *PLoS ONE*, *7*(7), e37558. https://doi.org/10.1371/journal.pone.0037558

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews. Genetics*, *12*(6), 443–451. https://doi.org/10.1038/nrg2986

Norman, A. J., Street, N. R., & Spong, G. (2013). De Novo SNP Discovery in the Scandinavian Brown Bear (Ursus arctos). *PLoS ONE*, *8*(11), e81012. https://doi.org/10.1371/journal.pone.0081012

Nosil, P., & Feder, J. L. (2012). Genomic divergence during speciation: causes and consequences. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1587), 332–342. https://doi.org/10.1098/rstb.2011.0263

Nosil, P., Funk, D. J., & Ortiz-Barrientos, D. (2009). Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, *18*(3), 375–402. https://doi.org/10.1111/j.1365-294X.2008.03946.x

O'Connor Lab. (2016). ScaffoldStitcher. Retrieved May 1, 2018, from https://bitbucket.org/dholab/scaffoldstitcher/src

O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., … Lyon, G. J. (2013). Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Medicine*, *5*(3), 28. https://doi.org/10.1186/gm432

Okonechnikov, K., Conesa, A., & García-Alcalde, F. (2015). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, *32*(2), btv566. https://doi.org/10.1093/bioinformatics/btv566

Ono, H., Hoshino, Y., Yasuo, S., Watanabe, M., Nakane, Y., Murai, A., … Yoshimura, T. (2008). Involvement of thyrotropin in photoperiodic signal transduction in mice. *Proceedings of the National Academy of Sciences, USA*, *105*(47), 18238–18242. https://doi.org/10.1073/pnas.0808952105

Ouborg, N. J., Pertoldi, C., Loeschcke, V., Bijlsma, R. K., & Hedrick, P. W. (2010). Conservation genetics in transition to conservation genomics. *Trends in Genetics*, *26*(4), 177–187. https://doi.org/10.1016/j.tig.2010.01.001

Overholtz, W. . (2002). The Gulf of Maine–Georges Bank Atlantic herring (Clupea harengus): spatial pattern analysis of the collapse and recovery of a large marine fish complex. *Fisheries Research*, *57*(3), 237–254. https://doi.org/10.1016/S0165-7836(01)00359-9

Ozsolak, F., & Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, *12*(2), 87–98. https://doi.org/10.1038/nrg2934

Pagani, F., & Baralle, F. E. (2004). Genomic variants in exons and introns: identifying the splicing spoilers. *Nature Reviews Genetics*, *5*(May), 389–396.

Palumbi, S. R. (1994). Genetic Divergence, Reproductive Isolation, and Marine Speciation. *Annual Review of Ecology and Systematics*, *25*(1), 547–572. https://doi.org/10.1146/annurev.es.25.110194.002555

Panagiotou, O. A., & Ioannidis, J. P. A. (2012). What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *International Journal of Epidemiology*, *41*(1), 273–286. https://doi.org/10.1093/ije/dyr178

Pardo-Diaz, C., Salazar, C., & Jiggins, C. D. (2015). Towards the identification of the loci of adaptive evolution. *Methods in Ecology and Evolution*, *6*(4), 445–464. https://doi.org/10.1111/2041-210X.12324

Parejo, M., Wragg, D., Gauthier, L., Vignal, A., Neumann, P., & Neuditschko, M. (2016). Using Whole-Genome Sequence Information to Foster Conservation Efforts for the European Dark Honey Bee, Apis mellifera mellifera. *Frontiers in Ecology and Evolution*, *4*(December), 1–15. https://doi.org/10.3389/fevo.2016.00140

Pasaniuc, B., Rohland, N., McLaren, P. J., Garimella, K., Zaitlen, N., Li, H., … Price, A. L. (2012). Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nature Genetics*, *44*(6), 631–635. https://doi.org/10.1038/ng.2283

Paten, B., Novak, A. M., Eizenga, J. M., & Garrison, E. (2017). Genome graphs and the evolution of genome inference. *Genome Research*, *27*(5), 665–676. https://doi.org/10.1101/gr.214155.116

Pearse, D E, & Pogson, G. H. (2000). Parallel evolution of the melanic form of the California legless lizard, Anniella pulchra, inferred from mitochondrial DNA sequence variation. *Evolution; International Journal of Organic Evolution*, *54*(3), 1041–1046.

Pearse, Devon E, Miller, M. R., Abadia-Cardoso, A., & Garza, J. C. (2014). Rapid parallel evolution of standing variation in a single, complex, genomic region is associated with life history in steelhead/rainbow trout. *Proceedings. Biological Sciences / The Royal Society*, *281*(1783), 20140012. https://doi.org/10.1098/rspb.2014.0012

Pedersen, B. S., Layer, R. M., Quinlan, A. R., Li, H., Wang, K., Li, M., … Kang, H. (2016). Vcfanno: fast, flexible annotation of genetic variants. *Genome Biology*, *17*(1), 118. https://doi.org/10.1186/s13059-016-0973-5

Pettersson, E., Lundeberg, J., & Ahmadian, A. (2009). Generations of sequencing technologies. *Genomics*, *93*(2), 105–111. https://doi.org/10.1016/j.ygeno.2008.10.003

Pfeifer, S. P. (2017). From next-generation resequencing reads to a high-quality variant data set. *Heredity*, *118*(2), 111–124. https://doi.org/10.1038/hdy.2016.102

Phan, V., Gao, S., Tran, Q., & Vo, N. S. (2014). How genome complexity can explain the hardness of aligning reads to genomes. *2014 IEEE 4th International Conference on Computational Advances in Bio and Medical Sciences, ICCABS 2014*, *16*(Suppl 17), 1–15. https://doi.org/10.1109/ICCABS.2014.6863916

Phillippy, A. M. (2017). New advances in sequence assembly. *Genome Research*, *27*(5), xi–xiii. https://doi.org/10.1101/gr.223057.117

Pompanon, F., Bonin, A., Bellemain, E., & Taberlet, P. (2005). Genotyping errors: causes, consequences and solutions. *Nature Reviews Genetics*, *6*(11), 847–859. https://doi.org/10.1038/nrg1707

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Auwera, G. A. Van der, … Banks, E. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv*, 1–22. https://doi.org/https://doi.org/10.1101/201178

Power, M. J., Clark, K. J., Fife, J. F., Knox, D., Melvin, G. D., & Stephenson, R. L. (2007). 2007 evaluation of 4VWX herring. In *Canadian Science Advisory Secretariat Research Document, 2007/040*.

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE*, *5*(3), e9490.

Primmer, C. R. (2009). From conservation genetics to conservation genomics. *Annals of the New York Academy of Sciences*, *1162*, 357–368. https://doi.org/10.1111/j.1749-6632.2009.04444.x

Pritchard, J. K., & Di Rienzo, A. (2010). Adaptation - not by sweeps alone. *Nature Reviews. Genetics*, *11*(10), 665–667. https://doi.org/10.1038/nrg2880

Pritchard, J., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., … Carroll, M. W. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature*, *530*(7589), 228–232. https://doi.org/10.1038/nature16996

R Core Development Team. (2019). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Rabbani, B., Tekin, M., & Mahdieh, N. (2014). The promise of whole-exome sequencing in medical genetics. *Journal of Human Genetics*, *59*(1), 5–15. https://doi.org/10.1038/jhg.2013.114

Rafati, N., Andersson, L. S., Mikko, S., Feng, C., Pettersson, J., Janecka, J., … Evan, E. (2016). Large Deletions at the SHOX Locus in the Pseudoautosomal Region are associated with Skeletal Atavism in Shetland ponies. *Genes|Genomes|Genetics*, *6*(July), 2213–2223. https://doi.org/10.1534/g3.116.029645

Raineri, E., Ferretti, L., Esteve-Codina, A., Nevado, B., Heath, S., & Pérez-Enciso, M. (2012). SNP calling by sequencing pooled samples. *BMC Bioinformatics*, *13*(1), 239. https://doi.org/10.1186/1471-2105-13-239

Rambaut, A. (2007). FigTree. Retrieved May 20, 2018, from http://tree.bio.ed.ac.uk/software/figtree/

Ravinet, M., Faria, R., Butlin, R. K., Galindo, J., Bierne, N., Rafajlović, M., … Westram, A. M. (2017). Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. *Journal of Evolutionary Biology*, *30*(8), 1450–1477. https://doi.org/10.1111/jeb.13047

Reid, R. N., Cargnelli, L. M., Griesbach, S. J., Packer, D. B., Johnson, D. L., Zetlin, C., … Berrien, P. L. (1999). Atlantic Herring, Clupea harengus, Life History and Habitat Characteristics, Available online at http://www.nefsc.noaa.gov/publications/tm/tm126/tm126.pdf.

Reinert, K., Langmead, B., Weese, D., & Evers, D. J. (2015). Alignment of Next-Generation Sequencing Reads. *Annual Review of Genomics and Human Genetics*, *16*, 133–151. https://doi.org/10.1146/annurev-genom-090413-025358

Reiss, H., Hoarau, G., Dickey-Collas, M., & Wolff, W. J. (2009). Genetic population structure of marine fish: mismatch between biological and fisheries management units. *Fish and Fisheries*, *10*(4), 361–395. https://doi.org/10.1111/j.1467-2979.2008.00324.x

Rellstab, C., Fischer, M. C., Zoller, S., Graf, R., Tedder, A., Shimizu, K. K., … Gugerli, F. (2016). Local adaptation (mostly) remains local: reassessing environmental associations of climate-related candidate SNPs in Arabidopsis halleri. *Heredity*, *118*(July), 1–9. https://doi.org/10.1038/hdy.2016.82

Revelle, W. (2018). *psych: Procedures for Personality and Psychological Research*. Retrieved from https://cran.r-project.org/package=psych

Richards, C. L., Bossdorf, O., & Pigliucci, M. (2010). What Role Does Heritable Epigenetic Variation Play in Phenotypic Evolution? *BioScience*, *60*(3), 232–237. https://doi.org/10.1525/bio.2010.60.3.9

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E., Getz, G., & Mesirov, J. P. (2011). Integrative Genomics Viewer. *Nature Biotechnology*, *29*(1), 24–26. https://doi.org/10.1038/nbt0111-24

Rockman, M. V. (2012). The QTN program and the alleles that matter for evolution: All that's gold does not glitter. *Evolution*, *66*(1), 1–17. https://doi.org/10.1111/j.1558-5646.2011.01486.x

Ronen, R., Udpa, N., Halperin, E., & Bafna, V. (2013). Learning natural selection from the site frequency spectrum. *Genetics*, *195*(1), 181–193. https://doi.org/10.1534/genetics.113.152587

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., … Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data. *Genome Biology*, *14*(5), R51. https://doi.org/10.1186/gb-2013-14-5-r51

Rousset, F. (1997). Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics*, *145*(April), 1219–1228.

Rubin, C.-J., Zody, M. C., Eriksson, J., Meadows, J. R. S., Sherwood, E., Webster, M. T., … Andersson, L. (2010). Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature*, *464*(7288), 587–591. https://doi.org/10.1038/nature08832

Ruffalo, M., Koyutürk, M., Ray, S., & LaFramboise, T. (2012). Accurate estimation of short read mapping quality for next-generation genome sequencing. *Bioinformatics*, *28*(18), 349–355. https://doi.org/10.1093/bioinformatics/bts408

Ruzzante, D. E., Mariani, S., Bekkevold, D., André, C., Mosegaard, H., Clausen, L. A. W., … Carvalho, G. R. (2006). Biocomplexity in a highly migratory pelagic marine fish, Atlantic herring. *Proceedings of the Royal Society B: Biological Sciences*, *273*(1593), 1459–1464. https://doi.org/10.1098/rspb.2005.3463

Ruzzante, D. E., Taggart, C. T., Lang, S., Cook, D., Applications, E., & Aug, N. (2000). Mixed-stock analysis of Atlantic cod near the Gulf of St. Lawrence based on microsatellite DNA. *Ecological Applications*, *10*(4), 1090–1109.

Ryman, N., & Palm, S. (2006). POWSIM: a computer program for assessing statistical power when testing for genetic differentiation. *Molecular Ecology Notes*, *6*(3), 600–602.

https://doi.org/10.1111/j.1471-8286.2006.01378.x

Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., … Yorke, J. a. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research*, *22*(3), 557–567. https://doi.org/10.1101/gr.131383.111

Sambrook, J. & Russel D.W. (2006) Purification of Nucleic Acids by Extraction with Phenol:Chloroform. *Cold Spring Harb Protoc. doi:10.1101/pdb.prot4455*

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, *74*(12), 5463–5467. https://doi.org/10.1073/pnas.74.12.5463

Satterthwaite, W. H., & Carlson, S. M. (2015). Weakening portfolio effect strength in a hatchery-supplemented Chinook salmon population complex. *Canadian Journal of Fisheries and Aquatic Sciences*, *72*(12), 1860–1875. https://doi.org/10.1139/cjfas-2015-0169

Savolainen, O., Lascoux, M., & Merilä, J. (2013). Ecological genomics of local adaptation. *Nature Reviews. Genetics*, *14*(11), 807–820. https://doi.org/10.1038/nrg3522

Schiffels, S., & Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, *46*(8), 919–925. https://doi.org/10.1038/ng.3015

Schindler, D. E., Hilborn, R., Chasco, B., Boatright, C. P., Quinn, T. P., Rogers, L. A., & Webster, M. S. (2010). Population diversity and the portfolio effect in an exploited species. *Nature*, *465*(7298), 609–612. https://doi.org/10.1038/nature09060

Schlötterer, C., Tobler, R., Kofler, R., & Nolte, V. (2014). Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. *Nature Reviews Genetics*, *15*(11), 749–763. https://doi.org/10.1038/nrg3803

Schluter, D. (2009). Evidence for Ecological Speciation and Its Alternative. *Science*, *323*(5915), 737–741. https://doi.org/10.1126/science.1160006

Schrider, D. R., & Kern, A. D. (2018). Supervised Machine Learning for Population Genetics: A New Paradigm. *Trends in Genetics*, *34*(4), 301–312. https://doi.org/10.1016/j.tig.2017.12.005

Scott, W. B., & Scott, M. G. (1988). *Atlantic fishes of Canada. Canadian Bulletin of Fisheries and Aquatic Sciences, bulletin 219*. Toronto, CA: University of Toronto Press.

Sedlackova, T., Repiska, G., Celec, P., Szemes, T., & Minarik, G. (2013). Fragmentation of DNA affects the accuracy of the DNA quantitation by the commonly used methods. *Biological Procedures Online*, *15*(1), 5. https://doi.org/10.1186/1480-9222-15-5

Shafer, A. B. A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., & Wolf, J. B. W. (2016). Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods in Ecology and Evolution*, 1–11. https://doi.org/10.1111/2041-210X.12700

Shafer, A. B. A., Wolf, J. B. W., Alves, P. C., Bergström, L., Bruford, M. W., Brännström, I., … Zieliński, P. (2015). Genomics and the challenging translation into conservation practice. *Trends in Ecology and Evolution*, *30*(2), 78–87. https://doi.org/10.1016/j.tree.2014.11.009

Shapiro, M. D., Marks, M. E., Peichel, C. L., Blackman, B. K., Nereng, K. S., Jónsson, B., … Kingsley, D. M. (2006). Corrigendum: Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature*, *439*(7079), 1014–1014. https://doi.org/10.1038/nature04500

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, *26*(10), 1135–1145. https://doi.org/10.1038/nbt1486

Simmonds, E. J. (2007). Comparison of two periods of North Sea herring stock management: success, failure, and monetary value. *ICES Journal of Marine Science*, *64*(4), 686–692. https://doi.org/10.1093/icesjms/fsm045

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews. Genetics*, *15*(2), 121–132. https://doi.org/10.1038/nrg3642

Sinclair-Waters, M. (2017). *Genomic perspectives for conservation and management of Atlantic cod in costal Labrador. (Unpublished master's thesis).* Dalhousie University, Halifax, Canada.

Sinclair, M. (1988). *Marine Populations: an Essay on Population Regulation and Speciation*. Seattle: Washington Sea Grant/Univ. Wash. Press.

Sinclair, M., & Iles, T. D. (1989). Population regulation and speciation in the oceans. *J. Cons. Int. Explor. Mer*, *45*, 165–175.

Sinclair, M., & Tremblay, M. J. (1984a). Timing of Spawning of Atlantic Herring (Clupea harengus harengus) Populations and the Match–Mismatch Theory. *Canadian Journal of Fisheries and Aquatic Sciences*, *41*(7), 1055–1065. https://doi.org/10.1139/f84-123

Sinclair, M., & Tremblay, M. J. (1984b). Timing of Spawning of Atlantic Herring (Clupea harengus harengus) Populations and the Match–Mismatch Theory. *Canadian Journal of Fisheries and Aquatic Sciences*, *41*(7), 1055–1065. https://doi.org/10.1139/f84-123

Skotte, L., Korneliussen, T. S., & Albrechtsen, A. (2013). Estimating individual admixture proportions from next generation sequencing data. *Genetics*, *195*(3), 693–702. https://doi.org/10.1534/genetics.113.154138

Slatkin, M. (1987). Gene flow and the geographic structure of natural populations. *Science*, *236*(4803), 787–792. https://doi.org/10.1126/science.3576198

Smith, P. J., & Jamieson, A. (1986). Stock discreteness in herrings: a conceptual revolution. *Fish. Res.*, 223–234.

Snyder-Mackler, N., Majoros, W. H., Yuan, M. L., Shaver, A. O., Gordon, J. B., Kopp, G. H., … Tung, J. (2016). Efficient genome-wide sequencing and low-coverage pedigree analysis from noninvasively collected samples. *Genetics*, *203*(2), 699–714. https://doi.org/10.1534/genetics.116.187492

Snyder, M. W., Adey, A., Kitzman, J. O., & Shendure, J. (2015). Haplotype-resolved genome sequencing: experimental methods and applications. *Nature Reviews Genetics*, *16*(6), 344–358. https://doi.org/10.1038/nrg3903

Stahl, G. (1983). Differences in the amount and distribution of genetic variation between natural populations and hatchery stocks of Atlantic salmon. *Aquaculture*, *33*(1–4), 23–32. https://doi.org/http://dx.doi.org/10.1016/0044-8486(83)90383-6

Stanley, R. R. E., DiBacco, C., Lowen, B., Beiko, R. G., Jeffery, N. W., Van Wyngaarden, M., … Bradbury, I. R. (2018). A climate-associated multispecies cryptic cline in the northwest Atlantic. *Science Advances*, *4*(3), eaaq0929. https://doi.org/10.1126/sciadv.aaq0929

Stanley, R. R. E., & Jeffery, N. W. (2017). CartDist: Re-projection tool for complex marine systems. https://doi.org/10.5281/zenodo.802875

Stanley, R. R. E., Jeffery, N. W., Wringe, B. F., DiBacco, C., & Bradbury, I. R. (2017). <scp>genepopedit</scp> : a simple and flexible tool for manipulating multilocus molecular data in R. *Molecular Ecology Resources*, *17*(1), 12–18. https://doi.org/10.1111/1755-0998.12569

Steiner, C. C., Putnam, A. S., Hoeck, P. E. A., & Ryder, O. A. (2013). Conservation genomics of threatened animal species. *Annual Review of Animal Biosciences*, *1*, 261–281. https://doi.org/10.1146/annurev-animal-031412-103636

Stephenson, R. L., Melvin, G. D., & Power, M. J. (2009). Population integrity and connectivity in Northwest Atlantic herring: a review of assumptions and evidence. *ICES Journal of Marine Science*, *66*(8), 1733–1739. https://doi.org/10.1093/icesjms/fsp189

Stern, D. L. (2013). The genetic causes of convergent evolution. *Nat Rev Genet*, *14*(11), 751–764.

Stetz, J. B., smith, S., Sawaya, M. A., Ramsey, A. B., Amish, S. J., Schwartz, M. K., & Luikart, G. (2016). Discovery of 20,000 RAD–SNPs and development of a 52-SNP array for monitoring river otters. *Conservation Genetics Resources*, *8*(3), 299–302. https://doi.org/10.1007/s12686-016-0558-3

Stobo, W. T. (1987). Atlantic herring (Clupea harengus) movement along the Scotian Shelf and management considerations. *Proceedings of the Conference on Forage Fishes of the Southeastern Bering Sea, Anchorage, Alaska, 4–5 November 1986, Pp. 75–85. US Department of the Interior, Minerals Management Services, Alaska OCS Region, MMS Report, 87-0017. 122 Pp.*

Straub, S. C. K., Fishbein, M., Livshultz, T., Foster, Z., Parks, M., Weitemier, K., … Liston, A. (2011). Building a model: developing genomic resources for common milkweed (Asclepias syriaca) with low coverage genome sequencing. *BMC Genomics*, *12*(1), 211. https://doi.org/10.1186/1471-2164-12-211

Sylvester, E. V. A., Beiko, R. G., Bentzen, P., Paterson, I., Horne, J. B., Watson, B., … Bradbury, I. R. (2018). Environmental extremes drive population structure at the northern range limit of Atlantic salmon in North America. *Molecular Ecology*, *27*(20), 4026–4040. https://doi.org/10.1111/mec.14849

Sylvester, E. V. A., Bentzen, P., Bradbury, I. R., Clément, M., Pearce, J., Horne, J., & Beiko, R. G. (2017). Applications of random forest feature selection for fine-scale genetic population assignment. *Evolutionary Applications*, (September 2016), 1–13. https://doi.org/10.1111/eva.12524

Tavares, H., Whibley, A., Field, D. L., Bradley, D., Couchman, M., Copsey, L., … Coen, E. (2018). Selection and gene flow shape genomic islands that control floral guides. *Proceedings of the National Academy of Sciences*, *115*(43), 11006–11011. https://doi.org/10.1073/pnas.1801832115

Teacher, A. G., André, C., Jonsson, P. R., & Merilä, J. (2013). Oceanographic connectivity and environmental correlates of genetic structuring in Atlantic herring in the Baltic Sea. *Evolutionary Applications*, *6*(3), 549–567. https://doi.org/10.1111/eva.12042

Teacher, A. G., André, C., Merilä, J., & Wheat, C. W. (2012). Whole mitochondrial genome scan for population structure and selection in the Atlantic herring. *BMC Evolutionary Biology*, *12*, 248. https://doi.org/10.1186/1471-2148-12-248

Temple, G. K., Cole, N. J., & Johnston, I. A. (2001). Embryonic temperature and the relative timing of muscle-specific genes during development in herring (Clupea harengus L.). *Journal of Experimental Biology*, *204*(21), 3629–3637.

The Computational Pan-genomics Consortium. (2016). Computational pan-genomics: status, promises and challenges. *Briefings in Bioinformatics*, (August), 1–18. https://doi.org/10.1093/bib/bbw089

The FAASG Consortium. (2016). *Functional Analysis of All Salmonid Genomes (FAASG): an international initiative supporting future salmonid research, conservation and aquaculture*. 1–18. https://doi.org/http://dx.doi.org/10.1101/095737

Therkildsen, N. O., & Palumbi, S. R. (2017). Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. *Molecular Ecology Resources*, *17*(2), 194–208. https://doi.org/10.1111/1755-0998.12593

Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, *14*(2), 178–192. https://doi.org/10.1093/bib/bbs017

Tibbo, S. N., Legare, J. E. H., Scatterwood, L. W., & Temple, R. F. (1958). On the occurrence and distribution of larval herring (Clupea harengus L.) in the Bay of Fundy and the Gulf of Maine. *Journal of the Fisheries Research Board of Canada*, *15*, 1451–1469.

Tiffin, P., & Ross-Ibarra, J. (2014). Advances and limits of using population genetics to understand local adaptation. *Trends in Ecology and Evolution*, *29*(12), 673–680. https://doi.org/10.1016/j.tree.2014.10.004

Tigano, A., & Friesen, V. L. (2016). Genomics of local adaptation with gene flow. *Molecular Ecology*, n/a-n/a. https://doi.org/10.1111/mec.13606

Townsend, D. W., Thomas, A. C., Mayer, L. M., Thomas, M. A., & Quinlan, J. A. (2004). Oceanography of the Northwest Atlantic Shelf (1, W). In A. R. Robinson & K. H. Brink (Eds.), *The Sea: The Global Coastal Ocean: Interdisciplinary Regional Studies and Syntheses* (pp. 1–57). Harvard University Press.

Travis, J. M. J., Munkemuller, T., Burton, O. J., Best, A., Dytham, C., & Johst, K. (2007). Deleterious Mutations Can Surf to High Densities on the Wave Front of an Expanding Population. *Molecular Biology and Evolution*, *24*(10), 2334–2343. https://doi.org/10.1093/molbev/msm167

Treangen, T. J., & Salzberg, S. L. (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, *13*(1), 36–46. https://doi.org/10.1038/nrg3117

Trowsdale, J., & Knight, J. C. (2013). Major Histocompatibility Complex Genomics and Human Disease. *Annual Review of Genomics and Human Genetics*, *14*(1), 301–323. https://doi.org/10.1146/annurev-genom-091212-153455

Tung, J., Zhou, X., Alberts, S. C., Stephens, M., & Gilad, Y. (2015). The genetic architecture of gene expression levels in wild baboons. *ELife*, *4*, 1–22. https://doi.org/10.7554/eLife.04729

Turner, S. D. (2014). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *BiorXiv*. https://doi.org/https://doi.org/10.1101/005165

Turner, T. L., Hahn, M. W., & Nuzhdin, S. V. (2005). Genomic Islands of Speciation in Anopheles gambiae. *PLoS Biology*, *3*(9), e285. https://doi.org/10.1371/journal.pbio.0030285

Vähä, J.-P., & Primmer, C. R. (2006). Efficiency of model-based Bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci. *Molecular Ecology*, *15*, 63–72. https://doi.org/10.1111/j.1365-294X.2005.02773.x

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., … DePristo, M. A. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. In *Current Protocols in Bioinformatics* (Vol. 11, pp. 11.10.1-11.10.33). https://doi.org/10.1002/0471250953.bi1110s43

van Overzee, H. M. J., & Rijnsdorp, A. D. (2015). Effects of fishing during the spawning period: implications for sustainable management. *Reviews in Fish Biology and Fisheries*, *25*(1), 65–83. https://doi.org/10.1007/s11160-014-9370-x

Vandergast, A. (2017). *Incorporating genetic sampling in long-term monitoring and adaptive management in the San Diego County Management Strategic Plan Area, Southern California*. https://doi.org/10.3133/ofr20171061

VanderMeer, J. E., & Ahituv, N. (2011). cis-regulatory mutations are a genetic cause of human limb malformations. *Developmental Dynamics*, *240*(5), 920–930. https://doi.org/10.1002/dvdy.22535

Varshney, G. K., Pei, W., LaFave, M. C., Idol, J., Xu, L., Gallardo, V., … Burgess, S. M. (2015). High-throughput gene targeting and phenotyping in zebrafish using CRISPR/Cas9. *Genome Research*, *25*(7), 1030–1042. https://doi.org/10.1101/gr.186379.114

Vatsiou, A. I., Bazin, E., & Gaggiotti, O. E. (2016). Detection of selective sweeps in structured populations: A comparison of recent methods. *Molecular Ecology*, *25*(1), 89–103. https://doi.org/10.1111/mec.13360

Veeckman, E., Ruttink, T., & Vandepoele, K. (2016). Are We There Yet? Reliably Estimating the Completeness of Plant Genome Sequences. *The Plant Cell*, *28*(8), 1759–1768. https://doi.org/10.1105/tpc.16.00349

Velasco, D., Hough, J., Aradhya, M., & Ross-Ibarra, J. (2016). Evolutionary Genomics of Peach and Almond Domestication. *Genes|Genomes|Genetics*, *6*(December), 3985–3993. https://doi.org/10.1534/g3.116.032672

Vieira, Filipe G., Albrechtsen, A., & Nielsen, R. (2016). Estimating IBD tracts from low coverage NGS data. *Bioinformatics*, *32*(14), 2096–2102. https://doi.org/10.1093/bioinformatics/btw212

Vieira, Filipe Garrett, Fumagalli, M., Albrechtsen, A., & Nielsen, R. (2013). Estimating inbreeding coefficients from NGS data: Impact on genotype calling and allele frequency estimation. *Genome Research*, *23*(11), 1852–1861. https://doi.org/10.1101/gr.157388.113

Vitti, J. J., Grossman, S. R., & Sabeti, P. C. (2013). Detecting Natural Selection in Genomic Data. *Annual Review of Genetics*, *47*(1), 97–120. https://doi.org/10.1146/annurev-genet-111212-133526

von der Heyden, S., Beger, M., Toonen, R. J., van Herwerden, L., Juinio-Meñez, M. A., Ravago-Gotanco, R., … Bernardi, G. (2014). The application of genetics to marine management and conservation: examples from the Indo-Pacific. *Bulletin of Marine Science*, *90*(1), 123–158. https://doi.org/10.5343/bms.2012.1079

vonHoldt, B. M., Cahill, J. A., Fan, Z., Gronau, I., Robinson, J., Pollinger, J. P., … Wayne, R. K. (2016). Whole-genome sequence analysis shows that two endemic species of North American wolf are admixtures of the coyote and gray wolf. *Science Advances*, *2*(7), e1501714–e1501714. https://doi.org/10.1126/sciadv.1501714

Wall, J. D., Schlebusch, S. A., Alberts, S. C., Cox, L. A., Snyder-Mackler, N., Nevonen, K. A.,

… Tung, J. (2016). Genomewide ancestry and divergence patterns from low-coverage sequencing data reveal a complex history of admixture in wild baboons. *Molecular Ecology*, *25*(14), 3469–3483. https://doi.org/10.1111/mec.13684

Wang, H., Xu, X., Vieira, F. G., Xiao, Y., Li, Z., Wang, J., … Chu, C. (2016). The Power of Inbreeding: NGS-Based GWAS of Rice Reveals Convergent Evolution during Rice Domestication. *Molecular Plant*, *9*(7), 975–985. https://doi.org/10.1016/j.molp.2016.04.018

Wang, J., Duncan, D., Shi, Z., & Zhang, B. (2013). WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Research*, *41*(W1), W77–W83. https://doi.org/10.1093/nar/gkt439

Wang, Jingwen, Skoog, T., Einarsdottir, E., Kaartokallio, T., Laivuori, H., Grauers, A., … Jiao, H. (2016). Investigation of rare and low-frequency variants using high-throughput sequencing with pooled DNA samples. *Scientific Reports*, *6*(August), 33256. https://doi.org/10.1038/srep33256

Wang, Z., Brickman, D., Greenan, B. J. W., & Yashayaev, I. (2016). An abrupt shift in the Labrador Current System in relation to winter NAO events. *Journal of Geophysical Research: Oceans*, *121*(7), 5338–5349. https://doi.org/10.1002/2016JC011721

Waples, R. K., Larson, W. A., & Waples, R. S. (2016). Estimating contemporary effective population size in non-model species using linkage disequilibrium across thousands of loci. *Heredity*, *117*(4), 233–240. https://doi.org/10.1038/hdy.2016.60

Waples, R. S. (1998). Separating the wheat from the chaff: patterns of genetic differentiation in high gene flow species. *Journal of Heredity*, *89*(5), 438–450. https://doi.org/10.1093/jhered/89.5.438

Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N., & Watson, M. (2015). Exome Sequencing: Current and Future Perspectives. *Genes|Genomes|Genetics*, *5*(8), 1543–1550. https://doi.org/10.1534/g3.115.018564

Watabe, S. (1999). Myogenic regulatory factors and muscle differentiation during ontogeny in fish. *Journal of Fish Biology*, *55*(A), 1–18. https://doi.org/10.1111/j.1095-8649.1999.tb01042.x

Waters, C. L., & Clark, K. J. (2005). 2005 summary of the weir herring tagging project with an update of the HSC/PRC/DFO herring tagging program. In *Canadian Science Advisory Secretariat Research Document, 2005/025*.

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, *38*(6), 1358. https://doi.org/10.2307/2408641

Wheeler, J. P., & Winters, G. . (1984a). Homing of Atlantic herring (Clupea harengus harengus) in Newfoundland waters as indicated by tagging data. *Can. J. Fish. Aquat. Sci.*, *41*, 108–117.

Wheeler, J. P., & Winters, G. H. (1984b). Migrations and stock relationships of east and southeast Newfoundland herring (Clupea harengus) as shown by tagging studies. *Journal of Northwest Atlantic Fishery Science*, *5*, 121–129.

Wiberg, R. A. W., Gaggiotti, O. E., Morrissey, M. B., & Ritchie, M. G. (2017). Identifying consistent allele frequency differences in studies of stratified populations. *Methods in Ecology and Evolution*, *8*(12), 1899–1909. https://doi.org/10.1111/2041-210X.12810

Winters, G. H., & Wheeler, J. P. (1987). Recruitment dynamics of spring-spawning herring in the Northwest Atlantic. *Canadian Journal of Fisheries and Aquatic Sciences*, *44*, 882–900.

Winters, G. H., & Wheeler, J. P. (1996). Environmental and phenotypic factors affecting the reproductive cycle of Atlantic herring. *ICES Journal of Marine Science*, *53*, 73–88. https://doi.org/10.1006/jmsc.1996.0007

Winters, G. H., Wheeler, J. P., & Dalley, E. L. (1986). Survival of a herring stock subjected to a catastrophic event and fluctuating environmental conditions. *Journal Du Conseil International Pour l'Exploration de La Mer*, *43*, v.

Winters, G. H., Wheeler, J. P., & Stansbury, D. (1993). Variability in the reproductive output of spring-spawning herring in the north-west atlantic. *ICES Journal of Marine Science*, Vol. 50, pp. 15–25. https://doi.org/10.1006/jmsc.1993.1003

Wittkopp, P. J., & Kalay, G. (2012). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics*, *13*(1), 59–69. https://doi.org/10.1038/nrg3095

Wong, P. B., Wiley, E. O., Johnson, W. E., Ryder, O. A., O'Brien, S. J., Haussler, D., … Murphy, R. W. (2012). Tissue sampling methods and standards for vertebrate genomics. *GigaScience*, *1*(1), 8. https://doi.org/10.1186/2047-217X-1-8

Wray, G. A. (2007). The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics*, *8*(3), 206–216. https://doi.org/10.1038/nrg2063

Wu, C.-I. (2001). The genic view of the process of speciation. *Journal of Evolutionary Biology*, *14*(6), 851–865. https://doi.org/10.1046/j.1420-9101.2001.00335.x

Xu, Q. -S., & Liang, Y. -Z. (2001). Monte–Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems*, *56*, 1–11.

Xue, Y., Prado-Martinez, J., Sudmant, P. H., Narasimhan, V., Ayub, Q., Szpak, M., … Scally, A. (2015). Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science*, *348*(6231), 242–245. https://doi.org/10.1126/science.aaa3952

Yang, H., & Wang, K. (2015). Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nature Protocols*, *10*(10), 1556–1566. https://doi.org/10.1038/nprot.2015.105

Yang, J., Li, W. R., Lv, F. H., He, S. G., Tian, S. L., Peng, W. F., … Liu, M. J. (2016). Whole-Genome Sequencing of Native Sheep Provides Insights into Rapid Adaptations to Extreme Environments. *Molecular Biology and Evolution*, *33*(10), 2576–2592. https://doi.org/10.1093/molbev/msw129

Ye, H., Meehan, J., Tong, W., & Hong, H. (2015). Alignment of short reads: A crucial step for application of next-generation sequencing data in precision medicine. *Pharmaceutics*, *7*(4), 523–541. https://doi.org/10.3390/pharmaceutics7040523

Yeaman, S., & Whitlock, M. C. (2011). The genetic architecture of adaptation under migration-selection balance. *Evolution*, *65*(7), 1897–1911. https://doi.org/10.1111/j.1558-5646.2011.01269.x

Zhang, G. (2015). Genomics: Bird sequencing project takes off. *Nature*, *522*(7554), 34–34. https://doi.org/10.1038/522034d

Zhao, S., Zheng, P., Dong, S., Zhan, X., Wu, Q., Guo, X., … Wei, F. (2012). Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation. *Nature Genetics*, *45*(1), 67–71. https://doi.org/10.1038/ng.2494

Zhou, X., Wang, B., Pan, Q., Zhang, J., Kumar, S., Sun, X., … Li, M. (2014). Whole-genome sequencing of the snub-nosed monkey provides insights into folivory and evolutionary history. *Nature Genetics*, *46*(12), 1303–1310. https://doi.org/10.1038/ng.3137

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, *1*(1), 3–14.

APPENDIX 1. DESCRIPTION OF ELECTRONIC SUPPLEMENTS

I have co-authored other relevant work as part of my PhD (Kerr et al. 2018) that is not included in this thesis. In addition, some of the tables generated in the thesis are too large for a printed format (contain thousands of rows). Therefore, all these files have been submitted as electronic supplementary material. Below a list of these files is provided:

Chapter 3:

Table S3.2 Loci showing strong genetic differentiation between spring- and autumn-spawning herring. Gene names are indicated if the SNP occurs within 5 kb upstream or 5 kb downstream of annotated genes. Loci significant in both NE and NW Atlantic populations are highlighted in green; loci significant only in the NW Atlantic populations are highlighted in pink.

Table S3.3 Previously identified loci showing the most consistent association with differences in salinity

Table S3.4 Genetic distance matrix used for building the phylogenetic tree among 26 herring populations used for Fig. 3.2. Details for sample IDs are given in Table 3.1.

Chapter 5:

Table S5.1. SNP loci that passed quality filters and constitute the SPW- and LAT-panels.

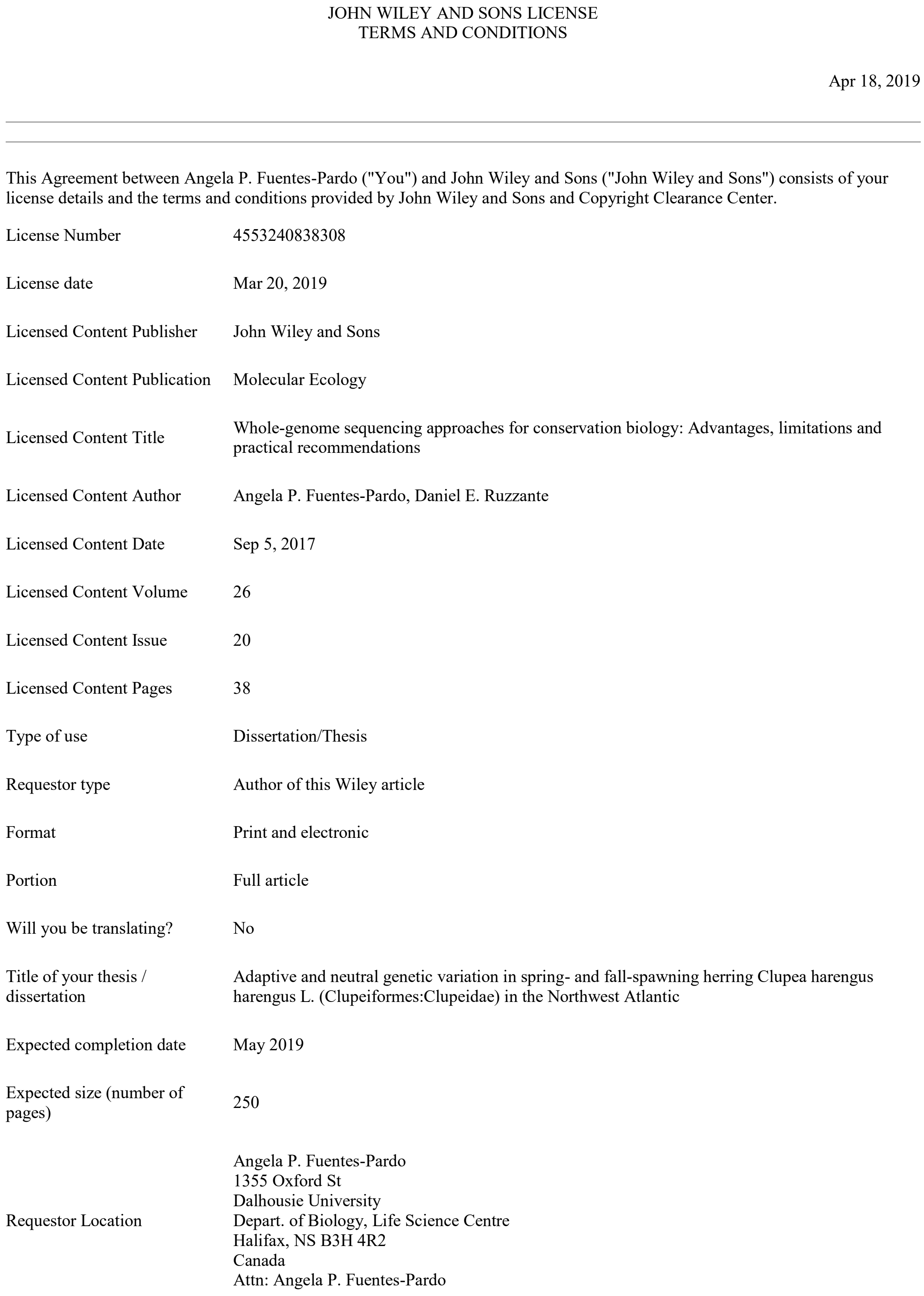
Table S5.2. Pairwise F_{ST} and P-values for the SPW-panel.

Table S5.3. Pairwise F_{ST} and P-values for the LAT-panel.

Kerr, Q., **Fuentes-Pardo, A. P.**, Kho, J., McDermid, J. L., & Ruzzante, D. E. (2018). Temporal stability and assignment power of adaptively divergent genomic regions between herring (*Clupea harengus*) seasonal spawning aggregations. *Ecology and Evolution*, (October), ece3.4768. https://doi.org/10.1002/ece3.4768

APPENDIX 2. COPYRIGHT LETTERS

For the 2017 Molecular Ecology paper of chapter 2:



JOHN WILEY AND SONS LICENSE
TERMS AND CONDITIONS

Apr 18, 2019

This Agreement between Angela P. Fuentes-Pardo ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number	4553240838308
License date	Mar 20, 2019
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	Molecular Ecology
Licensed Content Title	Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations
Licensed Content Author	Angela P. Fuentes-Pardo, Daniel E. Ruzzante
Licensed Content Date	Sep 5, 2017
Licensed Content Volume	26
Licensed Content Issue	20
Licensed Content Pages	38
Type of use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Print and electronic
Portion	Full article
Will you be translating?	No
Title of your thesis / dissertation	Adaptive and neutral genetic variation in spring- and fall-spawning herring Clupea harengus harengus L. (Clupeiformes:Clupeidae) in the Northwest Atlantic
Expected completion date	May 2019
Expected size (number of pages)	250
Requestor Location	Angela P. Fuentes-Pardo 1355 Oxford St Dalhousie University Depart. of Biology, Life Science Centre Halifax, NS B3H 4R2 Canada Attn: Angela P. Fuentes-Pardo

Publisher Tax ID	EU826007151
Total	0.00 CAD

Terms and Conditions

TERMS AND CONDITIONS

This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one of its group companies (each a"Wiley Company") or handled on behalf of a society with which a Wiley Company has exclusive publishing rights in relation to a particular work (collectively "WILEY"). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright Clearance Center Inc., ("CCC's Billing and Payment terms and conditions"), at the time that you opened your RightsLink account (these are available at any time at http://myaccount.copyright.com).

Terms and Conditions

- The materials you have requested permission to reproduce or reuse (the "Wiley Materials") are protected by copyright.

- You are hereby granted a personal, non-exclusive, non-sub licensable (on a stand-alone basis), non-transferable, worldwide, limited license to reproduce the Wiley Materials for the purpose specified in the licensing process. This license, **and any CONTENT (PDF or image file) purchased as part of your order,** is for a one-time use only and limited to any maximum distribution number specified in the license. The first instance of republication or reuse granted by this license must be completed within two years of the date of the grant of this license (although copies prepared before the end date may be distributed thereafter). The Wiley Materials shall not be used in any other manner or for any other purpose, beyond what is granted in the license. Permission is granted subject to an appropriate acknowledgement given to the author, title of the material/book/journal and the publisher. You shall also duplicate the copyright notice that appears in the Wiley publication in your use of the Wiley Material. Permission is also granted on the understanding that nowhere in the text is a previously published source acknowledged for all or part of this Wiley Material. Any third party content is expressly excluded from this permission.

- With respect to the Wiley Materials, all rights are reserved. Except as expressly granted by the terms of the license, no part of the Wiley Materials may be copied, modified, adapted (except for minor reformatting required by the new Publication), translated, reproduced, transferred or distributed, in any form or by any means, and no derivative works may be made based on the Wiley Materials without the prior permission of the respective copyright owner.**For STM Signatory Publishers clearing permission under the terms of the STM Permissions Guidelines only, the terms of the license are extended to include subsequent editions and for editions in other languages, provided such editions are for the work as a whole in situ and does not involve the separate exploitation of the permitted figures or extracts,** You may not alter, remove or suppress in any manner any copyright, trademark or other notices displayed by the Wiley Materials. You may not license, rent, sell, loan, lease, pledge, offer as security, transfer or assign the Wiley Materials on a stand-alone basis, or any of the rights granted to you hereunder to any other person.

- The Wiley Materials and all of the intellectual property rights therein shall at all times remain the exclusive property of John Wiley & Sons Inc, the Wiley Companies, or their respective licensors, and your interest therein is only that of having possession of and the right to reproduce the Wiley Materials pursuant to Section 2 herein during the continuance of this Agreement. You agree that you own no right, title or interest in or to the Wiley Materials or any of the intellectual property rights therein. You shall have no rights hereunder other than the license as provided for above in Section 2. No right, license or interest to any trademark, trade name, service mark or other branding ("Marks") of WILEY or its licensors is granted hereunder, and you agree that you shall not assert any such right, license or interest with respect thereto

- NEITHER WILEY NOR ITS LICENSORS MAKES ANY WARRANTY OR REPRESENTATION OF ANY KIND TO YOU OR ANY THIRD PARTY, EXPRESS, IMPLIED OR STATUTORY, WITH RESPECT TO THE MATERIALS OR THE ACCURACY OF ANY INFORMATION CONTAINED IN THE MATERIALS, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTY OF MERCHANTABILITY, ACCURACY, SATISFACTORY QUALITY, FITNESS FOR A PARTICULAR PURPOSE, USABILITY, INTEGRATION OR NON-INFRINGEMENT AND ALL SUCH WARRANTIES ARE HEREBY EXCLUDED BY WILEY AND ITS LICENSORS AND WAIVED BY YOU.

- WILEY shall have the right to terminate this Agreement immediately upon breach of this Agreement by you.

- You shall indemnify, defend and hold harmless WILEY, its Licensors and their respective directors, officers, agents and employees, from and against any actual or threatened claims, demands, causes of action or proceedings arising from any breach of this Agreement by you.

- IN NO EVENT SHALL WILEY OR ITS LICENSORS BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR ENTITY FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, EXEMPLARY OR PUNITIVE DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, PROVISIONING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

- Should any provision of this Agreement be held by a court of competent jurisdiction to be illegal, invalid, or unenforceable, that provision shall be deemed amended to achieve as nearly as possible the same economic effect as the original provision, and the legality, validity and enforceability of the remaining provisions of this Agreement shall not be affected or impaired thereby.

- The failure of either party to enforce any term or condition of this Agreement shall not constitute a waiver of either party's right to enforce each and every term and condition of this Agreement. No breach under this agreement shall be deemed waived or excused by either party unless such waiver or consent is in writing signed by the party granting such waiver or consent. The waiver by or consent of a party to a breach of any provision of this Agreement shall not operate or be construed as a waiver of or consent to any other or subsequent breach by such other party.

- This Agreement may not be assigned (including by operation of law or otherwise) by you without WILEY's prior written consent.

- Any fee required for this permission shall be non-refundable after thirty (30) days from receipt by the CCC.

- These terms and conditions together with CCC's Billing and Payment terms and conditions (which are incorporated herein) form the entire agreement between you and WILEY concerning this licensing transaction and (in the absence of fraud) supersedes all prior agreements and representations of the parties, oral or written. This Agreement may not be amended except in writing signed by both parties. This Agreement shall be binding upon and inure to the benefit of the parties' successors, legal representatives, and authorized assigns.

- In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall prevail.

- WILEY expressly reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

- This Agreement will be void if the Type of Use, Format, Circulation, or Requestor Type was misrepresented during the licensing process.

- This Agreement shall be governed by and construed in accordance with the laws of the State of New York, USA, without regards to such state's conflict of law rules. Any legal action, suit or proceeding arising out of or relating to these Terms and Conditions or the breach thereof shall be instituted in a court of competent jurisdiction in New York County in the State of New York in the United States of America and each party hereby consents and submits to the personal jurisdiction of such court, waives any objection to venue in such court and consents to service of process by registered or certified mail, return receipt requested, at the last known address of such party.

WILEY OPEN ACCESS TERMS AND CONDITIONS

Wiley Publishes Open Access Articles in fully Open Access Journals and in Subscription journals offering Online Open. Although most of the fully Open Access journals publish open access articles under the terms of the Creative Commons Attribution (CC BY) License only, the subscription journals and a few of the Open Access Journals offer a choice of Creative Commons Licenses. The license type is clearly identified on the article.

The Creative Commons Attribution License

The Creative Commons Attribution License (CC-BY) allows users to copy, distribute and transmit an article, adapt the article and make commercial use of the article. The CC-BY license permits commercial and non-

Creative Commons Attribution Non-Commercial License

The Creative Commons Attribution Non-Commercial (CC-BY-NC)License permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.(see below)

Creative Commons Attribution-Non-Commercial-NoDerivs License

The Creative Commons Attribution Non-Commercial-NoDerivs License (CC-BY-NC-ND) permits use, distribution and reproduction in any medium, provided the original work is properly cited, is not used for commercial purposes and no modifications or adaptations are made. (see below)

Use by commercial "for-profit" organizations

Use of Wiley Open Access articles for commercial, promotional, or marketing purposes requires further explicit permission from Wiley and will be subject to a fee.

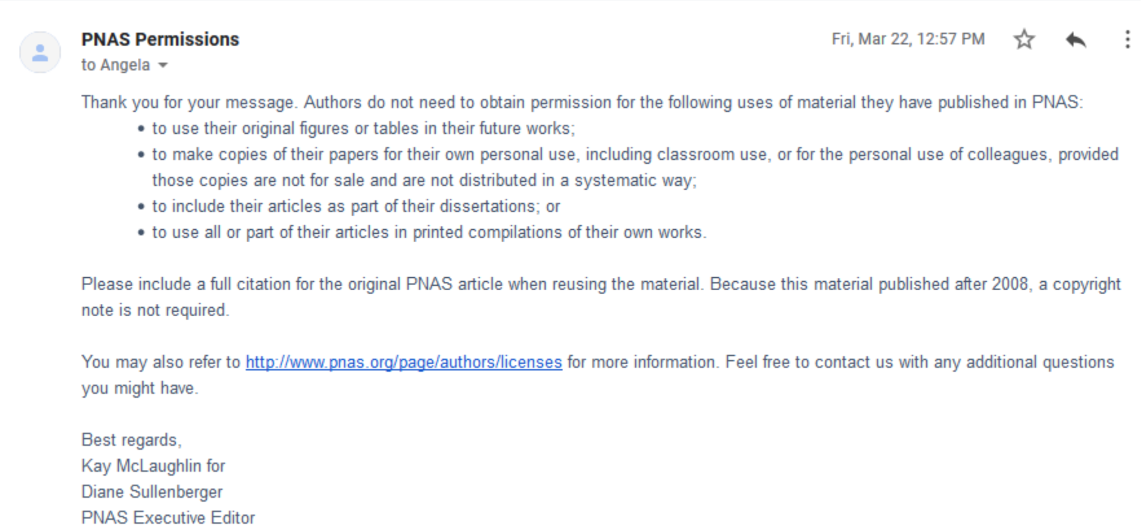
Further details can be found on Wiley Online Library http://olabout.wiley.com/WileyCDA/Section/id-410895.html

Other Terms and Conditions:

v1.10 Last updated September 2015

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

For the 2017 PNAS paper of chapter 3 is not necessary to obtain permission from the journal (see email below):



PNAS Permissions
to Angela

Fri, Mar 22, 12:57 PM

Thank you for your message. Authors do not need to obtain permission for the following uses of material they have published in PNAS:

- to use their original figures or tables in their future works;
- to make copies of their papers for their own personal use, including classroom use, or for the personal use of colleagues, provided those copies are not for sale and are not distributed in a systematic way;
- to include their articles as part of their dissertations; or
- to use all or part of their articles in printed compilations of their own works.

Please include a full citation for the original PNAS article when reusing the material. Because this material published after 2008, a copyright note is not required.

You may also refer to http://www.pnas.org/page/authors/licenses for more information. Feel free to contact us with any additional questions you might have.

Best regards,
Kay McLaughlin for
Diane Sullenberger
PNAS Executive Editor

FGS staff recommended to obtain personal permission from the co-authors of the paper. Permission letters of all coauthors can be found in the next pages:

April, 2019

Dear coauthor,

I am preparing my PhD thesis for submission to the Faculty of Graduate Studies at Dalhousie University, Halifax, Nova Scotia, Canada. I am seeking your permission to include a manuscript version of the following paper as a chapter in the thesis:

> Parallel adaptive evolution of geographically distant herring populations on both sides of the North Atlantic Ocean
> Sangeet Lamichhaney, Angela P. Fuentes-Pardo, Nima Rafati, Nils Ryman, Gregory R. McCracken, Christina Bourne, Rabindra Singh, Daniel E. Ruzzante, and Leif Andersson
> PNAS April 25, 2017 114 (17) E3452-E3461, https://doi.org/10.1073/pnas.1617728114, 2017

Canadian graduate theses are reproduced by the Library and Archives of Canada (formerly National Library of Canada) through a non-exclusive, world-wide license to reproduce, loan, distribute, or sell theses. I am also seeking your permission for the material described above to be reproduced and distributed by the LAC(NLC). Further details about the LAC(NLC) thesis program are available on the LAC(NLC) website (www.nlc-bnc.ca).

Full publication details and a copy of this permission letter will be included in the thesis.

Yours sincerely,

Angela Patricia Fuentes Pardo
PhD candidate, Department of Biology
Dalhousie University

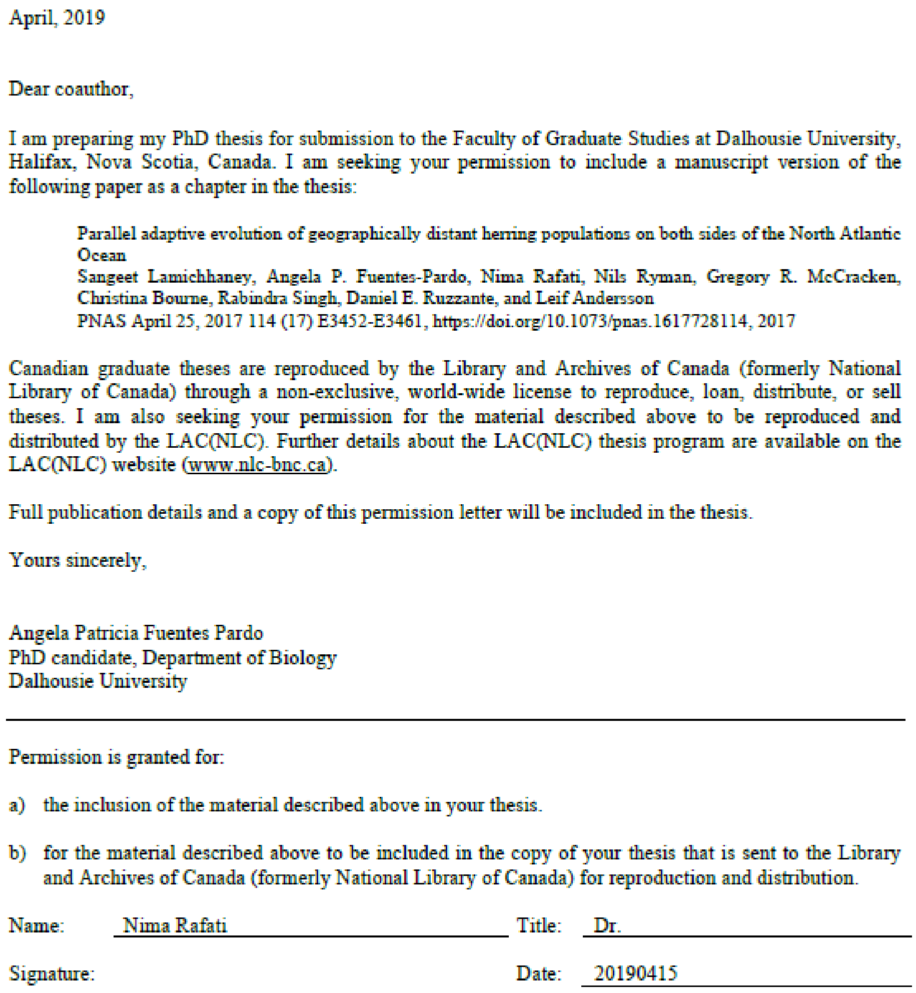
Permission is granted for:

a) the inclusion of the material described above in your thesis.

b) for the material described above to be included in the copy of your thesis that is sent to the Library and Archives of Canada (formerly National Library of Canada) for reproduction and distribution.

Name: Dr. Sangeet Lamichhaney Title: Wenner-Gren Fellow, Harvard University

Signature: Date: May 1, 2019



April, 2019

Dear coauthor,

I am preparing my PhD thesis for submission to the Faculty of Graduate Studies at Dalhousie University, Halifax, Nova Scotia, Canada. I am seeking your permission to include a manuscript version of the following paper as a chapter in the thesis:

> Parallel adaptive evolution of geographically distant herring populations on both sides of the North Atlantic Ocean
> Sangeet Lamichhaney, Angela P. Fuentes-Pardo, Nima Rafati, Nils Ryman, Gregory R. McCracken, Christina Bourne, Rabindra Singh, Daniel E. Ruzzante, and Leif Andersson
> PNAS April 25, 2017 114 (17) E3452-E3461, https://doi.org/10.1073/pnas.1617728114, 2017

Canadian graduate theses are reproduced by the Library and Archives of Canada (formerly National Library of Canada) through a non-exclusive, world-wide license to reproduce, loan, distribute, or sell theses. I am also seeking your permission for the material described above to be reproduced and distributed by the LAC(NLC). Further details about the LAC(NLC) thesis program are available on the LAC(NLC) website (www.nlc-bnc.ca).

Full publication details and a copy of this permission letter will be included in the thesis.

Yours sincerely,

Angela Patricia Fuentes Pardo
PhD candidate, Department of Biology
Dalhousie University

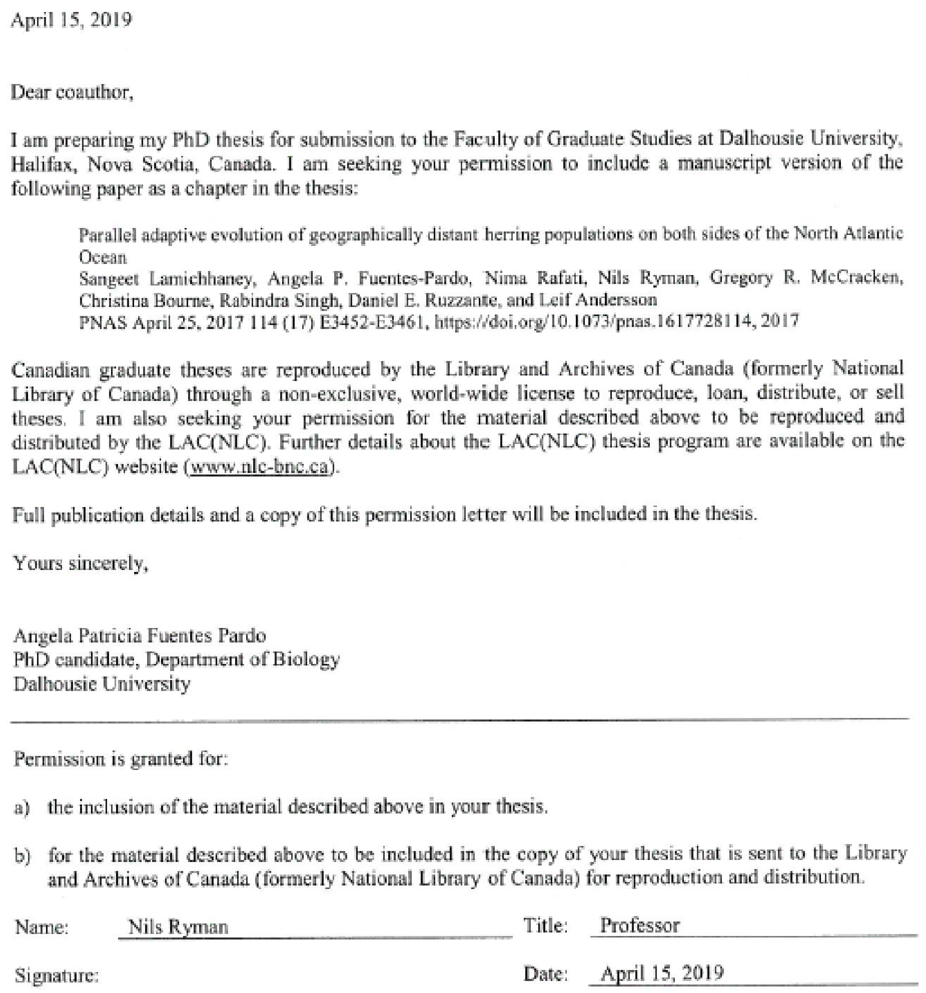
Permission is granted for:

a) the inclusion of the material described above in your thesis.

b) for the material described above to be included in the copy of your thesis that is sent to the Library and Archives of Canada (formerly National Library of Canada) for reproduction and distribution.

Name: Nima Rafati Title: Dr.

Signature: Date: 20190415



April 15, 2019

Dear coauthor,

I am preparing my PhD thesis for submission to the Faculty of Graduate Studies at Dalhousie University, Halifax, Nova Scotia, Canada. I am seeking your permission to include a manuscript version of the following paper as a chapter in the thesis:

> Parallel adaptive evolution of geographically distant herring populations on both sides of the North Atlantic Ocean
> Sangeet Lamichhaney, Angela P. Fuentes-Pardo, Nima Rafati, Nils Ryman, Gregory R. McCracken, Christina Bourne, Rabindra Singh, Daniel E. Ruzzante, and Leif Andersson
> PNAS April 25, 2017 114 (17) E3452-E3461, https://doi.org/10.1073/pnas.1617728114, 2017

Canadian graduate theses are reproduced by the Library and Archives of Canada (formerly National Library of Canada) through a non-exclusive, world-wide license to reproduce, loan, distribute, or sell theses. I am also seeking your permission for the material described above to be reproduced and distributed by the LAC(NLC). Further details about the LAC(NLC) thesis program are available on the LAC(NLC) website (www.nlc-bnc.ca).

Full publication details and a copy of this permission letter will be included in the thesis.

Yours sincerely,

Angela Patricia Fuentes Pardo
PhD candidate, Department of Biology
Dalhousie University

Permission is granted for:

a) the inclusion of the material described above in your thesis.

b) for the material described above to be included in the copy of your thesis that is sent to the Library and Archives of Canada (formerly National Library of Canada) for reproduction and distribution.

Name: Nils Ryman   Title: Professor

Signature: Date: April 15, 2019

April, 2019

Dear coauthor,

I am preparing my PhD thesis for submission to the Faculty of Graduate Studies at Dalhousie University, Halifax, Nova Scotia, Canada. I am seeking your permission to include a manuscript version of the following paper as a chapter in the thesis:

> Parallel adaptive evolution of geographically distant herring populations on both sides of the North Atlantic Ocean
> Sangeet Lamichhaney, Angela P. Fuentes-Pardo, Nima Rafati, Nils Ryman, Gregory R. McCracken, Christina Bourne, Rabindra Singh, Daniel E. Ruzzante, and Leif Andersson
> PNAS April 25, 2017 114 (17) E3452-E3461, https://doi.org/10.1073/pnas.1617728114, 2017

Canadian graduate theses are reproduced by the Library and Archives of Canada (formerly National Library of Canada) through a non-exclusive, world-wide license to reproduce, loan, distribute, or sell theses. I am also seeking your permission for the material described above to be reproduced and distributed by the LAC(NLC). Further details about the LAC(NLC) thesis program are available on the LAC(NLC) website (www.nlc-bnc.ca).

Full publication details and a copy of this permission letter will be included in the thesis.

Yours sincerely,

Angela Patricia Fuentes Pardo
PhD candidate, Department of Biology
Dalhousie University

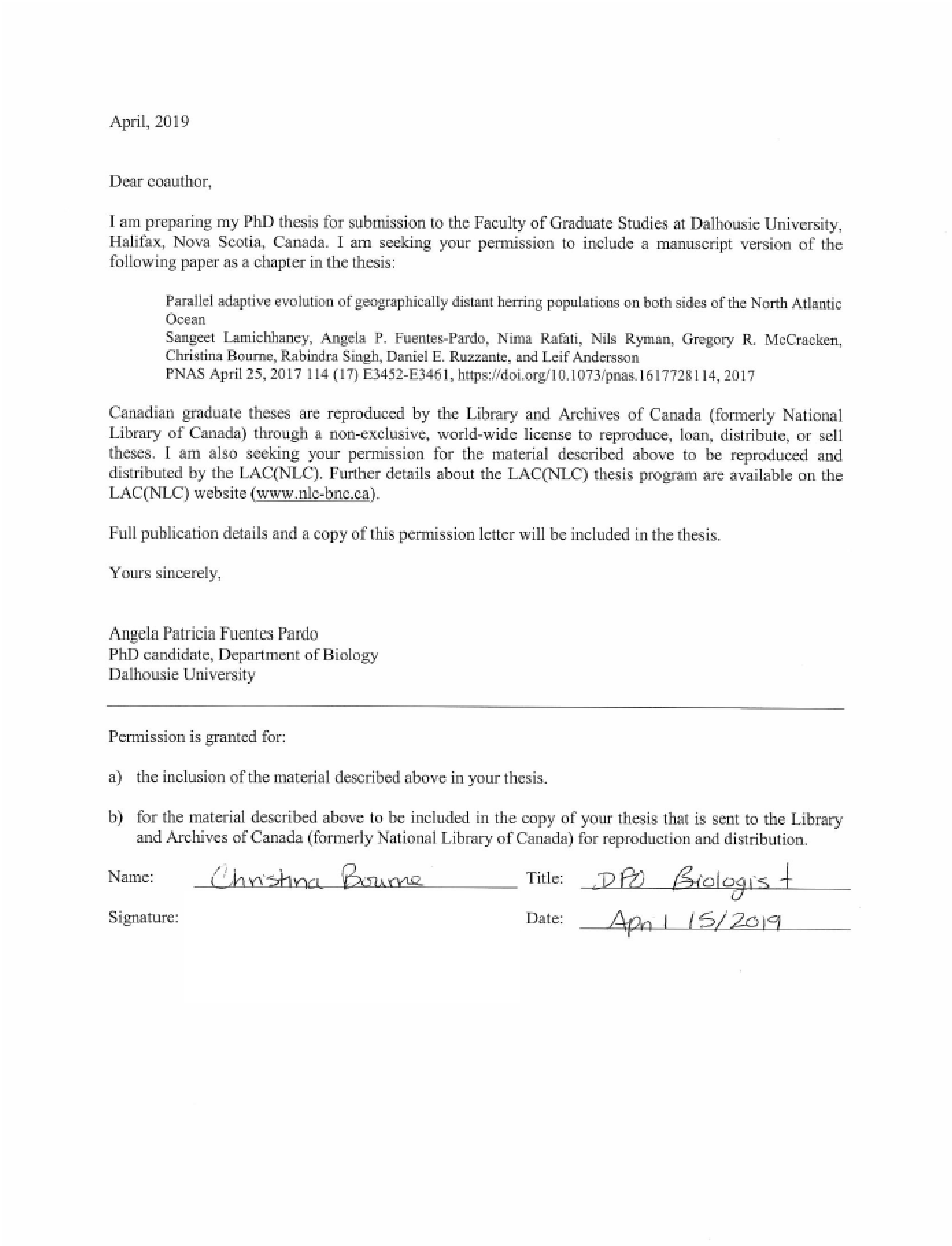
Permission is granted for:

a) the inclusion of the material described above in your thesis.

b) for the material described above to be included in the copy of your thesis that is sent to the Library and Archives of Canada (formerly National Library of Canada) for reproduction and distribution.

Name: Gregory McCracken Title: Research Associate

Signature: Date: April 15/2019



April, 2019

Dear coauthor,

I am preparing my PhD thesis for submission to the Faculty of Graduate Studies at Dalhousie University, Halifax, Nova Scotia, Canada. I am seeking your permission to include a manuscript version of the following paper as a chapter in the thesis:

> Parallel adaptive evolution of geographically distant herring populations on both sides of the North Atlantic Ocean
> Sangeet Lamichhaney, Angela P. Fuentes-Pardo, Nima Rafati, Nils Ryman, Gregory R. McCracken, Christina Bourne, Rabindra Singh, Daniel E. Ruzzante, and Leif Andersson
> PNAS April 25, 2017 114 (17) E3452-E3461, https://doi.org/10.1073/pnas.1617728114, 2017

Canadian graduate theses are reproduced by the Library and Archives of Canada (formerly National Library of Canada) through a non-exclusive, world-wide license to reproduce, loan, distribute, or sell theses. I am also seeking your permission for the material described above to be reproduced and distributed by the LAC(NLC). Further details about the LAC(NLC) thesis program are available on the LAC(NLC) website (www.nlc-bnc.ca).

Full publication details and a copy of this permission letter will be included in the thesis.

Yours sincerely,

Angela Patricia Fuentes Pardo
PhD candidate, Department of Biology
Dalhousie University

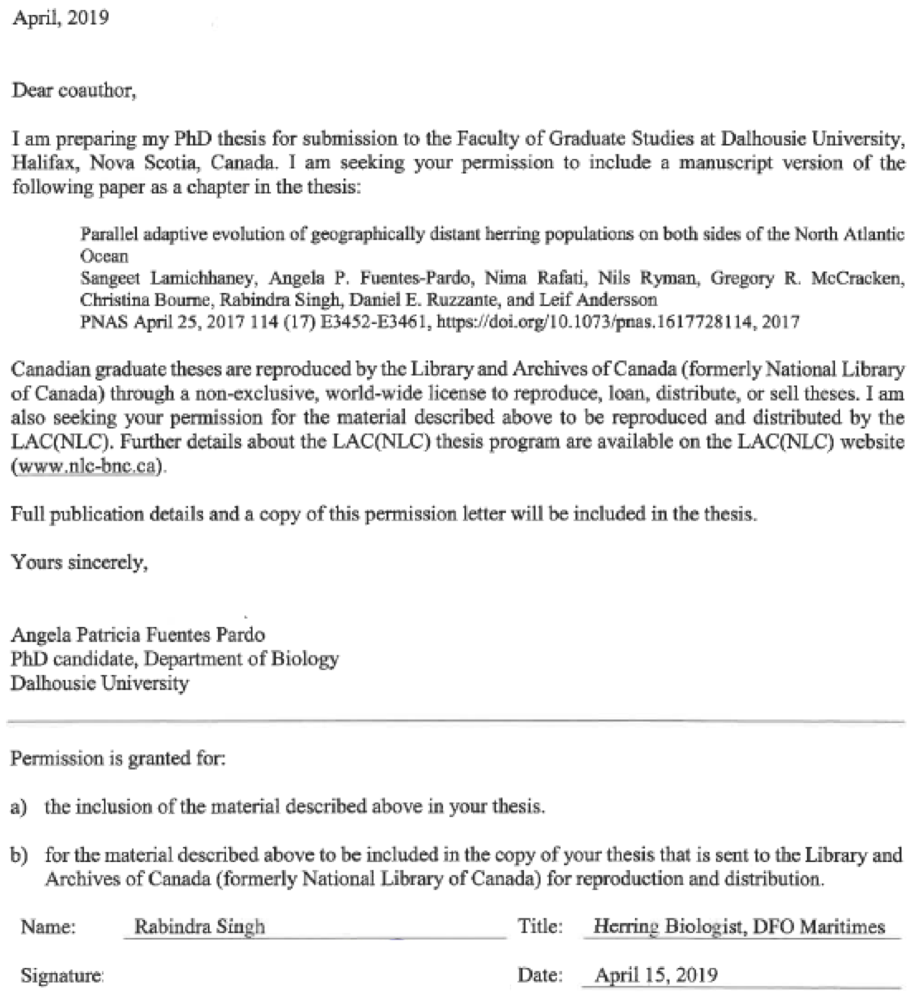
Permission is granted for:

a) the inclusion of the material described above in your thesis.

b) for the material described above to be included in the copy of your thesis that is sent to the Library and Archives of Canada (formerly National Library of Canada) for reproduction and distribution.

Name: Christina Bourne Title: DFO Biologist

Signature: Date: April 15/2019



April, 2019

Dear coauthor,

I am preparing my PhD thesis for submission to the Faculty of Graduate Studies at Dalhousie University, Halifax, Nova Scotia, Canada. I am seeking your permission to include a manuscript version of the following paper as a chapter in the thesis:

> Parallel adaptive evolution of geographically distant herring populations on both sides of the North Atlantic Ocean
> Sangeet Lamichhaney, Angela P. Fuentes-Pardo, Nima Rafati, Nils Ryman, Gregory R. McCracken, Christina Bourne, Rabindra Singh, Daniel E. Ruzzante, and Leif Andersson
> PNAS April 25, 2017 114 (17) E3452-E3461, https://doi.org/10.1073/pnas.1617728114, 2017

Canadian graduate theses are reproduced by the Library and Archives of Canada (formerly National Library of Canada) through a non-exclusive, world-wide license to reproduce, loan, distribute, or sell theses. I am also seeking your permission for the material described above to be reproduced and distributed by the LAC(NLC). Further details about the LAC(NLC) thesis program are available on the LAC(NLC) website (www.nlc-bnc.ca).

Full publication details and a copy of this permission letter will be included in the thesis.

Yours sincerely,

Angela Patricia Fuentes Pardo
PhD candidate, Department of Biology
Dalhousie University

Permission is granted for:

a) the inclusion of the material described above in your thesis.

b) for the material described above to be included in the copy of your thesis that is sent to the Library and Archives of Canada (formerly National Library of Canada) for reproduction and distribution.

Name: Rabindra Singh Title: Herring Biologist, DFO Maritimes

Signature: Date: April 15, 2019

April, 2019

Dear coauthor,

I am preparing my PhD thesis for submission to the Faculty of Graduate Studies at Dalhousie University, Halifax, Nova Scotia, Canada. I am seeking your permission to include a manuscript version of the following paper as a chapter in the thesis:

> Parallel adaptive evolution of geographically distant herring populations on both sides of the North Atlantic Ocean
> Sangeet Lamichhaney, Angela P. Fuentes-Pardo, Nima Rafati, Nils Ryman, Gregory R. McCracken, Christina Bourne, Rabindra Singh, Daniel E. Ruzzante, and Leif Andersson
> PNAS April 25, 2017 114 (17) E3452-E3461, https://doi.org/10.1073/pnas.1617728114, 2017

Canadian graduate theses are reproduced by the Library and Archives of Canada (formerly National Library of Canada) through a non-exclusive, world-wide license to reproduce, loan, distribute, or sell theses. I am also seeking your permission for the material described above to be reproduced and distributed by the LAC(NLC). Further details about the LAC(NLC) thesis program are available on the LAC(NLC) website (www.nlc-bnc.ca).

Full publication details and a copy of this permission letter will be included in the thesis.

Yours sincerely,

Angela Patricia Fuentes Pardo
PhD candidate, Department of Biology
Dalhousie University

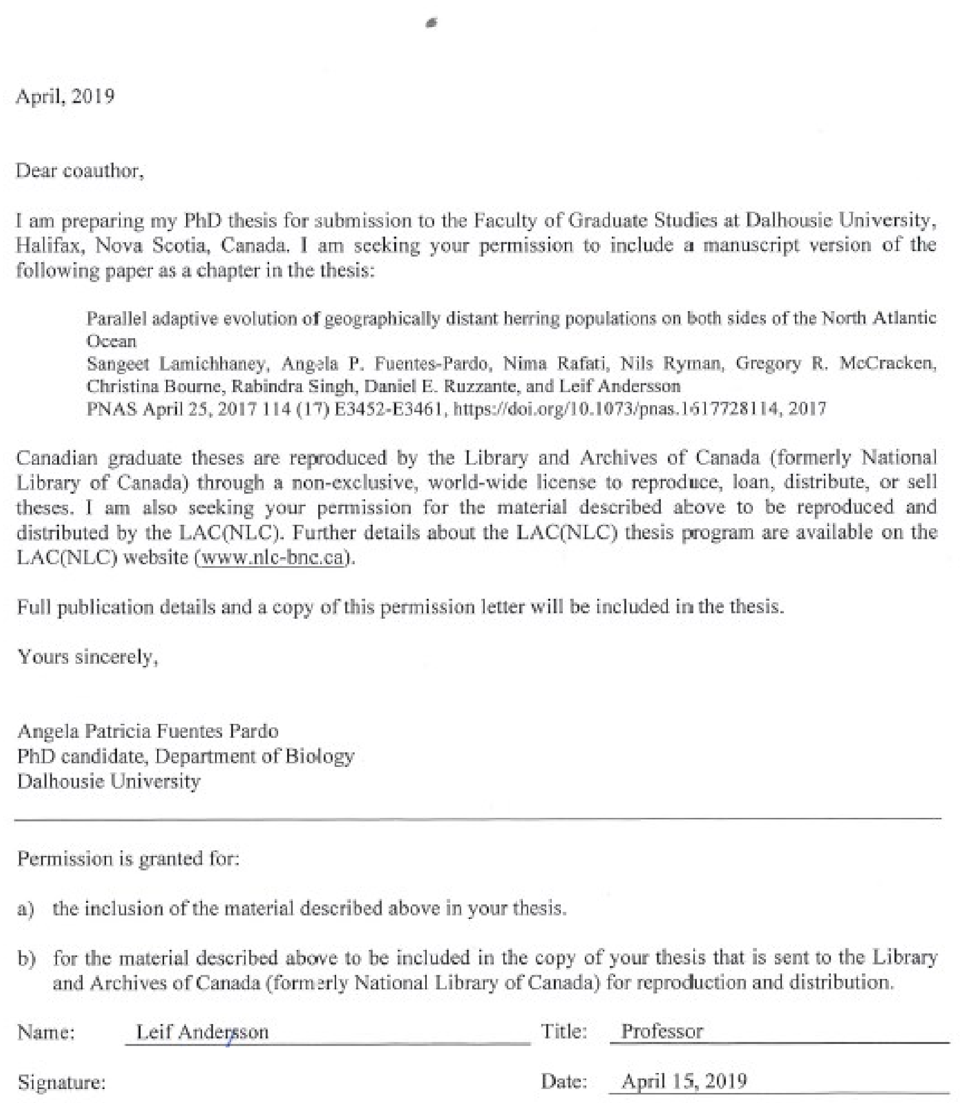
Permission is granted for:

a) the inclusion of the material described above in your thesis.

b) for the material described above to be included in the copy of your thesis that is sent to the Library and Archives of Canada (formerly National Library of Canada) for reproduction and distribution.

Name: Daniel Ruzzante Title: Professor

Signature: Date: April 15, 2019



April, 2019

Dear coauthor,

I am preparing my PhD thesis for submission to the Faculty of Graduate Studies at Dalhousie University, Halifax, Nova Scotia, Canada. I am seeking your permission to include a manuscript version of the following paper as a chapter in the thesis:

> Parallel adaptive evolution of geographically distant herring populations on both sides of the North Atlantic Ocean
> Sangeet Lamichhaney, Angela P. Fuentes-Pardo, Nima Rafati, Nils Ryman, Gregory R. McCracken, Christina Bourne, Rabindra Singh, Daniel E. Ruzzante, and Leif Andersson
> PNAS April 25, 2017 114 (17) E3452-E3461, https://doi.org/10.1073/pnas.1617728114, 2017

Canadian graduate theses are reproduced by the Library and Archives of Canada (formerly National Library of Canada) through a non-exclusive, world-wide license to reproduce, loan, distribute, or sell theses. I am also seeking your permission for the material described above to be reproduced and distributed by the LAC(NLC). Further details about the LAC(NLC) thesis program are available on the LAC(NLC) website (www.nlc-bnc.ca).

Full publication details and a copy of this permission letter will be included in the thesis.

Yours sincerely,

Angela Patricia Fuentes Pardo
PhD candidate, Department of Biology
Dalhousie University

Permission is granted for:

a) the inclusion of the material described above in your thesis.

b) for the material described above to be included in the copy of your thesis that is sent to the Library and Archives of Canada (formerly National Library of Canada) for reproduction and distribution.

Name: Leif Andersson Title: Professor

Signature: Date: April 15, 2019