



Norwegian University of Life Sciences  
Faculty of Biosciences  
Department of Animal and Aquacultural Sciences

Philosophiae Doctor (PhD)  
Thesis 2023:26

# Genomic structural variations as drivers of adaptation in salmonid fishes

Strukturell variasjon som påvirker genetisk  
miljøtilpasning i laksefisk

Kristina Severine Rudskjær Stenløkk



# Genomic structural variations as drivers of adaptation in salmonid fishes

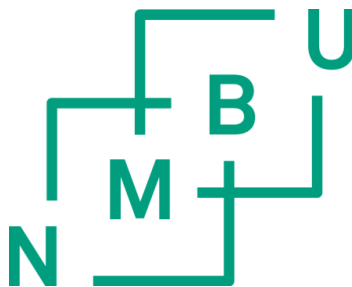
Strukturell variasjon som påvirker genetisk miljøtilpasning i laksefisk

Philosophiae Doctor (PhD) Thesis

Kristina Severine Rudskjær Stenløkk

Norwegian University of Life Sciences  
Faculty of Biosciences  
Department of Animal and Aquacultural Sciences

Ås 2023





# Supervisors and Evaluation Committee

## PhD supervisors

### **Prof. Sigbjørn Lien**

Faculty of Biosciences,  
Department of Animal and  
Aquacultural Sciences, CIGENE,  
Norwegian University of Life Sciences,  
P.O. Box 5003 NMBU,  
1432 Ås, Norway.  
[sigbjorn.lien@nmbu.no](mailto:sigbjorn.lien@nmbu.no)

### **Dr. Michel Moser**

Faculty of Biosciences,  
Department of Animal and  
Aquacultural Sciences, CIGENE,  
Norwegian University of Life Sciences,  
P.O. Box 5003 NMBU,  
1432 Ås, Norway.  
[michel.moser@nmbu.no](mailto:michel.moser@nmbu.no)

### **Dr. Marie Saitou**

Faculty of Biosciences,  
Department of Animal and  
Aquacultural Sciences, CIGENE,  
Norwegian University of Life Sciences,  
P.O. Box 5003 NMBU,  
1432 Ås, Norway.  
[marie.saitou@nmbu.no](mailto:marie.saitou@nmbu.no)

### **Prof. Simen Rød Sandve**

Faculty of Biosciences,  
Department of Animal and  
Aquacultural Sciences, CIGENE,  
Norwegian University of Life Sciences,  
P.O. Box 5003 NMBU,  
1432 Ås, Norway.  
[simen.sandve@nmbu.no](mailto:simen.sandve@nmbu.no)

### **Dr. Nicola Barson**

Faculty of Biosciences,  
Department of Animal and  
Aquacultural Sciences, CIGENE,  
Norwegian University of Life Sciences,  
P.O. Box 5003 NMBU,  
1432 Ås, Norway.  
[nicola.barson@nmbu.no](mailto:nicola.barson@nmbu.no)

## PhD Evaluation Committee

### **Assoc. Prof. Maren Wellenreuther**

School of Biological Sciences,  
University of Auckland,  
Biology building - Bldg  
106, 5 Symonds St, Auckland Central,  
Auckland, 1010,  
New Zealand.  
[m.wellenreuther@auckland.ac.nz](mailto:m.wellenreuther@auckland.ac.nz)

### **Prof. Martien A. M. Groenen**

Department of Animal Sciences,  
Wageningen University,  
PO Box 338,  
6700AH Wageningen,  
The Netherlands.  
[martien.groenen@wur.nl](mailto:martien.groenen@wur.nl)

### **Prof. Dag Inge Våge**

Faculty of Biosciences,  
Department of Animal and  
Aquacultural Sciences, CIGENE,  
Norwegian University of Life Sciences,  
P.O. Box 5003 NMBU,  
1432 Ås, Norway.  
[daginge.vage@nmbu.no](mailto:daginge.vage@nmbu.no)



# Acknowledgements

The work presented in this thesis has been carried out at the Faculty of Biosciences at the Norwegian University of Life Sciences (NMBU).

First and foremost, I would like to thank my main supervisor Prof. Sigbjørn Lien for being a fantastic supervisor throughout these years. The supervision meetings have been a great place to get new ideas and discuss problems, but all our “quick” chats (usually ending up as hour long discussions) have also been extremely valuable to me. Your devotion to science and salmon is astonishing and I am very grateful that you trusted me with this project and the amazing dataset. A big thanks to my co-supervisors Dr. Michel Moser, Dr. Marie Saitou, Dr. Nicola Barson and Prof. Simen Sandve for excellent guidance and support. You have helped me become a better bioinformatician, evolutionary biologist, scientific writer and taught me how to become a researcher. Also, I would like to thank all collaborators, in particular I would like to thank Matthew Kent for your long-read enthusiasm, Torfinn Nome for all bioinformatic first aid, and Claire Mérot and Louis Bernatchez for including me in the Whitefish project and being so kind and welcoming. A special thanks to Anna Sofie Kjelstrup who contributed with impressive work in genome-graphs and being a pleasure to work with. I would also like to thank Øystein Monsen, our collaboration was a lot of fun and I highly appreciate all your reflections about science, TEs and life.

I would also like to thank all employees at CIGENE for providing a welcoming and fun workplace. Thanks to my office colleagues Darshan, Noman, Øyvind and Marius. Sharing office with you has been a joy and a much-needed arena for sharing frustrations and achievements. A very special thanks to Martin Paliocha for your friendship and your immense knowledge about most things. You are a huge inspiration to me.

I owe my family a big thanks for supporting me through this time and I am also very lucky to have such amazing friends, many of which sharing this PhD experience with me. Marie, Guro, Sofie, Thea, Kristin, Lars, Sara, and Pelle – thank you so much for your friendship and support! And of course, a big thanks to my favourite person in the world (and partner) Lars. I know living with a stressed PhD student isn't always the best, but I am eternally grateful for your patience and compassion, especially through these last months.

Lastly, I would like to thank Simon, Arnold, Alto, Tanner, Tess, Klopp, Barry, Maxine, Bond, Brian and Louis, our long-read sequenced Atlantic salmon, for their (unknowing) sacrifice to science, and whom I at times have spent more time with than my friends and family.

Drammen, January 2023  
Kristina S. R. Stenløkk



# Table of Contents

Supervisors and Evaluation Committee .....	ii
Acknowledgements.....	iv
<b>1 List of papers .....</b>	<b>1</b>
<b>2 Abstract .....</b>	<b>2</b>
<b>3 Norsk sammendrag.....</b>	<b>4</b>
<b>4 Synopsis.....</b>	<b>6</b>
4.1 Introduction.....	6
4.1.1 Genomic structural variations (SVs).....	6
4.1.1.1 SVs contribute to phenotypic variation and local adaptation.....	7
4.1.1.2 Methods for identifying genomic variation underlying local adaptation .....	9
4.1.1.3 SVs facilitate speciation.....	10
4.1.1.4 SVs can evolve into supergenes.....	12
4.1.2 Means of SV discovery and genotyping.....	13
4.1.2.1 Advances in sequencing technologies increase our ability to construct high-quality genome assemblies and detect SVs.....	14
4.1.2.2 Applications of genome graphs .....	15
4.1.3 Long-reads provide a paradigm shift for constructing genome assemblies.....	16
4.1.4 Salmonids .....	16
4.1.4.1 Atlantic salmon ( <i>Salmo salar</i> ) .....	18
4.1.4.2 Lake whitefish ( <i>Coregonus clupeaformis</i> ) .....	20
4.1.5 Aims and objectives.....	21
4.2 Brief paper summaries .....	22
4.2.1 Paper I.....	22
4.2.2 Paper II .....	23
4.2.3 Paper III.....	24
4.3 Discussion and future perspectives.....	25
<b>5 References.....</b>	<b>29</b>

# 1 List of papers

## Paper I

Kristina Stenløkk, Michel Moser, Øystein Monsen, Anna Sofie Kjelstrup, Mariann Árnýasi, Torfinn Nome, Simen Sandve, Matthew Kent, Nicola Barson. Sigbjørn Lien. **Atlantic salmon pan-genome reveals hidden genomic variation impacting environmental adaptation.** Manuscript.

## Paper II

Kristina Stenløkk, Marie Saitou, Live Rud-Johansen, Torfinn Nome, Michel Moser, Mariann Árnýasi, Matthew Kent, Nicola Jane Barson and Sigbjørn Lien (2022). **The emergence of supergenes from inversions in Atlantic salmon.** *Philosophical Transactions of the Royal Society B*, 377(1856):20210195. doi: [10.1098/rstb.2021.0195](https://doi.org/10.1098/rstb.2021.0195)

## Paper III

Claire Mérot, Kristina Stenløkk, Clare Venney, Martin Laporte, Michel Moser, Eric Normandeau, Mariann Árnýasi, Matthew Kent, Clément Rougeux, Jullien M. Flynn, Sigbjørn Lien & Louis Bernatchez (2022). **Genome assembly, structural variants, and genetic differentiation between lake whitefish young species pairs (*Coregonus* sp.) with long and short reads.** *Molecular Ecology*. doi: [10.1111/mec.16468](https://doi.org/10.1111/mec.16468)

## 2 Abstract

Structural variations (SVs), e.g. deletions, insertions, inversions and duplications of sequences, are a major source of genomic variation affecting more base pairs in the genome than single nucleotide polymorphisms (SNPs). Despite their increasingly recognised importance in adaptive evolution and species diversification, SVs are vastly understudied in most species. Long-read sequencing, together with recently developed bioinformatic tools, have provided step-change improvements in the precision and recall of SV detection and allow us to increase the detected SVs manyfold across the species range. In addition, long-reads represent a major shift in our ability to build continuous genome assemblies as fundamental resources for most genome wide studies. The work in this thesis utilises long-read data to generate multiple genome sequences for the two salmonid species Atlantic salmon (*Salmo salar*) and lake whitefish (*Coregonus clupeaformis*).

We present the first pan-genome for Atlantic salmon, comprising 11 long-read-based assemblies across the species range. Among these, the highest quality genome has 2.55 Gbp assembled into chromosome sequences, 259 Mbp more sequence than in the previous Atlantic salmon reference genome. The genome has a highly improved continuity with contig N50 increasing from 58 kbp to 28.06 Mbp (484-fold). The detection of SVs in these 11 individuals, revealed 1,061,452 SVs, with an average of ~77.4 Mbp of sequence differing per sample. The Atlantic salmon has adapted to different river environment across a large geographical distribution. To investigate genomic variation underlying these adaptations, we associated SVs and environmental data in a dataset of 366 short-read samples genotyped using genome graph analyses. These analyses highlighted multiple SVs contributing to environmental adaptations, including an 18 kbp deletion encompassing a polymorphic segmental duplication of three genes associated with annual precipitation.

Next, we use the Atlantic salmon pan-genome to study the emergence of supergenes. Because supergenes can be maintained over millions of years by balancing selection and typically exhibit strong recombination suppression, their underlying functional variants and how they are formed are largely unknown. Inversions are type of rearrangement commonly associated with supergenes, and by directly comparing

multiple highly continuous genome assemblies we were able to detect a number of large inversions in Atlantic salmon. A 3 Mb inversion, estimated to be ~15,000-year-old, and segregating in North American populations, displayed supergene signatures with adaptive variation captured within the standard arrangement of the inversion, as well as other adaptive variation accumulating after the inversion occurred. Characterization of other inversions with matched repeat structures at the breakpoints did not show any supergene signatures, suggesting that shared breakpoint repeats may obstruct the supergene formation.

Lastly, we created long-read based genome assemblies for sympatric species pairs (Dwarf and Normal) belonging to lake whitefish (*Coregonus clupeaformis*). The species pairs offer a suitable model system for studying genomic patterns of differentiation and in particular the role of SVs in speciation. By combining long-reads, direct assembly, and short-read methods we detect 89,909 high-confidence SVs in the species pair across two lakes, covering five times more sequence in the genome compared to SNPs. In the study, we highlight shared outliers of differentiation between the lakes, indicating that they contribute to speciation. Interestingly, we find that more than 70% of SVs differentiating between the Normal and Dwarf species pairs of lake whitefish are overlapping transposable elements. This work demonstrates that SVs may play an important role for the differentiation and speciation of sympatric species pairs in lake whitefish.

### 3 Norsk sammendrag

Strukturell variasjon (SVer), for eksempel delesjoner, insersjoner, inversjoner og duplikasjoner av sekvens, er en viktig kilde til genomisk variasjon som samlet sett påvirker flere basepar i genomet enn punktmutasjoner (SNPs). Til tross for en økende annerkjennelse for at SVer spiller en viktig rolle i genetisk tilpassing til ulikt miljø og artsdannelse har denne typen variasjon vært lite studert i mange arter. Ny DNA-sekvenseringsteknologi med lengre leselengder (long-read sequencing), samt utvikling av nye bioinformatiske verktøy, har ført til drastiske forbedringer i deteksjonen av SVer. 'Long-read' sekvensering gjør det også mulig å lage mer komplette og sammenhengende genomsekvenser enn tidligere. I denne avhandlingen benytter vi oss av 'long-read' data til å lage flere genomsekvenser av høy kvalitet for to ulike laksefiskarter: Atlanterhavslaks (*Salmo salar*) og en Nordamerikansk type sik 'lake whitefish' (*Coregonus clupeaformis*).

Her rapporterer vi det første pan-genomet for Atlanterhavslaks. Det består av 11 assembler basert på 'long-read' sekvensering av individer fra fire ulike fylogeografiske grupper av villaks. Assembleret av høyest kvalitet inkluderer 2,55 Gbp sekvens i kromosomer, 259 Mbp mer enn det forrige referansegnetomet til Atlanterhavslaks. I tillegg ble andelen sammenhengende sekvens, målt som contig N50, økt fra 58 kbp til 28,06 Mbp (484 ganger høyere).

Vi fant 1.061.452 SVer på tvers av de 11 individene med ~77,4 Mbp gjennomsnittlig sekvensforskjell per prøve. Atlanterhavslaksen har over tid tilpasset miljøet i ulike elver. For å studere underliggende genetisk variasjon for denne tilpasningen assosierte vi SVer med ulike miljøvariabler i et datasett bestående av 366 'short-read' sekvenserte prøver ved bruk av en genom-graf. Ved hjelp av disse analysene fant vi flere SVer som bidrar til miljøtilpasning, blant annet en 18 kbp lang delesjon som inneholder tre gener assosiert med mengden nedbør i området.

Vi brukte så pan-genomet for Atlanterhavslaks til å studere dannelsen av 'supergener'. Supergener er en sammenkobling av genetisk variasjon i koblingsulikevekt som for eksempel kan oppstå ved hjelp av store inversjoner. Her utnyttet vi 11 genomassembler til å identifisere og karakterisere en rekke store inversjoner i Atlanterhavslaks. En av inversjonene på 3 Mbp, estimert til å være ~15.000 år

gammel, viste signaturer for utvikling som supergen. For de andre inversjonene som var flankert av repetert DNA fant vi ikke karakteristiske trekk på supergener, noe som tyder på at det repetitive DNA forhindrer en dannelse av supergener.

Til slutt lagde vi genomsekvenser for ulike former ('Normal' og 'Dwarf') av 'lake whitefish' (*Coregonus clupeaformis*) som lever i de samme innsjøene i Nord-Amerika. Genomsekvensene muliggjør studier av genomiske mekanismene bak artsdannelse i denne laksefisken. Ved å kombinere 'long-read' data, direkte sammenlikning av assembler, og 'short-read' data fant vi 89,909 SVer som skilte de to formene av 'lake whitefish' i to innsjøer. SVene omfatter mer enn fem ganger flere basepar i genomet sammenlignet med SNPs. I studiet fant vi flere SVer med avvikende forekomst ('outliers') i de to formene av 'lake whitefish', noe som indikerer at disse SVene bidrar til artsdannelse. Videre fant vi at 70 % av SVene overlappet en form av repetert DNA kalt transposable elementer. Dette arbeidet understreker at SVer kan spille en viktig rolle for artsdannelse i 'lake whitefish'.

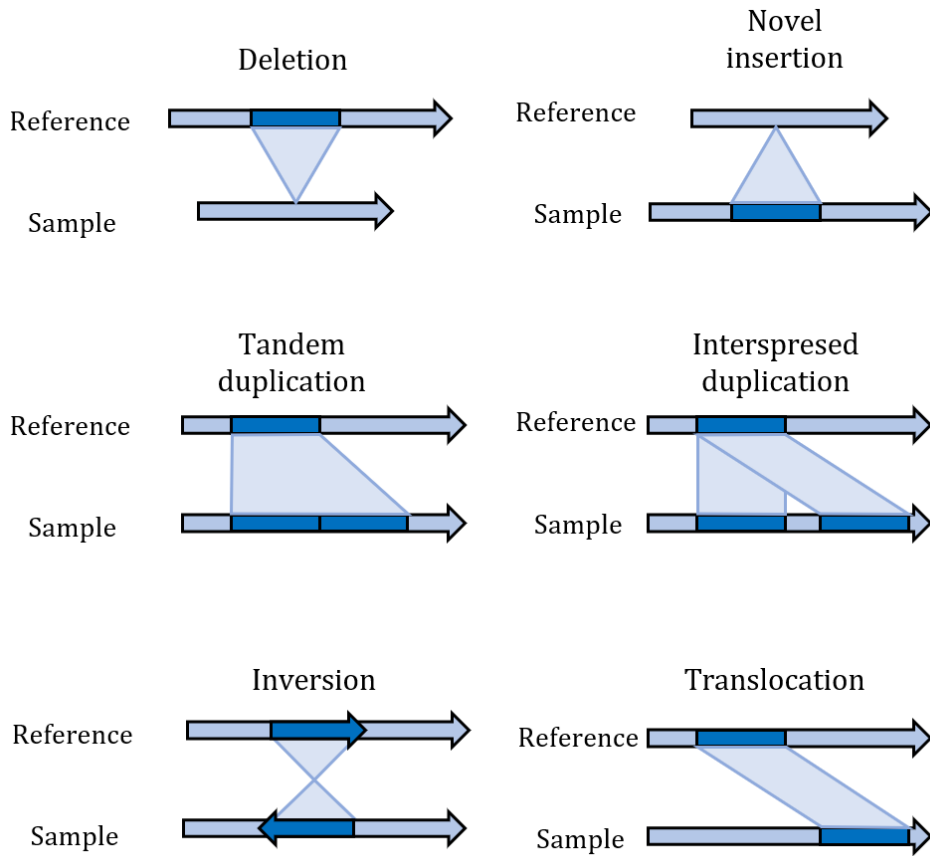
# 4 Synopsis

## 4.1 Introduction

### 4.1.1 Genomic structural variations (SVs)

Over the past decade the most widely studied type of genomic variation has been single nucleotide polymorphisms (SNPs) (Jiang et al. 2016; Mérot et al. 2020). SNPs are efficiently detected by generating short-read data from multiple individuals and mapping them to a single reference genome. This approach has worked well across species, and today SNP-data is extensively used in genome-wide studies to investigate their importance in a multitude of disciplines, including population genomics (Zimmerman et al. 2020), genetic diseases (Li et al. 2015) and genomic selection (Reshma & Das 2021). Although this approach has resulted in striking advances, SNPs are only one form of genomic variation and are shown to affect fewer base pairs in the genome compared to larger genomic variations in multiple species (Catanach et al. 2019; Hämälä et al. 2021; Pang et al. 2010).

Structural variations (SVs) are differences in the genome somewhat arbitrary defined as being larger than or equal to 50 bp (Mahmoud et al. 2019). The most common types of SVs are deletions and insertions. Insertions can be novel or a duplication of an already existing sequence and is then referred to as a duplication. Duplications can be copied in from another location in the genome (interspersed duplication) or duplicated in tandem. The sequence can also be inverted or moved to another location (translocation), see Figure 1. Though most SVs are short (50-100 bp), they may also cover mega bases of sequence.



**Figure 1:** Schematic representation of common classes of SVs.

#### 4.1.1.1 SVs contribute to phenotypic variation and local adaptation

SVs can have functional impact in many ways that might lead to phenotypic changes in an organism. This includes direct altering of protein coding sequence, introducing changes in regulatory elements, which affect gene expression, or modifying gene dosage (Mahmoud et al. 2019). Recent findings in humans suggests that SVs have larger effects compared to SNPs (Chiang et al. 2017; Hsieh et al. 2019).

Several SV studies has been conducted in economically important agriculture species, including tomato (Alonge et al. 2020), cattle (Low et al. 2020) and soybean (Liu et al. 2020). These studies reveal huge contributions of genetic variation from SVs and find links to many traits of interest for animal breeding and crop improvement, including



disease resistance, fruit flavour and fruit size. By studying SVs in wild populations, we can acquire knowledge of how SVs contribute to natural selection in the processes of local adaptation and speciation.

Genomic variation can give rise to local adaptation which occurs when some individuals have higher fitness in their local environment compared to others (Savolainen et al. 2013). Most of our knowledge concerning the genomic mechanisms underlying local adaptation comes from SNPs or microsatellites, but SVs has also been shown to play an important role (Mérot et al. 2020). Notably, a recent study on how SVs facilitates adaptation in wild chocolate trees predicts that most SVs are detrimental (Hämälä et al. 2021). Despite the potential functional consequences of SVs, results documenting these are still rare because to date most studies are based on short-read sequencing data with limited ability to resolve SVs within or close to repeats (Sudmant et al. 2015). Old inversions that have developed strong linkage disequilibrium (LD) have been subject to many studies because they can be detected through indirect methods, while SVs caused by repeats are less represented as they are more difficult to resolve. In fact, most studies investigating adaptive SVs focus on one type or few classes of SVs and more work is needed to gain a better understanding of the contribution of the whole range of size and types of SVs on local adaptation (Mérot et al. 2020).

An emerging number of studies are considering a wide range of SVs in the context of local adaptation. For example Faria et al. (2019) studied two different ecotypes of the snail *Littorina saxatilis* and found variable frequencies of 17 polymorphic inversions between two microhabitats despite gene flow, suggesting that these SVs are involved with local adaptation. Another example is the study by Hämälä et al. (2021) detecting more than hundred SVs bearing signatures of local adaptation in the chocolate tree (*Theobroma cacao*), of which several were associated with genes differentially expressed between populations. This study also finds that SVs can contribute to local adaptation by sheltering locally beneficial alleles from gene flow by preventing or reducing the formation of viable crossovers within chromosomal heterozygotes.

#### **4.1.1.2 Methods for identifying genomic variation underlying local adaptation**

Scanning the genome for variants underlying local adaptation has become a widely used approach in evolutionary and ecological studies (Hoban et al. 2016). Estimating associations between genomic variants (such as SNPs and SVs) and environmental variables, called genotype-environment associations (GEA), is a particularly promising way of detecting adaptive variants (Rellstab et al. 2015). Selecting a suitable sampling design is crucial for the subsequent analysis. Some common designs include sampling along environmental gradients, categorical sampling (for example low vs. high temperature) or aiming to sample as broadly within the species environmental niche as possible (Rellstab et al. 2015). Another sampling design that has shown to be especially useful for detecting weak selection, as may occur on polygenic trait variation, is to sample random pairs of closely located populations with distinct differences in environmental conditions (for example two closely located rivers of drastically different size) (Lotterhos & Whitlock 2015). It is also important to consider how the underlying population structure is accounted for as this might mask the true adaptation to environment with neutral genomic variability between locations stemming from factors such as drift, creating false positive associations. Many different statistical approaches for GEA have been developed with no single method dominating in the field, as different use cases might require different methods. This includes methods for testing categorical factors, logistic regression, matrix correlations, linear and mixed effect models (Rellstab et al. 2015), but for the purpose of this thesis we will focus on the mixed effect model methods as these are powerful methods providing a unified statistical framework for controlling the effects of population structure.

The term mixed effect refers to the inclusion of both fixed (affects the response variable in a non-random manner) and random factors (affects the response variable in a random manner) in the models. For the purpose of GEA, these statistical models treat allele frequencies as the response variable, environmental variables as the fixed factor and the neutral genetic structure as the random factor (Rellstab et al. 2015). One commonly used mixed model GEA tool is BAYENV (Coop et al. 2010), which has been shown through simulations to have a low false positive rate (De Mita et al. 2013). One potential limitation of this tool is the dependency on a good estimate of population allele frequency, which depends on the number of individuals sampled within each genetic population. Latent factor mixed model (LFMM) (Frichot et al.

2013) does not require an estimate of population frequency as it is individual based, which can be advantageous where samples are distributed across the environment rather than clustered into local populations. LFMM estimates the neutral genetic structure (random factors) as latent factors, which is a method of reducing the data dimensionality (similar to PCA) of the neutral genetic structure. Another advantage of this method is that the effects of environmental factors and neutral genetic structure are simultaneously estimated, reducing the impact of selected loci on the estimation of population structure. However, this method requires the number of latent factors as input, which equals the expected number of genetic populations ( $K$ ) in the data and this, therefore, needs to be estimated prior to the analysis.

#### **4.1.1.3 SVs facilitate speciation**

One field of research that is central to evolutionary biology is the emergence of new species through the process of speciation (Weissing et al. 2011). An essential element in understanding speciation is identifying the genetic basis of reproductive isolation, which can arise from various genetic changes (Zhang et al. 2021). Often, speciation involves multiple genes and understanding the mutations involved is key for understanding the speciation process. SVs may be a particularly important type of mutation for speciation, in that many genes can be affected simultaneously. This idea has been strengthened by multiple recent studies made possible with advances in detection and genotyping of SVs, for example in deer mouse (Hager et al. 2022), songbirds (Weissensteiner et al. 2020) and killifish (Berdan et al. 2021b). Though the signatures of differentiation are the result of complex interactions between gene flow, recombination, demography and selection, there are some proposed models for the impacts of SVs on speciation (Zhang et al. 2021). We will briefly discuss some of these models.

The hybrid-sterility model suggests that heterokaryotypes of an SV will have reduced fitness compared to the homokaryotypes and was first suggested by (Wright 1978). This is because the SV might cause mispairing during meiosis and hence produce non-functional gametes (Homolka et al. 2007). Another model, the suppressed-recombination model, suggests that instead of directly reducing the fitness of hybrids, SVs might promote reproductive isolation through suppressed recombination (Navarro & Barton 2003; Noor et al. 2001; Rieseberg 2001), reviewed by Zhang et al. (2021). For example, an inversion can limit recombination among sets of alleles

related to local adaptation and reproductive isolation (Zhang et al. 2021). Rieseberg (2001) suggested that SVs suppressing recombination could have an increased effect in combination with genes located in the SV breakpoints. A third model specifically tackles gene duplications and suggests that these SVs can cause postzygotic isolation, which was first described by Haldane (1933). This can either happen through loss of function in one copy or sub-functionalisation of the copies between different species.

In addition to these three suggested models for how SVs might contribute to reproductive isolation, SVs can cause large phenotypic effects in more direct ways. For example, by deletion of multiple genes linked to a trait or an inversion disrupting the reading frame. Another interesting case is when a transposable element affects the expression of nearby genes by insertion of a promoter or other regulatory region in the genome. For example, in songbirds where a gene affecting premating isolation was found to be downregulated by insertion of a LTR retrotransposon (Weissensteiner et al. 2020).

To detect regions in the genome associated with speciation we can estimate the genetic differentiation between populations. The most common measure of differentiation is the fixation index ( $F_{ST}$ ), which is among the most widely used descriptive statistics in population and evolutionary genetics (Holsinger & Weir 2009).  $F_{ST}$  measures the variance of allele frequencies between populations. This is often estimated as the difference in the average number of pairwise differences (nucleotide diversity) between two individuals sampled from different populations ( $\pi_{between}$ ) and the same population ( $\pi_{within}$ ) divided by the average number of pairwise differences between ( $\pi_{between}$ ) two individuals sampled from different populations (Hudson et al. 1992).

$$F_{ST} = \frac{\pi_{between} - \pi_{within}}{\pi_{between}}$$

This means that a low  $F_{ST}$  value signifies that the allele frequencies within each population are similar and a large  $F_{ST}$  value signifies that the allele frequencies are different. If one allele is favoured over the other at a locus in some populations, the  $F_{ST}$  will be higher than the genome-wide average and we can, therefore, use  $F_{ST}$  to detect loci under positive selection. Though  $F_{ST}$  is the most widely applied population statistic, there exists many other measures that can be used.

#### 4.1.1.4 SVs can evolve into supergenes

An interesting evolutionary trajectory some SVs follow is evolution into so called ‘supergenes’. This happens when alternative phenotypes in balanced polymorphisms segregate as if controlled by a single locus because of tight linkage among multiple functional loci (Thompson & Jiggins 2014). Supergenes can arise from SVs effectively blocking recombination, for example, between the two orientations of an inversion (Gutiérrez-Valencia et al. 2021). Over time, these alternative variants can accumulate linkage disequilibrium (LD), and thus evolve into a segment behaving as a single non-recombining unit. Supergenes have been an area of interest to scientist for many decades, especially for local adaptation and speciation (Gutiérrez-Valencia et al. 2021; Thompson & Jiggins 2014). Some of the best-studied supergene systems stem from inversions, including polymorphic wing mimicry in butterflies (Clarke et al. 1968), and different ecotypes of the yellow monkeyflower *Mimulus guttatus* (Hall et al. 2006). Multiple inversion supergenes have been studied in fish species, such as a supergene found in cod (*Gadus morhua*) associated with migration phenotype (Kirubakaran et al. 2016; Kirubakaran et al. 2020; Matschiner et al. 2022) and Atlantic herring (*Clupea harengus*) relating to local adaptation to water temperatures (Pettersson et al. 2019). Another example is that of a large (55 Mbp) double inversion supergene studied by Pearse et al. (2019) in the salmonid rainbow trout (*Oncorhynchus mykiss*) mediating a sex specific migratory tendency with evidence of environment dependent selection.

Despite many examples of inversion supergenes, questions remain about their emergence, which can be efficiently detected by *de novo* assembly comparisons (see section 4.1.2.1). In addition, existing assemblies might have restricted genomic resolution in repeat regions which are often associated with inversion breakpoints (Gutiérrez-Valencia et al. 2021; Pettersson et al. 2019). In particular, young inversions will likely not yet have developed strong LD, which makes them harder to detect with SNP-markers alone. Older supergenes are likely to have lost many signatures important for the emergence of the supergene, through fixation of adaptive variation, and gained neutral or deleterious variation because of reduced efficiency of purifying selection resulting from reduced effective population size (Berdan et al. 2021a). These processes make differentiating adaptive variation that was important for supergene formation and trait linked variation challenging (Jay et al. 2018) There is still little empirical evidence regarding the mechanisms during formation of supergenes (Charlesworth & Barton 2018), and one question related to

this is whether the adaptive loci are captured during the supergene formation or gained through mutation subsequently (Berdan et al. 2022).

To investigate capture vs. gain of adaptive loci following the initial inversion event, we can study young inversions. This is because adaptive loci likely will become fixed over time within inversion orientations, and the two cases will become indistinguishable. However, by observing the inversion soon after the inversion event a case of capture would be characterised by polymorphic adaptive alleles in the ancestral arrangement, while only containing one of the alleles in the inverted arrangement. On the other hand, if we observe a case of gain, the adaptive alleles will only be polymorphic in the inverted arrangement, while fixed in the ancestral, as this variation has been derived through mutation after the inversion event. Recently simulations have predicted the importance of both gain and capture of adaptive alleles for supergene formation (Schaal et al. 2022), but empirical evidence for either has been lacking.

#### **4.1.2 Means of SV discovery and genotyping**

Many recent studies are reporting impacts of SVs in adaptive evolution and species diversification (e.g. (Faria et al. 2019; Hämälä et al. 2021; Tong et al. 2022; Weissensteiner et al. 2020), but the full range of SVs remains understudied in most species (Mahmoud et al. 2019). In theory, detecting genomic differences between sequences sound trivial, but in practice the detection can be challenging. The detection is complicated by sequencing- and mapping errors (Mahmoud et al. 2019), especially when reads used for detection are shorter than the SV (Sedlazeck et al. 2018). Some types of SVs can be difficult to differentiate, for example, distinguishing a novel insertion from a tandem duplication. Further complications appear with nested SVs, which might be impossible to resolve with read mapping and may require *de novo* genome assembly comparisons to be determined. Fortunately, the introduction of new long-read sequencing technology and improved computational methods are enhancing our ability to detect and genotype SVs. The most used approaches for SV detection and genotyping will be discussed in the following sections.

#### **4.1.2.1 Advances in sequencing technologies increase our ability to construct high-quality genome assemblies and detect SVs**

In the beginning of the 2000's the National Human Genome Research Institute (NHGRI) initiated a program to bring the cost of whole-genome sequencing down to US\$1000 in 10 years (Schloss 2008). At that time, the dominating sequencing technology was the expensive and labour-intensive Sanger sequencing. The '\$1000 genome' initiative led to the development of several more cost effective and less time-consuming technologies, referred to as Next Generation Sequencing (NGS), second generation sequencing or short-read sequencing. A major benefit of this technology, besides reduced costs, time, and labour, is the high sequencing accuracy (van Dijk et al. 2018). However, a major limitation is the limited read lengths, often in the range 100-300 bp. For appliances like detection and genotyping of SNPs and short indels these read lengths are sufficient, but for other applications, like assembly and read mapping in repeat rich regions of the genome, they turn out to be less efficient. Consequently, many genome assemblies constructed based on short-reads are fragmented. To build complete genome assemblies, longer reads spanning the repeat rich regions is needed (see below). When it comes to SV detection, short-reads are often not long enough to cover the entire length of the SVs (Mahmoud et al. 2019) making it harder to resolve the variation (Smolka et al. 2015). So far, no short-read based bioinformatic tool has been able to detect all SV-types and sizes reliably (Mahmoud et al. 2019; Sedlazeck et al. 2018), with mid to large size insertions being particularly challenging to identify. Overall, the recall of short-read based SV detection tools has been found to be in the range 10% to 70%, while the false positive rate is reported to be as high as 89 % (Mahmoud et al., 2019).

A few years after the introduction of short-read sequencing, a new generation of sequencing technology emerged called third-generation sequencing or long-read sequencing. These technologies based on single-molecule sequencing (van Dijk et al. 2018) are not bound to a set read length and produce much longer reads (average read length often >10kbp) than short-read sequencing (100-300 bp). The first commercially available long-read technology, was Pacific Biosciences (PacBio) released in 2011, followed by Oxford Nanopore Technologies (ONT) announced in 2014. Although these technologies are relatively costly and possess error-rates exceeding that of short-reads, they are proven particularly useful for improving contiguity of genome assemblies. Especially for duplicated and repeat rich genomes, the introduction of long-reads has been revolutionary (Huddleston et al. 2014). Long-

read sequencing was awarded “method of the year” in 2022 by Nature (Method of the Year 2022: long-read sequencing 2023). Our ability to detect different types of SVs has also changed substantially with the development of long-read sequencing technology, achieving high sensitivity and specificity by spanning both SVs and their flanking sequences (Dierckxsens et al. 2021; Mahmoud et al. 2019; Sedlazeck et al. 2018). Depending on the organism investigated, the number of SV detected using long-reads are reported to be in the range of 2 to 8.33 times higher than found with short-read data (reviewed by Mahmoud et al. (2019)).

Another strategy for SV-detection is through direct comparison of *de novo* assembled genomes. Such alignments have the potential to detect any form of structural variations and have been especially useful for detecting longer insertions and inversions (Mahmoud et al. 2019). An advantage of direct *de novo* assembly comparisons is that the approach does not suffer from reference sequence bias in which reads containing non-reference alleles are less likely to be mapped than those containing reference alleles (Nattestad & Schatz 2016; Tian et al. 2018). A major drawback connected with the approach is that assemblies are costly to produce which typically limits their applicability.

#### **4.1.2.2 Applications of genome graphs**

Although long-reads are superior for SV-detection, they remain prohibitively expensive for population-scale SV genotyping. One way to produce population scale SV datasets could be to split the discovery and genotyping steps (Huddleston et al. 2017) and use long-reads for the accurate SV-detection and short-reads for large scale genotyping. Such a combined approach has been made possible by the development of genome graph-based methodology (Chen et al. 2019; Eggertsson et al. 2019; Garrison et al. 2018). The goal of genome graphs is to represent the complete genome, including all genomic variation, of a given species or clade (Eizenga et al. 2020). This is often referred to as the pan-genome as an alternative to the linear reference genome. The number bioinformatics tools developed for genome graph construction has expanded over the last 5-10 years (Armstrong et al. 2020; Chen et al. 2019; Eggertsson et al. 2019; Garrison et al. 2018; Li et al. 2020). Normally the methods can be grouped into two categories; (i) graphs based on a linear reference genome with added variation, and (ii) alignment-based graphs. Generally, the tools using the former category will combine a genome assembly with a VCF-file containing known variants to build a directed acyclic graph (DAG) (Eizenga et al. 2020). Thus far,



the most popular genome graph methods for genotyping of SVs have been variation graph (vg) (Garrison et al. 2018), GraphTyper2 (Eggertsson et al. 2019) and Paragraph (Chen et al. 2019).

By using assembly-based graphs, we can theoretically genotype the full scope of SVs. These tools are still under active development, and appears somewhat in the developing stage, though some examples of use are emerging. For example, Crysanto et al. (2021) used the tool Minigraph (Li et al. 2020) to create a pan-genome of cattle (*Bos taurus*), detecting ~68 k SVs. Another example is the human pan-genome reported by Liao et al. (2022) increasing the number of SVs per haplotype by 104% compared to the linear reference. The work is part of the Human Pangenome Reference Consortium initiative, aiming at creating a complete human pan-genome with telomere-to-telomere representation of global genomic diversity (Wang et al. 2022). Assembly-based graph is still under active development but remains challenging to implement for large and complex genomes with high repeat content (Wang et al. 2022).

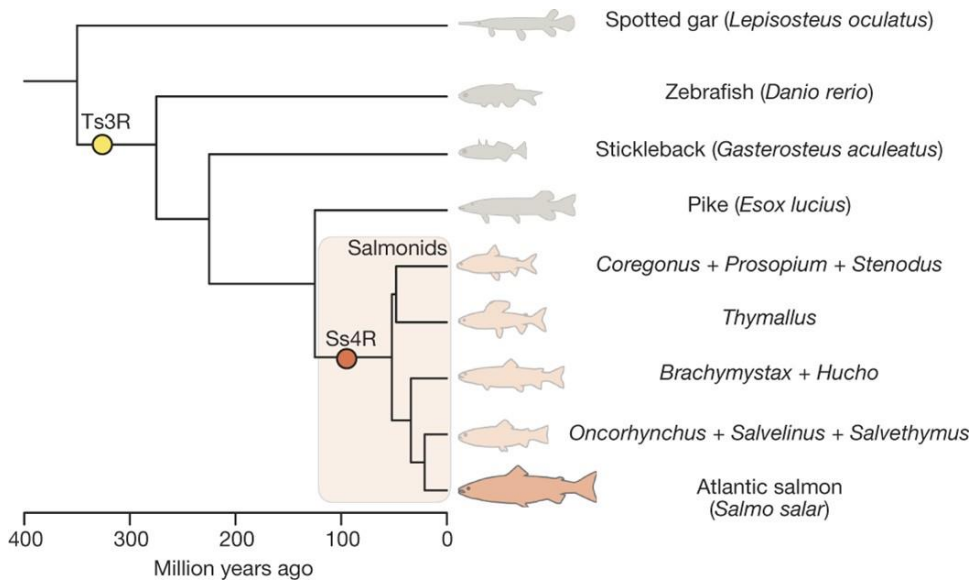
### **4.1.3 Long-reads provide a paradigm shift for constructing genome assemblies**

The detection and genotyping of SVs are highly dependent of the completeness and quality of genome assemblies, which may be particularly challenging to obtain in repeat-rich and duplicated genomes. For the last two decades, short-read based approaches have been dominating the field of *de novo* genome assembly construction (Hotaling et al. 2021). However, short-reads have limited ability to read through repeats which may lead to fragmented assemblies. This is particularly concerning when constructing assemblies from species with large and duplicated genomes. Long-reads offers a paradigm shift for *de novo* genome assemblies by vastly increasing both sequence continuity and completeness (Sohn & Nam 2018; Su et al. 2021). Commonly used pipelines to build genome assemblies with long-reads often include correction of the sequence with short-reads to further increase sequence quality though a process referred to as ‘polishing’ (Sohn & Nam 2018).

### **4.1.4 Salmonids**

The salmonids are a group of ray finned fish encompassing three subfamilies: Coregoninae, Thymallinae and Salmoninae with approximately 70 species (Nelson et

al. 2006). The salmonids include the genus *Salmo* (where we find Atlantic salmon), *Oncorhynchus* (trouts), *Salvelinus* (chars), *Coregonus* (whitefishes), *Prosopium* (lake whitefish, but not the species lake whitefish), *Thymallus* (graylings), *Hucho* (taimens) and *Brachymystax* (lenoks), as shown by the phylogeny in Figure 2. Salmonids are one of the most important and influential fish species in the northern hemisphere (Crawford & Muir 2008; Johnsson & Näslund 2018), and have been subject to numerous genetic studies (Houston & Macqueen 2019; Houston et al. 2020).



**Figure 2:** Phylogenetic relationship of salmonids and selected teleost lineages. Originally published in Lien et al. (2016).

The salmonid ancestor went through a whole genome duplication (Ss4R) 89–125 million years ago (MYA) (Gundappa et al. 2022), following an earlier WGD (300–350 MYA) in the teleost common ancestor (Lien et al. 2016). The ancestor is thought to have a diploid chromosome number (2N) of 48–50 (Mank & Avise 2006; Phillips & Rab 2001), similar to what the closest relatives Esocidae (e.g. pike (*Esox Lucius*)) has today: 50 chromosomes (2N) (Phillips et al. 2009). Through the WGD, the chromosome number was doubled to 96–100. Today, the salmonids tend to fall into one of two karyotypic categories: A) The diploid chromosome number (2N) being close to 80 with approximately 100 chromosome arms or B) the diploid chromosome number being close to 60 (2N) with approximately 100 chromosome arms (Phillips & Rab 2001). In the Coregoninae and Salmoninae subfamilies, the chromosomes have evolved through Robertsonian fusions, fusing chromosomes together and

subsequently reducing the chromosome number (Phillips & Rab 2001). The group of salmonid fishes is noted for having a considerable phenotypic plasticity thought to stem from the whole genome duplication. Whole genome duplication events are particularly prone to favour rapid diversification (Landis et al. 2018) and with an extra set of genes, some might evolve to attain new functions (Ohno 1970). For the work in this thesis, we are focusing two salmonid species: the Atlantic salmon (*Salmo salar*) and lake whitefish (*Coregonus clupeaformis*).

#### **4.1.4.1 Atlantic salmon (*Salmo salar*)**

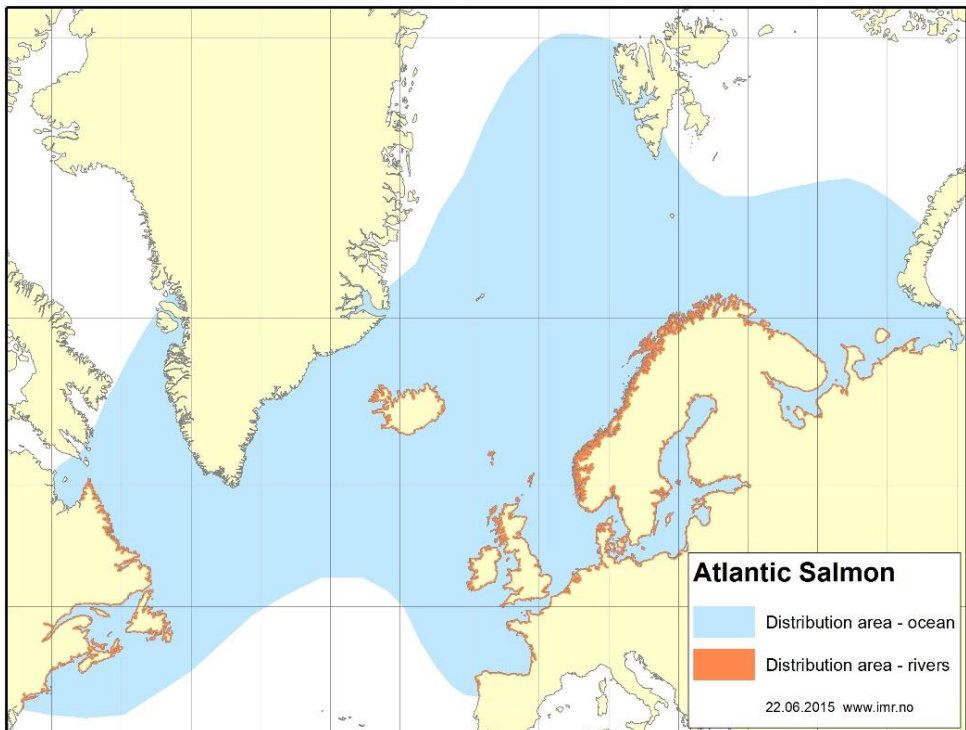
The latest report from the UN Food and Aquaculture Organization FAO (FAO 2022), states that the Atlantic salmon is among the most important species farmed in marine environment, with a global production of 2.71 million tons in 2020 or 32.6 % of all fish in aquaculture. Further, the Atlantic salmon is an important species for fisheries but has over the recent decades seen a significantly decline of wild populations and has therefore been subject to efforts of conservation and management (Verspoor et al. 2007).

The Atlantic salmon is profoundly anadromous, although there exist both land-locked populations and other systems where the fish complete their life cycle without marine migration (Verspoor et al. 2007). Anadromy refers to a life-cycle strategy where the fish is born in freshwater, mature in the ocean, and later returns to spawn (mate) in freshwater. The early life stages are in freshwater, including eggs, alevins, fry and parr, and transpires for a variable number of years (1-6) (Verspoor et al. 2007). The salmon then undergoes a series of physiological and morphological changes, called smoltification, to adapt to the marine environment.

The marine phase of the lifecycle is characterised by periods of rapid growth and eventually sexual maturation before returning to freshwater to spawn. During this period, lasting between 1-5 years, the salmon can grow up to be as large as 32 kg. Mortality during the marine phase can be as high as 70-99% depending on the geographical location (Verspoor et al. 2007). The surviving adult fish will then return to freshwater to spawn, which interestingly tends to be the same location as they were born. This means that the river populations are reproductively isolated and can over time differentiate to form locally adapted populations. There is considerable variability between rivers regarding the number of reproducing individuals and environmental factors affecting this adaptation, but overall, the Atlantic salmon show

a lot of variability regarding local adaptation dependent on habitat heterogeneity (Verspoor et al. 2007).

The geographical distribution of the Atlantic salmon is the North Atlantic and associated coastal drainages (see Figure 3). The greatest genetic divergence is found between North America and Europe, with an estimated divergence time of at least 600,000 years (King et al. 2007). Within Europe, three distinct phylogeographic groups are found: Atlantic, Barents/White Sea and Baltic Sea, with Baltic having a lower measured genetic diversity than the two other groups (Bourret et al. 2013). These three groups are consistent with main postglacial colonization routes (Tonteri et al. 2007; Tonteri et al. 2009). There is also significant genetic differentiation between anadromous and landlocked populations (Verspoor et al. 2007).



**Figure 3:** Geographical range of the Atlantic salmon. Figure downloaded from the institute of marine research (<https://www.hi.no/en/hi/temasider/species/salmon-atlantic>, 27.11.22).

#### 4.1.4.2 Lake whitefish (*Coregonus clupeaformis*)

The *Coregonus* genus is the most speciose within the family Salmonidae and fish within this genus are considered an attractive study system for adaptive radiation, fast speciation, and species reversal (Frei et al. 2022; Lundsgaard-Hansen et al. 2013). For example, De-Kayne et al. (2022) report that the Alpine whitefish contains more than 30 species in Swiss lakes adapting to local environments after the last glacial maximum. The species complex includes multiple species parallelly adapting to the lacustrine water depth gradient in distinct forms, called eco-morphs, and has been widely used as a model system to study speciation (De-Kayne et al. 2022; Frei et al. 2022; Schluter 2000).

A similar species complex is found for lake whitefish throughout a number of lakes in North America. The most studied species complex includes two sympatric forms that are mostly reproductively isolated, referred to as *C. clupeaformis* sp. *Normal* and *C. clupeaformis* sp. *Dwarf* (Bernatchez et al. 2010). Although these two species are found in the same lakes, they occupy different niches concordant to the water depth gradient. In addition to having a differentiated morphology, including different body sizes as their names indicate, the Normal form is adapted to the benthic habitat (bottom of the body of water), while the Dwarf form occupy the limnetic zone (open water area where light is accessible), where it has evolved to make use of the planktonic trophic niche. There is limited gene flow between the forms, but hybrids do occur (Renaut et al. 2009). Based on demographic modelling and analysis of mitochondrial DNA it has been estimated that the two forms started divergence during the last glaciation, roughly 60,000 years ago and that secondary contact occurred approximately 12,000 years ago (Bernatchez & Dodson 1990; Jacobsen et al. 2012; Rougeux et al. 2017). Other data suggest that the Dwarf form has derived from the Normal form multiple times (Renaut et al. 2011). The lake whitefish species complex provides pairs of nascent and sympatric forms which is a suitable system for studying the genetic architecture of speciation, and more specifically for this thesis, the role of SVs in speciation.

#### **4.1.5 Aims and objectives**

The principal objective of this thesis is to utilize the benefits of long-read sequencing to create highly continuous genome assemblies and investigate the role of structural variations (SVs) in local adaptation and speciation in salmonid fishes. That work is presented in three chapters with the following sub-goals:

- 1) Establish pipelines for long-read based assemblies and SV detection and use these pipelines to provide pan-genomic resources providing an extensive description of SVs across phylogeographical groups in Atlantic salmon and sympatric species pairs in lake whitefish (paper I and III).
- 2) Expanding the SV dataset by using a hybrid approach of long-reads for SV discovery and short-reads for genotyping through genome graph methods (paper I and III).
- 3) Uncover structural variants associated with local adaptation on Atlantic salmon (paper I, II) and speciation in lake whitefish (paper III).

## 4.2 Brief paper summaries

### 4.2.1 Paper I

The Atlantic salmon has high economic importance and has been widely studied. Construction of the previous Atlantic salmon reference genome sequence (ICSASG\_v2) was based on short-reads, and resulted in a rather fragmented assembly, especially in regions with high sequence similarity and repeats. Also, previous SV detection in Atlantic salmon was based on the use of short-read data, which in other species has been shown to give be a reduced and biased representation towards short deletions. To make step-change improvement in the genomic resources for Atlantic salmon, we created a salmon pan-genome consisting of individual assemblies from 11 Atlantic salmon sampled from a wide phylogeographic range. The genome assemblies a great improvement in continuity and additional sequence anchored to chromosome sequence compared to ICSASG\_v2.

We developed a pipeline for SV detection using a consensus-approach, based on three independent SV calling software. We found a total of 1,061,452 SVs detected independently by multiple pipelines, on average affecting ~3 % of the genome per fish. A large proportion of SVs (632,193) were found in one sample only, reflecting that analyses were based on data from a limited number of individuals sampled from a broad phylogeographic distribution. The insertions and deletions overlapping coding sequence affect 2,725 genes, with a significant enrichment of SVs overlapping duplicated versus singleton genes, supporting the results found by (Bertolotti et al. 2020). This suggest that the functional redundancy of duplicated genes allows for the accumulation of deleterious SVs in duplicated genes. We found that TEs showed an overall depletion of sequences overlapping deletions (24.02%) and insertions (21.15%) compared to the genome wide TE-content (40.61%), but enriched in TR-sequence, which is also strongly correlated with distance to telomeres.

To expand our SV data, detected by long-reads into a population wide dataset, we genotyped 366 short-read sequenced individuals using genome graphs with the GraphTyper2 software. After filtering, the dataset consisted of 304,407 genotyped SVs which we associated with environmental variables to reveal SVs contributing to local adaptation. We found an 18 kbp deletion encompassing a segmental duplication of three genes associated with annual precipitation, in addition to enrichment to several enriched KEGG pathways including the GnRH signalling pathway known to be one of

the main regulators of reproductive function in vertebrates Our results demonstrate how long-reads and the use of genome-graphs can reveal previously hidden complex genetic variants and that these likely have consequences for fitness and adaptations in the wild.

## 4.2.2 Paper II

Supergenes link alleles into non-recombining units acting as a single locus known to play essential roles in maintaining adaptive genetic variation. However, their underlying mechanisms facilitating adaptations has been poorly characterised in most species, including Atlantic salmon. Further, there are unanswered questions about the emergence of supergenes, including whether the adaptive variants are captured or gained. Also, inversion breakpoints are poorly characterised due to repeat sequence, which has been difficult to resolve with short-reads alone.

By a combining our long-read mapping and direct assembly comparisons of the Atlantic salmon pan-genome, we identified 11 large (> 50 kbp) inversions. By inspecting the breakpoint sequences, we found that none of the inversions with repeat blocks in their breakpoints had developed haplotype structures, suggesting that these repeats might hinder the build-up of linkage disequilibrium and later supergene formation. Amongst the inversions with no obvious repeat structures, we found one large multigene inversion tagged by a haplotype across 482 SNPs matching the inversion genotype (chr18inv). We estimated this inversion to be approximately 15,000 years old, making this a young inversion compared to iconic supergenes reported for other species. By associating variation within the inversions to environmental variables, we found three inversions had associations to multiple environments, indicating adaptive variation. Chr18inv has three strongly differentiated variants (LFMM  $p < 0.05$  and  $F_{ST} > 0.8$ ) which provided evidence for both capture and gain of adaptive alleles. The upstream breakpoint of chr18inv hits a gene (*MRC2-like*) with two nearby copies possibly compensating for the eventual functional consequences of the gene disruption. Overall, our results suggests that multiple processes contribute to the formation of supergenes from inversions, that is both capture and gain of adaptive alleles and tolerated breakpoint mutations.



### 4.2.3 Paper III

The sympatric and nascent species pairs within lake whitefish, Normal and Dwarf, offers an attractive system for studying the genetic mechanisms of speciation. Especially, the role of SVs in speciation is largely unknown. Here, we combine short- and long-read sequencing to investigate a wide range of variants role in speciation between the species pair. We created the first reference genome assemblies for lake whitefish Normal and Dwarf, with high levels of completeness. By combining different software, including methods for long-read mapping, short-read mapping, and direct assembly comparison, we found 89,909 high-confidence SVs. In total, these SVs cover five times more base pairs than SNPs. These SVs were then used to genotype a set of 32 short-read sequenced fish using a genome graph. By investigating the genomic patterns of differentiation between Dwarf and Normal species pairs using both SVs and SNPs, we highlighted a large fraction of the differentiated SVs overlapped transposable elements (TEs), suggesting that TE- accumulation may represent a key component of genetic divergence between the lake whitefish Normal and Dwarf species. Our results suggest that SVs may play an important role in speciation and that, by combining short- and long-read sequencing, we now can integrate SVs into speciation genomics.

### 4.3 Discussion and future perspectives

The construction of reference genomes establishes the foundation for advanced genomics exploring how variation in the genome affect the evolution of species, populations, and how it translates to an individual's phenotype. The ultimate goal of genome sequencing is to produce error-free continuous sequences spanning entire chromosomes. Unfortunately, in many species this is not an easy endeavour because the genome of interest is large, duplicated and loaded with repeated DNA that are difficult assemble (Sohn & Nam 2018). Long-read sequencing technologies, which have potential to read through repetitive DNA, represent a game changer for the construction of continuous and complete genome sequences (Amarasinghe et al. 2020; van Dijk et al. 2018). In this thesis, we generated long-read nanopore data and build 13 high-quality *de novo* assemblies for two salmonid species; 11 assemblies for Atlantic salmon and two for lake whitefish. The lake whitefish genome assemblies are the first for *C. clupeaformis* and establishes the foundation for genome wide investigations targeting adaptive evolution and species diversification in this species. The 11 *de novo* assemblies constructed from all four phylogeographical lineages of Atlantic salmon represent the first pan-genomic resource reported for this species. The salmon pan-genome moves away from traditional reliance on a single linear reference genome towards a more comprehensive representation of the genomic diversity of Atlantic salmon.

Pan-genomes can be represented by genome graphs, where the relationship among the linear references is shown. As shown in other eukaryote species like humans (Liao et al. 2022) and cattle (Crysnanto et al. 2021), representation in genome graphs provide benefits such as reduced reference bias and improved SV detection as a result of coherent representation of alleles (Eizenga et al. 2020). However, the field of pan-genome graphs is still emerging and is the construction and analysis of such genome graphs are still far from trivial. There is no coherent format for graphs, and most of the available tools seem to be in development. So far, creation, indexing and alignment steps tend to be slower than for linear reference genomes (Eizenga et al. 2020).

SVs are widely recognized as a major source of genomic variation impacting adaptive evolution and species diversification. Traditionally, SVs are underrepresented in genetic studies due to both the traditional reliance on a single linear reference genome and technological limitations in the commonly used approaches for high-

throughput population level genotyping. With the introduction of long-read sequencing we are much better equipped to detect and genotype SVs, and multiple studies have recently revealed large numbers of novel SVs, including tomato (Alonge et al. 2020), chocolate tree (Hämälä et al. 2021), humans (Beyter et al. 2021), songbirds (Weissensteiner et al. 2020) and silkworms (Tong et al. 2022). The content of SVs in genomes is much greater than previously believed, for example, they found that SVs cover more than 7% of the cattle genome (Gao et al. 2022), ~16% on average of soybean genomes (Liu et al. 2020) and as much as 10% of sequence in humans with African descent were not included in the linear human reference genome (Sherman et al. 2019). In paper I, we find a total of 1,061,452 SVs jointly covering ~15,6% (or ~3 % on average per genome) of the Atlantic salmon genome. This is a substantial amount in line with the expectations of large genetic variability in a wild Atlantic salmon dataset sampled across a wide phylogeographic range. The number of SVs detected across lake whitefish Normal and Dwarf subspecies in paper III is also substantial, with 104k SVs being labelled as highly confident (detected with multiple independent tools), in total covering more than five times more base pairs than SNPs.

In paper I and III, we found the contribution of SVs in the genomes of Atlantic salmon and lake whitefish to be substantial and mainly associated with repeats. Atlantic salmon SVs are significantly enriched with TRs, especially toward the telomere ends, a trend possibly strengthened by the duplicated nature of the Atlantic salmon genome with regions of high similarity and possible interchange between homeologs. The enrichment of SVs and TRs towards the telomere ends is also seen in other species, including humans (Audano et al. 2019). In paper II, we found that TRs were present in many breakpoints of large inversions. Inversions flanked by TRs did not show clear patterns of LD and haplotype structures which may suggest that the repeats might prevent the inversion to develop into supergenes. In the lake whitefish (paper III), we found that most SVs were caused by TEs (72 % vs. 60 % of genome average), with four groups (Tc1-Mariner, Line-L2, Gypsy and ERV1) being still active with distinct peaks in the SV-length distribution. The SVs contributing to differentiation between the species pair were enriched with TEs, which indicate that TEs may be important for speciation. This finding is in concordance with de Boer et al. (2007), postulating that TEs could contribute to speciation. Overall, we see that repeats and SVs are tightly interconnected, which underlines the benefits of applying long-read sequencing to explore the full SV landscape.

There are currently no clear “best practice” pipelines for long-read based SV detection. So far, the benchmarks that has been performed suggests that finding the consensus call from multiple tools provides the highest precision and recall (e.g. (Dierckxsens et al. 2021; Liu et al. 2022)). One problem we encountered when applying the consensus approach in paper I and III, using the three methods Sniffles (Sedlazeck et al. 2018), SVIM (Heller & Vingron 2019) and NanoVar (Tham et al. 2020), was the variable representation of insertion sequences. For example, SVIM does not output insertion sequence at all, while Sniffles collects the insertion sequence from one possibly noisy read. More recently, Sniffles2 has been released which specifically tackles this problem through creating a consensus call. Additionally, the program has overall increased accuracy and speed (Smolka et al. 2022), and might be a good single method that could match results from the consensus approach.

Current analyses of long-read data suggests that longer insertions and inversions are not fully resolved (Mahmoud et al. 2019; Nattestad & Schatz 2016; Tian et al. 2018). We found (in paper II) that direct assembly comparisons outperformed long-read mapping for the detection of longer inversions (>100 kbp). Inversions in this size range will likely not have reads covering both breakpoints, and hence are more difficult to accurately resolve. When using both direct assembly comparisons and long-read mapping methods to detect SVs in lake whitefish (paper III), we found that most larger insertions could only be detected with assembly comparisons. In addition to being more costly than read mapping methods, direct assembly methods normally lack haplotype representation as diploid genomes are collapsed into one sequence and hence will miss potential variation. Still, when both data types are available, the combination of long-reads and direct assembly comparison methods might supplement each other for the analysis of a complete set of SVs.

Massive amounts of short-read datasets are readily available in online databases after decades of whole genome resequencing. This, together with the development of bioinformatics tools utilizing short-reads to genotype SVs (Huddleston et al. 2017), suggest that combining long-reads for detection and short-reads for genotyping could be a viable option for exploring the impact of SVs on the population scale (Eggertsson et al. 2019). Multiple tools have been developed to this end, including Graphtyper2 (Eggertsson et al. 2019) used in paper I, and the Variation Graph toolkit (Garrison et al. 2018) used in paper III. Although we found bias towards more deletions than insertions when genotyping with Graphtyper2 in paper I, we were able to genotype

304,407 SVs in 366 wild salmon and pinpoint numerous SVs associated with environmental adaptation. In paper III, we genotyped ~90k SV in 32 fish with the vg pipeline (Garrison et al. 2018), which provided insights into the genomic architecture of recent speciation by SVs. These results show that combining short- and long-read sequencing in a genome graph has great potential for unveiling previously hidden genomic variation. However, there are other emerging methods that could be used to improve SV-genotyping. For example, adaptive sampling (also known as Read Until) allows for long-read sequencing of many samples in real time to a pre-defined list of targets (Loose et al. 2016). In that way, the nanopore instrument can limit the sequencing to regions of interest, which could be genome sequences with known SVs, covering either the full SV or SV-breakpoints.

The work presented in this thesis involves the detection and analysis of hundreds of thousands of novel SVs in Atlantic salmon and lake whitefish. However, as a limited number of samples were included for the SV-detection by long-reads (11 salmon and two lake whitefish) this probably represent only a fraction of the full repertoire of SVs. Thus, additional sequencing is needed to better describe the SV-landscape in these species. In Atlantic salmon this could include samples of farmed salmon to supplement the present SV catalogue. These SVs could then be used to conduct genome wide association studies (GWAS) associating SVs with disease resistance and other important traits in aquaculture. Further analysis of the functional effects of SVs would benefit from functional annotation of the salmon genome (e.g. ATAC-Seq, ChIP-Seq), as results from other species have shown that many SVs could affect regulatory regions rather than protein coding sequence directly (Alonge et al. 2020). Investigating effects on the transcriptome level would also be valuable to understand the functional effects of SVs, as well as methods of genome editing (e.g. CRISPR-Cas9).

## 5 References

- Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., Suresh, H., Ramakrishnan, S., Maumus, F., Ciren, D., et al. (2020). Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell*, 182 (1): 145-161.e23.
- Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E. & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, 21 (1): 30.
- Armstrong, J., Hickey, G., Diekhans, M., Fiddes, I. T., Novak, A. M., Deran, A., Fang, Q., Xie, D., Feng, S., Stiller, J., et al. (2020). Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature*, 587 (7833): 246-251.
- Audano, P. A., Sulovari, A., Graves-Lindsay, T. A., Cantsilieris, S., Sorensen, M., Welch, A. E., Dougherty, M. L., Nelson, B. J., Shah, A., Dutcher, S. K., et al. (2019). Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell*, 176 (3): 663-675.e19.
- Berdan, E. L., Blanckaert, A., Butlin, R. K. & Bank, C. (2021a). Deleterious mutation accumulation and the long-term fate of chromosomal inversions. *PLOS Genetics*, 17 (3): e1009411.
- Berdan, E. L., Fuller, R. C. & Kozak, G. M. (2021b). Genomic landscape of reproductive isolation in *Lucania killifish*: The role of sex loci and salinity. *Journal of Evolutionary Biology*, 34 (1): 157-174.
- Berdan, E. L., Flatt, T., Kozak, G. M., Lotterhos, K. E. & Wielstra, B. (2022). *Genomic architecture of supergenes: connecting form and function*, 377, 1856: The Royal Society. p. 20210192.
- Bernatchez, L. & Dodson, J. J. (1990). Allopatric origin of sympatric populations of lake whitefish (*Coregonus clupeaformis*) as revealed by mitochondrial-DNA restriction analysis. *Evolution*, 44 (5): 1263-1271.
- Bernatchez, L., Renaut, S., Whiteley, A. R., Derome, N., Jeukens, J., Landry, L., Lu, G., Nolte, A. W., Østbye, K., Rogers, S. M., et al. (2010). On the origin of species: insights from the ecological genomics of lake whitefish. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365 (1547): 1783-1800.
- Bertolotti, A. C., Lauer, R. M., Gundappa, M. K., Gallagher, M. D., Pehlivanoglu, E., Nome, T., Robledo, D., Kent, M. P., Røsæg, L. L., Holen, M. M., et al. (2020). The structural variation landscape in 492 Atlantic salmon genomes. *Nature Communications*, 11 (1): 5176.
- Beyter, D., Ingimundardottir, H., Oddsson, A., Eggertsson, H. P., Bjornsson, E., Jonsson, H., Atlason, B. A., Kristmundsdottir, S., Mehringer, S., Hardarson, M. T., et al. (2021). Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nature Genetics*, 53 (6): 779-786.

- Bourret, V., Kent, M. P., Primmer, C. R., Vasemägi, A., Karlsson, S., Hindar, K., McGinnity, P., Verspoor, E., Bernatchez, L. & Lien, S. (2013). SNP-array reveals genome-wide patterns of geographical and potential adaptive divergence across the natural range of Atlantic salmon (*Salmo salar*). *Molecular Ecology*, 22 (3): 532-551.
- Catanach, A., Crowhurst, R., Deng, C., David, C., Bernatchez, L. & Wellenreuther, M. (2019). The genomic pool of standing structural variation outnumbers single nucleotide polymorphism by threefold in the marine teleost *Chrysophrys auratus*. *Molecular Ecology*, 28 (6): 1210-1223.
- Charlesworth, B. & Barton, N. H. (2018). The Spread of an Inversion with Migration and Selection. *Genetics*, 208 (1): 377-382.
- Chen, S., Krusche, P., Dolzhenko, E., Sherman, R. M., Petrovski, R., Schlesinger, F., Kirsche, M., Bentley, D. R., Schatz, M. C., Sedlazeck, F. J., et al. (2019). Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biology*, 20 (1): 291.
- Chiang, C., Scott, A. J., Davis, J. R., Tsang, E. K., Li, X., Kim, Y., Hadzic, T., Damani, F. N., Ganel, L., Montgomery, S. B., et al. (2017). The impact of structural variation on human gene expression. *Nature Genetics*, 49 (5): 692-699.
- Clarke, C. A., Sheppard, P. M. & Thornton, I. W. (1968). The genetics of the mimetic butterfly *Papilio memnon* L. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 254 (791): 37-89.
- Coop, G., Witonsky, D., Di Rienzo, A. & Pritchard, J. K. (2010). Using environmental correlations to identify loci underlying local adaptation. *Genetics*, 185 (4): 1411-1423.
- Crawford, S. S. & Muir, A. M. (2008). Global introductions of salmon and trout in the genus *Oncorhynchus*: 1870–2007. *Reviews in Fish Biology and Fisheries*, 18 (3): 313-344.
- Crysnanto, D., Leonard, A. S., Fang, Z.-H. & Pausch, H. (2021). Novel functional sequences uncovered through a bovine multiassembly graph. *Proceedings of the National Academy of Sciences*, 118 (20): e2101056118.
- De-Kayne, R., Selz, O. M., Marques, D. A., Frei, D., Seehausen, O. & Feulner, P. G. D. (2022). Genomic architecture of adaptive radiation and hybridization in Alpine whitefish. *Nature Communications*, 13 (1): 4479.
- de Boer, J. G., Yazawa, R., Davidson, W. S. & Koop, B. F. (2007). Bursts and horizontal evolution of DNA transposons in the speciation of pseudotetraploid salmonids. *BMC Genomics*, 8 (1): 422.
- De Mita, S., Thuillet, A. C., Gay, L., Ahmadi, N., Manel, S., Ronfort, J. & Vigouroux, Y. (2013). Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Molecular ecology*, 22 (5): 1383-1399.
- Dierckxsens, N., Li, T., Vermeesch, J. R. & Xie, Z. (2021). A benchmark of structural variation detection by long reads through a realistic simulated model. *Genome Biology*, 22 (1): 342.
- Eggertsson, H. P., Kristmundsdottir, S., Beyter, D., Jonsson, H., Skuladottir, A., Hardarson, M. T., Gudbjartsson, D. F., Stefansson, K., Halldorsson, B. V. & Melsted, P. (2019). GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nature Communications*, 10 (1): 5402.

- Eizenga, J. M., Novak, A. M., Sibbesen, J. A., Heumos, S., Ghaffaari, A., Hickey, G., Chang, X., Seaman, J. D., Rounthwaite, R., Ebler, J., et al. (2020). Pangenome Graphs. *Annual Review of Genomics and Human Genetics*, 21 (1): 139-162.
- FAO. (2022). The State of World Fisheries and Aquaculture 2022. Towards Blue Transformation.
- Faria, R., Chaube, P., Morales, H. E., Larsson, T., Lemmon, A. R., Lemmon, E. M., Rafajlović, M., Panova, M., Ravinet, M., Johannesson, K., et al. (2019). Multiple chromosomal rearrangements in a hybrid zone between *Littorina saxatilis* ecotypes. *Molecular Ecology*, 28 (6): 1375-1393.
- Frei, D., De-Kayne, R., Selz, O. M., Seehausen, O. & Feulner, P. G. D. (2022). Genomic variation from an extinct species is retained in the extant radiation following speciation reversal. *Nature Ecology & Evolution*, 6 (4): 461-468.
- Frichot, E., Schoville, S. D., Bouchard, G. & François, O. (2013). Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular biology and evolution*, 30 (7): 1687-1699.
- Gao, Y., Ma, L. & Liu, G. E. (2022). Initial Analysis of Structural Variation Detections in Cattle Using Long-Read Sequencing Methods. *Genes*, 13 (5).
- Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., Jones, W., Garg, S., Markello, C., Lin, M. F., et al. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, 36 (9): 875-879.
- Gundappa, M. K., To, T.-H., Grønvold, L., Martin, S. A. M., Lien, S., Geist, J., Hazlerigg, D., Sandve, S. R. & Macqueen, D. J. (2022). Genome-Wide Reconstruction of Rediploidization Following Autopolyploidization across One Hundred Million Years of Salmonid Evolution. *Molecular Biology and Evolution*, 39 (1): msab310.
- Gutiérrez-Valencia, J., Hughes, P. W., Berdan, E. L. & Slotte, T. (2021). The Genomic Architecture and Evolutionary Fates of Supergenes. *Genome Biology and Evolution*, 13 (5): evab057.
- Hager, E. R., Harringmeyer, O. S., Wooldridge, T. B., Theingi, S., Gable, J. T., McFadden, S., Neugeboren, B., Turner, K. M., Jensen, J. D. & Hoekstra, H. E. (2022). A chromosomal inversion contributes to divergence in multiple traits between deer mouse ecotypes. *Science*, 377 (6604): 399-405.
- Haldane, J. (1933). The part played by recurrent mutation in evolution. *The American Naturalist*, 67 (708): 5-19.
- Hall, M. C., Basten, C. J. & Willis, J. H. (2006). Pleiotropic Quantitative Trait Loci Contribute to Population Divergence in Traits Associated With Life-History Variation in *Mimulus guttatus*. *Genetics*, 172 (3): 1829-1844.
- Hämälä, T., Wafula, E., Guiltinan, M., Ralph, P., dePamphilis, C. & Tiffin, P. (2021). Genomic structural variants constrain and facilitate adaptation in natural populations of *Theobroma cacao*, the chocolate tree. *Proceedings of the National Academy of Sciences*, 118: e2102914118.
- Heller, D. & Vingron, M. (2019). SVIM: structural variant identification using mapped long reads. *Bioinformatics*, 35 (17): 2907-2915.
- Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., Poss, M. L., Reed, L. K., Storfer, A. & Whitlock, M. C. (2016). Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. *The American Naturalist*, 188 (4): 379-397.



- Holsinger, K. E. & Weir, B. S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nature Reviews Genetics*, 10 (9): 639-650.
- Homolka, D., Ivanek, R., Capkova, J., Jansa, P. & Forejt, J. (2007). Chromosomal rearrangement interferes with meiotic X chromosome inactivation. *Genome Research*, 17 (10): 1431-1437.
- Hotaling, S., Kelley, J. L. & Frandsen, P. B. (2021). Toward a genome sequence for every animal: Where are we now? *Proceedings of the National Academy of Sciences*, 118 (52): e2109019118.
- Houston, R. D. & Macqueen, D. J. (2019). Atlantic salmon (*Salmo salar* L.) genetics in the 21st century: taking leaps forward in aquaculture and biological understanding. *Animal Genetics*, 50 (1): 3-14.
- Houston, R. D., Bean, T. P., Macqueen, D. J., Gundappa, M. K., Jin, Y. H., Jenkins, T. L., Selly, S. L. C., Martin, S. A. M., Stevens, J. R., Santos, E. M., et al. (2020). Harnessing genomics to fast-track genetic improvement in aquaculture. *Nature Reviews Genetics*, 21 (7): 389-409.
- Hsieh, P., Vollger, M. R., Dang, V., Porubsky, D., Baker, C., Cantsilieris, S., Hoekzema, K., Lewis, A. P., Munson, K. M., Sorensen, M., et al. (2019). Adaptive archaic introgression of copy number variants and the discovery of previously unknown human genes. *Science*, 366 (6463): eaax2083.
- Huddleston, J., Ranade, S., Malig, M., Antonacci, F., Chaisson, M., Hon, L., Sudmant, P. H., Graves, T. A., Alkan, C. & Dennis, M. Y. (2014). Reconstructing complex regions of genomes using long-read sequencing technology. *Genome research*, 24 (4): 688-696.
- Huddleston, J., Chaisson, M. J., Steinberg, K. M., Warren, W., Hoekzema, K., Gordon, D., Graves-Lindsay, T. A., Munson, K. M., Kronenberg, Z. N. & Vives, L. (2017). Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome research*, 27 (5): 677-685.
- Hudson, R. R., Slatkin, M. & Maddison, W. P. (1992). Estimation of levels of gene flow from DNA sequence data. *Genetics*, 132 (2): 583-589.
- Jacobsen, M. W., Hansen, M. M., Orlando, L., Bekkevold, D., Bernatchez, L., Willerslev, E. & Gilbert, M. T. P. (2012). Mitogenome sequencing reveals shallow evolutionary histories and recent divergence time between morphologically and ecologically distinct European whitefish (*Coregonus* spp.). *Molecular Ecology*, 21 (11): 2727-2742.
- Jay, P., Whibley, A., Frézal, L., de Cara, M. Á. R., Nowell, R. W., Mallet, J., Dasmahapatra, K. K. & Joron, M. (2018). Supergene evolution triggered by the introgression of a chromosomal inversion. *Current Biology*, 28 (11): 1839-1845. e3.
- Jiang, Z., Wang, H., Michal, J. J., Zhou, X., Liu, B., Woods, L. C. S. & Fuchs, R. A. (2016). Genome wide sampling sequencing for SNP genotyping: methods, challenges and future development. *International Journal of Biological Sciences*, 12 (1): 100.
- Johnsson, J. I. & Näslund, J. (2018). Studying behavioural variation in salmonids from an ecological perspective: observations questions methodological considerations. *Reviews in Fish Biology and Fisheries*, 28 (4): 795-823.

- King, T. L., Verspoor, E., Spidle, A. P., Gross, R., Phillips, R. B., Koljonen, M. L., Sanchez, J. A. & Morrison, C. L. (2007). Biodiversity and Population Structure. In *The Atlantic Salmon*, pp. 117-166.
- Kirubakaran, T. G., Grove, H., Kent, M. P., Sandve, S. R., Baranski, M., Nome, T., De Rosa, M. C., Righino, B., Johansen, T. & Otterå, H. (2016). Two adjacent inversions maintain genomic differentiation between migratory and stationary ecotypes of Atlantic cod. *Molecular ecology*, 25 (10): 2130-2143.
- Kirubakaran, T. G., Andersen, Ø., Moser, M., Árnýasi, M., McGinnity, P., Lien, S. & Kent, M. (2020). A Nanopore Based Chromosome-Level Assembly Representing Atlantic Cod from the Celtic Sea. *G3 Genes/Genomes/Genetics*, 10 (9): 2903-2910.
- Landis, J. B., Soltis, D. E., Li, Z., Marx, H. E., Barker, M. S., Tank, D. C. & Soltis, P. S. (2018). Impact of whole-genome duplication events on diversification rates in angiosperms. *American Journal of Botany*, 105 (3): 348-363.
- Li, H., Feng, X. & Chu, C. (2020). The design and construction of reference pangenome graphs with minigraph. *Genome Biology*, 21 (1): 265.
- Li, P., Guo, M., Wang, C., Liu, X. & Zou, Q. (2015). An overview of SNP interactions in genome-wide association studies. *Briefings in Functional Genomics*, 14 (2): 143-155.
- Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J. K., Monlong, J., Abel, H. J., et al. (2022). A Draft Human Pangenome Reference. *bioRxiv*: 2022.07.09.499321.
- Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Kent, M. P., Nome, T., Hvidsten, T. R., Leong, J. S., Minkley, D. R., Zimin, A., et al. (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature*, 533 (7602): 200-205.
- Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G.-A., Zhang, H., Liu, Z., Shi, M., et al. (2020). Pan-Genome of Wild and Cultivated Soybeans. *Cell*, 182 (1): 162-176.e13.
- Liu, Y. H., Luo, C., Golding, S., Ioffe, J. & Zhou, X. (2022). Methods for structural variant detection with long-read sequencing data.
- Loose, M., Malla, S. & Stout, M. (2016). Real-time selective sequencing using nanopore technology. *Nature Methods*, 13 (9): 751-754.
- 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.
- Low, W. Y., Tearle, R., Liu, R., Koren, S., Rhie, A., Bickhart, D. M., Rosen, B. D., Kronenberg, Z. N., Kingan, S. B., Tseng, E., et al. (2020). Haplotype-resolved genomes provide insights into structural variation and gene content in Angus and Brahman cattle. *Nature Communications*, 11 (1): 2071.
- Lundsgaard-Hansen, B., Matthews, B., Vonlanthen, P., Taverna, A. & Seehausen, O. (2013). Adaptive plasticity and genetic divergence in feeding efficiency during parallel adaptive radiation of whitefish (*Coregonus* spp.). *Journal of Evolutionary Biology*, 26 (3): 483-498.
- Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C. & Sedlazeck, F. J. (2019). Structural variant calling: the long and the short of it. *Genome Biology*, 20 (1): 246.

- Mank, J. E. & Avise, J. C. (2006). Phylogenetic conservation of chromosome numbers in Actinopterygian fishes. *Genetica*, 127 (1): 321-327.
- Matschiner, M., Barth, J. M. I., Tørresen, O. K., Star, B., Baalsrud, H. T., Briec, M. S. O., Pampoulie, C., Bradbury, I., Jakobsen, K. S. & Jentoft, S. (2022). Supergene origin and maintenance in Atlantic cod. *Nature Ecology & Evolution*, 6 (4): 469-481.
- Mérot, C., Oomen, R. A., Tigano, A. & Wellenreuther, M. (2020). A roadmap for understanding the evolutionary significance of structural genomic variation. *Trends in Ecology & Evolution*, 35 (7): 561-572.
- Method of the Year 2022: long-read sequencing. (2023). *Nature Methods*, 20 (1): 1-1.
- Nattestad, M. & Schatz, M. C. (2016). Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics*, 32 (19): 3021-3023.
- Navarro, A. & Barton, N. H. (2003). Accumulating postzygotic isolation genes in parapatry: a new twist on chromosomal speciation. *Evolution*, 57 (3): 447-459.
- Nelson, J., Grande, T. & Wilson, M. (2006). Fishes of the world 4th ed. *John Wiley and Sons, New York, USA*: 1-624.
- Noor, M. A., Grams, K. L., Bertucci, L. A. & Reiland, J. (2001). Chromosomal inversions and the reproductive isolation of species. *Proceedings of the National Academy of Sciences*, 98 (21): 12084-12088.
- Ohno, D. S. (1970). *Evolution by Gene Duplication*. Springer Berlin Heidelberg.
- Pang, A. W., MacDonald, J. R., Pinto, D., Wei, J., Rafiq, M. A., Conrad, D. F., Park, H., Hurler, M. E., Lee, C., Venter, J. C., et al. (2010). Towards a comprehensive structural variation map of an individual human genome. *Genome Biology*, 11 (5): R52.
- Pearse, D. E., Barson, N. J., Nome, T., Gao, G., Campbell, M. A., Abadía-Cardoso, A., Anderson, E. C., Rundio, D. E., Williams, T. H., Naish, K. A., et al. (2019). Sex-dependent dominance maintains migration supergene in rainbow trout. *Nature Ecology & Evolution*, 3 (12): 1731-1742.
- Pettersson, M. E., Rochus, C. M., Han, F., Chen, J., Hill, J., Wallerman, O., Fan, G., Hong, X., Xu, Q. & Zhang, H. (2019). A chromosome-level assembly of the Atlantic herring genome—detection of a supergene and other signals of selection. *Genome research*, 29 (11): 1919-1928.
- Phillips, R. & Rab, P. (2001). Chromosome evolution in the Salmonidae (Pisces): an update. *Biological Reviews*, 76 (1): 1-25.
- Phillips, R. B., Keatley, K. A., Morasch, M. R., Ventura, A. B., Lubieniecki, K. P., Koop, B. F., Danzmann, R. G. & Davidson, W. S. (2009). Assignment of Atlantic salmon (*Salmo salar*) linkage groups to specific chromosomes: Conservation of large syntenic blocks corresponding to whole chromosome arms in rainbow trout (*Oncorhynchus mykiss*). *BMC Genetics*, 10 (1): 46.
- Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M. & Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology*, 24 (17): 4348-4370.
- Renaut, S., Nolte, A. W. & Bernatchez, L. (2009). Gene Expression Divergence and Hybrid Misexpression between Lake Whitefish Species Pairs (*Coregonus* spp. Salmonidae). *Molecular Biology and Evolution*, 26 (4): 925-936.

- Renaut, S., Nolte, A. W., Rogers, S. M., Derome, N. & Bernatchez, L. (2011). SNP signatures of selection on standing genetic variation and their association with adaptive phenotypes along gradients of ecological speciation in lake whitefish species pairs (*Coregonus* spp.). *Molecular Ecology*, 20 (3): 545-559.
- Reshma, R. S. & Das, D. N. (2021). Chapter 9 - Molecular markers and its application in animal breeding. In Mondal, S. & Singh, R. L. (eds) *Advances in Animal Genomics*, pp. 123-140: Academic Press.
- Rieseberg, L. H. (2001). Chromosomal rearrangements and speciation. *Trends in Ecology & Evolution*, 16 (7): 351-358.
- Rougeux, C., Bernatchez, L. & Gagnaire, P.-A. (2017). Modeling the Multiple Facets of Speciation-with-Gene-Flow toward Inferring the Divergence History of Lake Whitefish Species Pairs (*Coregonus clupeaformis*). *Genome Biology and Evolution*, 9 (8): 2057-2074.
- Savolainen, O., Lascoux, M. & Merilä, J. (2013). Ecological genomics of local adaptation. *Nature Reviews Genetics*, 14 (11): 807-820.
- Schaal, S. M., Haller, B. C. & Lotterhos, K. E. (2022). Inversion invasions: when the genetic basis of local adaptation is concentrated within inversions in the face of gene flow. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377 (1856): 20210200.
- Schloss, J. A. (2008). How to get genomes at one ten-thousandth the cost. *Nature biotechnology*, 26 (10): 1113-1115.
- Schluter, D. (2000). *The ecology of adaptive radiation*: OUP Oxford.
- Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A. & Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, 15 (6): 461-468.
- Sherman, R. M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., Boorgula, M. P., Chavan, S., Vergara, C., Ortega, V. E., et al. (2019). Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nature Genetics*, 51 (1): 30-35.
- Smolka, M., Rescheneder, P., Schatz, M. C., von Haeseler, A. & Sedlazeck, F. J. (2015). Teaser: Individualized benchmarking and optimization of read mapping results for NGS data. *Genome biology*, 16 (1): 1-10.
- Smolka, M., Paulin, L. F., Grochowski, C. M., Mahmoud, M., Behera, S., Gandhi, M., Hong, K., Pehlivan, D., Scholz, S. W., Carvalho, C. M. B., et al. (2022). Comprehensive Structural Variant Detection: From Mosaic to Population-Level. *bioRxiv*: 2022.04.04.487055.
- Sohn, J.-i. & Nam, J.-W. (2018). The present and future of de novo whole-genome assembly. *Briefings in Bioinformatics*, 19 (1): 23-40.
- Su, X., Wang, B., Geng, X., Du, Y., Yang, Q., Liang, B., Meng, G., Gao, Q., Yang, W., Zhu, Y., et al. (2021). A high-continuity and annotated tomato reference genome. *BMC Genomics*, 22 (1): 898.
- Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., et al. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, 526 (7571): 75-81.

- Tham, C. Y., Tirado-Magallanes, R., Goh, Y., Fullwood, M. J., Koh, B. T. H., Wang, W., Ng, C. H., Chng, W. J., Thiery, A., Tenen, D. G., et al. (2020). NanoVar: accurate characterization of patients' genomic structural variants using low-depth nanopore sequencing. *Genome Biology*, 21 (1): 56.
- Thompson, M. J. & Jiggins, C. D. (2014). Supergenes and their role in evolution. *Heredity*, 113 (1): 1-8.
- Tian, S., Yan, H., Klee, E. W., Kalmbach, M. & Slager, S. L. (2018). Comparative analysis of de novo assemblers for variation discovery in personal genomes. *Briefings in bioinformatics*, 19 (5): 893-904.
- Tong, X., Han, M.-J., Lu, K., Tai, S., Liang, S., Liu, Y., Hu, H., Shen, J., Long, A., Zhan, C., et al. (2022). High-resolution silkworm pan-genome provides genetic insights into artificial selection and ecological adaptation. *Nature Communications*, 13 (1): 5619.
- Tonteri, A., Veselov, A. J., Titov, S., Lumme, J. & Primmer, C. (2007). The effect of migratory behaviour on genetic diversity and population divergence: a comparison of anadromous and freshwater Atlantic salmon *Salmo salar*. *Journal of Fish Biology*, 70: 381-398.
- Tonteri, A., Veselov, A. J., Zubchenko, A. V., Lumme, J. & Primmer, C. R. (2009). Microsatellites reveal clear genetic boundaries among Atlantic salmon (*Salmo salar*) populations from the Barents and White seas, northwest Russia. *Canadian Journal of Fisheries and Aquatic Sciences*, 66 (5): 717-735.
- van Dijk, E. L., Jaszczyszyn, Y., Naquin, D. & Thermes, C. (2018). The Third Revolution in Sequencing Technology. *Trends in Genetics*, 34 (9): 666-681.
- Verspoor, E., Stradmeyer, L. & Nielsen, J. (2007). The Atlantic salmon. *The Atlantic salmon: genetics, conservation and management*, 1: 17-56.
- Wang, T., Antonacci-Fulton, L., Howe, K., Lawson, H. A., Lucas, J. K., Phillippy, A. M., Popejoy, A. B., Asri, M., Carson, C., Chaisson, M. J. P., et al. (2022). The Human Pangenome Project: a global resource to map genomic diversity. *Nature*, 604 (7906): 437-446.
- Weissensteiner, M. H., Bunikis, I., Catalán, A., Francoijs, K.-J., Knief, U., Heim, W., Peona, V., Pophaly, S. D., Sedlazeck, F. J., Suh, A., et al. (2020). Discovery and population genomics of structural variation in a songbird genus. *Nature Communications*, 11 (1): 3403.
- Weissing, F. J., Edelaar, P. & van Doorn, G. S. (2011). Adaptive speciation theory: a conceptual review. *Behavioral Ecology and Sociobiology*, 65 (3): 461-480.
- Wright, S. (1978). Modes of Speciation. Michael JD White WH Freeman and Co., San Francisco. 1978. VIII+ 456 pp. illus. \$27.50. *Paleobiology*, 4 (3): 373-379.
- Zhang, L., Reifová, R., Halenková, Z. & Gompert, Z. (2021). How Important Are Structural Variants for Speciation? *Genes*, 12 (7).
- Zimmerman, S. J., Aldridge, C. L. & Oyler-McCance, S. J. (2020). An empirical comparison of population genetic analyses using microsatellite and SNP data for a species of conservation concern. *BMC Genomics*, 21 (1): 382.



# PAPER I

# Atlantic salmon pan-genome reveals hidden genomic variation impacting environmental adaptation

Kristina Severine Rudskjær Stenløkk, Michel Moser, Øystein Monsen, Anna Sofie Kjelstrup, Mariann Árnýasi, Torfinn Nome, Simen Sandve, Matthew Kent, Nicola Barson, Sigbjørn Lien

Centre for Integrative Genetics (CIGENE) and Department of Animal and Aquacultural Sciences, Faculty of Biosciences, Norwegian University of Life Sciences.

## Abstract

Structural variations (SVs) are widely recognized as a major source of genomic variation impacting adaptive evolution and species diversification. However, their functional importance is poorly understood largely because they are challenging to detect and genotype at the population-scale involving large numbers of samples. Here we present the first pan-genome for Atlantic salmon, comprising 11 long-read-based assemblies from all four phylogeographical lineages. High quality chromosome-level assemblies were constructed for three of these lineages. We detected 1,061,452 SVs capturing 367 Mbp of sequence, wherein 13,038 SVs overlapped the coding sequence of 2,725 unique genes, implicating that they directly affect gene functions. Repeat annotation revealed a marked enrichment of tandem repeats (TRs) in chromosome regions towards both extant and historic telomeres. For SVs, the enrichment was much more pronounced in extant than historical telomeric regions, suggesting that the salmon TRs have become less variable when translocated to intra-chromosomal positions. We found a highly significant enrichment of SVs in duplicated versus singleton genes suggesting that duplicate retention has played a role in shielding the impact of deleterious SVs in the salmon genome. We genotyped 304,407 SVs using graph genome analysis and short-read data for 366 salmon sampled from contrasting riverine environments across the natural range. Genotype-environment association analyses revealed GO enrichment of neurological processes and highlighted 13 KEGG pathways including the GnRH signalling pathway known to be one of the main regulators of reproductive function in vertebrates. SVs overlapping protein coding sequences of genes associated with environmental adaptation highlight the importance of several polymorphic immunoglobulin regions and an 18 kbp deletion on chromosome 28 encompassing a segmental duplication of three genes. These three genes are luteinizing hormone subunit beta (*LHB*) involved in late stages of maturation, and two genes involved in cellular respiration; G-protein coupled receptor (*GPR4*) and Cytochrome C Oxidase Assembly Factor (*COX20*). Our results demonstrate how long-read pan-genomics can reveal previously hidden complex genetic variants and that these likely have consequences for fitness and adaptations in the wild. We anticipate the revealing of this previously hidden functional variation to inform future management of vulnerable wild populations of Atlantic salmon and contribute to sustainable aquaculture.



## Introduction

Structural genomic variations (SVs) represent a distinct type of genomic variation that is far more diverse in both conformation and size, and which cumulatively affect more sequence than single nucleotide polymorphisms (SNPs) (Kosugi *et al.* (2019); (Sudmant *et al.* 2015). SVs may also cause large phenotypic effects, particularly if they directly affect gene functions through disrupting coding regions or regulatory elements (Alonge *et al.* 2020; Chiang *et al.* 2017). Despite the potential functional consequences of SVs, results documenting this are still rare because most of these studies are based on short-read sequencing data with limited ability to resolve SVs within or close to repeats (Alkan *et al.* 2011; Lappalainen *et al.* 2019; Sudmant *et al.* 2015).

The tendency to create large effect mutations means that SVs are also likely to be highly deleterious. This is reflected in their depletion in coding sequence in humans (Beyter *et al.* 2021), and causal involvement in human disease phenotypes (Beroukhim *et al.* 2010; Sebat *et al.* 2007; Talkowski *et al.* 2012). However, some SVs with large effects appear to be beneficial mutations that are subject to balancing selection, suggesting they are important for adaptation (Hämälä *et al.* 2021; Yan *et al.* 2021). This is particularly evident for inversions where many iconic adaptive polymorphisms have been found to be controlled by inversion polymorphisms (Dobzhansky 1947; Pearse *et al.* 2019; Villoutreix *et al.* 2020). Currently, the role of other types of SVs in adaptation is much less well studied and it is not clear how large a role they play.

Repetitive DNA, often grouped into tandem repeats (TRs) and transposable elements (TEs), may constitute as much as 50-90% of eukaryote genomes (de Koning *et al.* 2011; Garrido-Ramos 2017; Liu *et al.* 2019; Mehrotra & Goyal 2014; Platt *et al.* 2016). TRs are defined as adjacently repeated stretches of DNA where the length of the repeated unit and sequence composition can vary widely (Lu *et al.* 2021; Sulovari *et al.* 2019). A common class of TRs is satellite DNA (satDNA), which are head-to-tail tandemly repeated non-coding DNA sequences primarily organized in long arrays (Garrido-Ramos 2017). TRs are typically enriched around telomeric and centromeric regions and playing important roles in cell division-related processes such as recombination and cytokinesis (Garrido-Ramos 2017). While many TRs appear to be nonpolymorphic, some minisatellites, often termed Variable Number Tandem Repeats (VNTRs), exhibit high copy number variability (Eslami Rasekh *et al.* 2021). Changes in VNTR copy number have been proposed to arise by template slippage or switching (Course *et al.* 2020) strand mispairing (Taylor & Breden 2000), unequal crossover (Jeffreys *et al.* 1998), and by gene conversion/tandem duplication (Farnoud *et al.* 2019; Pâques *et al.* 1998). The second major category of repetitive DNA sequences; TEs, are mobile, self-replicating elements present in eukaryotic genomes (Bourque *et al.* 2018). The role of TEs in generating SV is mostly linked to their transposition activity, resulting in insertion and deletion variations (Mun *et al.* 2021). Scattered and highly similar TE-copies can also result in ectopic recombination giving rise to inversions or translocations (Kent *et al.* 2017). However, the extent to which, and how, repeats contribute to SVs formation and distribution in vertebrate genomes is largely unexplored.

Atlantic salmon (*Salmo salar*) is an important fish species in the northern hemisphere and a high value resource for both aquaculture, wild stock fisheries and recreational sport fisheries OECD (2017). The fishes are also an important resource for indigenous cultures and artisanal fisheries (Lam & Borch 2011). Atlantic salmon colonized its current geographic distribution following the retreat of the glaciers

10,000 – 15,000 years ago. This recolonization initiated extensive adaptation to different riverine environments across broad latitudinal clines (Rougemont & Bernatchez 2018) and today the species is divided into four phylogeographic groups: Atlantic, Barents/White sea, Baltic and North America (Bourret *et al.* 2013), with the greatest divergence found between European and North American lineages, which separated more than 600,000 years before present (King *et al.* 2007).

The ancestor of Atlantic salmon experienced a whole genome duplication (Ss4R) event ~89-125 Mya (Gundappa *et al.* 2022), meaning that the present Atlantic salmon genome is characterized by extensive gene copy expansions and high redundancy (Lien *et al.* 2016). This redundancy could act to ameliorate the impacts of large effect mutations (Sinclair-Waters *et al.* 2022) and enhance the role of SVs in phenotypic diversity and adaptation in salmon. Large rearrangements, coinciding with bursts of repeat expansions, have been suggested as a mechanism for reverting the ancestral autotetraploid Atlantic salmon genome into disomic inheritance through rediploidization (Lien *et al.* 2016). Thus, Atlantic salmon represent an attractive species to contextualize not only the process of repeat expansion following WGD, but also how repeats impact the SV-landscape, and how such variation influences environmental and life history adaptation.

In an earlier study, short-read sequencing data was used to study SVs in 493 wild and farmed Atlantic salmon (Bertolotti *et al.* 2020). However, among the almost 165,000 SVs identified, only 15,483 (9%) were classified as high confidence SVs with the remainder being false-positives or low-confidence. The relatively small fraction of SVs that could be discovered and genotyped with confidence from short-reads highlights the need for the development and implementation of new approaches. Long-read sequencing technologies, which have potential to read through repetitive DNA, represent a game changer in SV-detection (Ho *et al.* 2020), with recent studies reporting hundreds of thousands of novel SVs across a broad size range, for example in tomatoes (Alonge *et al.* 2020), humans (Beyter *et al.* 2021) and songbirds (Weissensteiner *et al.* 2020). In this study we use nanopore long-read sequencing technology to generate a pan-genomic resource for Atlantic salmon including 11 high-quality assemblies from individuals sampled across the natural distribution of the species. This pan-genome were used to describe the SV-landscape and the role of SVs in environmental adaptation in Atlantic salmon.

## Results

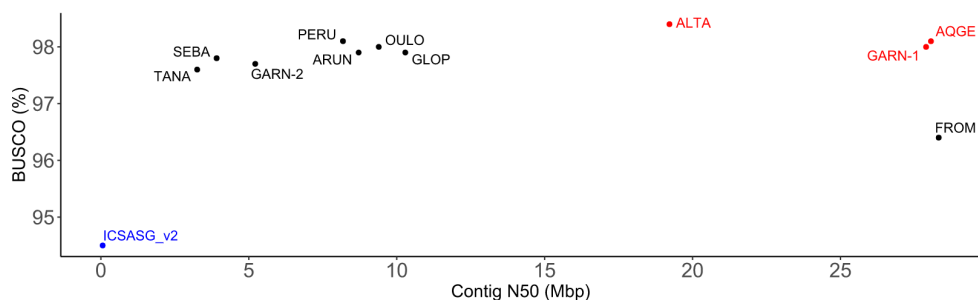
### Construction of chromosome-scale reference genomes for Atlantic salmon

To capture the genomic diversity of SVs within Atlantic salmon, we sequenced 11 fish (Table S1) from four phylogeographic groups covering the natural distribution of the species (Bourret *et al.* 2013); Atlantic (ATL), Barents/White Sea (BWS), Baltic (BAL) and North American (NAM). The samples were sequenced with nanopore long-read technology (individual read depth 16-72x, Table S2), and *de novo* assembled into contigs with a mean N50 of 13.86 Mbp (range 3.26-28.32 Mb) (Table S3, Figure 1). The genomes were polished with sample-derived Illumina short-reads. Three of the highest quality genomes, representing the phylogeographic groups ATL, BWS and NAM, were assembled into chromosome sequences using chromosome conformation data generated using Hi-C or Pore-C protocols (Figure S1-S3). Among these, the highest quality genome from the ATL group (AQGE, sampled from an aquaculture strain) was chosen as a reference genome; Ssal\_v3.1 (GCA\_905237065.2). Ssal\_v3.1 has 2.50 Gbp assembled into chromosome sequences, 259 Mbp more sequence than in the previous salmon

reference genome (ICSASG\_v2) which was constructed from Sanger sequencing and Illumina short-reads (Lien *et al.* 2016). The genome has a highly improved continuity, with the number of contigs decreasing from 368,060 to 4,222 (57-fold), and the contig N50 (ctgN50) increasing from 58 kbp to 28.06 Mbp (484-fold) (see Table S3). The completeness of the genome, as measured by Benchmarking Universal Single-Copy Orthologs (BUSCO) (Manni *et al.* 2021), increased from 94.5% to 98.1% (Figure 1).

Substantial karyotype differences are reported between Atlantic salmon of European and North American origin. The European type has 29 pairs of chromosomes and 74 chromosome arms, while the North American type is variable, but typically has 27 chromosome pairs and an NF of 72 (Phillips & Rab 2001). The reduction from 29 to 27 chromosome pairs in North America are due to large chromosomal polymorphic rearrangements (Watson *et al.* 2022) involving two fusions (ssa08/29 and ssa26/28) and a translocation (ssa01p/23) (Brenna-Hansen *et al.* 2012). To describe these rearrangements on the genomic level, we constructed a highly continuous (ctgN50 = 27.89 Mbp) chromosome anchored assembly from a wild North American Atlantic salmon from the Garnish River in Canada (GARN-1). A contact map constructed from Pore-C data revealed that the salmon being sequenced contained the ssa26/28 fusion, lacked the ssa08/29 fusion and was polymorphic for the ssa01p/23 translocation (Figure S1).

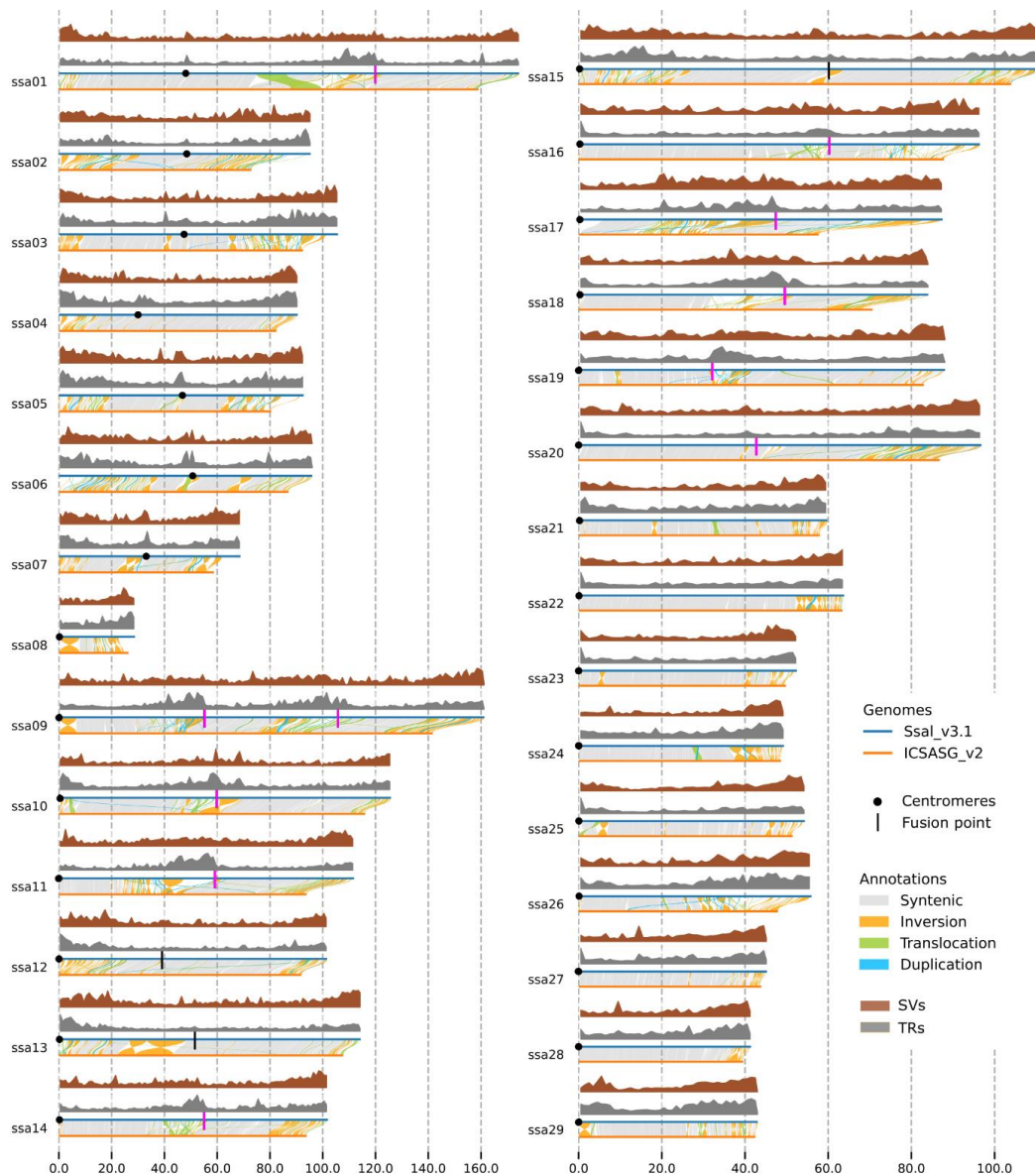
A highly continuous (ctgN50 = 19.22 Mbp) and chromosome anchored assembly was also constructed for a male sampled from the Alta River in the North of Norway; representing the Barents/White Sea phylogeographic group (Table S3). A contact map constructed from Pore-C data (Data S2) revealed that this salmon had a normal European karyotype with 29 chromosomes.



**Figure 1:** Assembly quality measured in BUSCO score (%) and contig N50 for the 11 long-read Atlantic salmon genome assemblies and the previous short-read based assembly ICSASG\_v2 (in blue). Chromosome-level assemblies were constructed for three of the phylogeographical groups; Atlantic (AQGE), North America (GARN-1) and Barents/White Sea (ALTA) marked in red.

To resolve duplicated regions in the Atlantic salmon genome, we aligned the chromosome sequences in the Ssal\_v3.1 assembly against each other and identified 147 homeologous (duplicated) blocks with high collinearity (Data S1). Together, the blocks account for 2.47 Gbp (98.8%) of chromosome-anchored sequences (Figure S4; Data S1). A considerable proportion of the blocks showed high sequence similarity between duplicates, with 849 Mbp (34.4% of the genome) showing a similarity >90% and

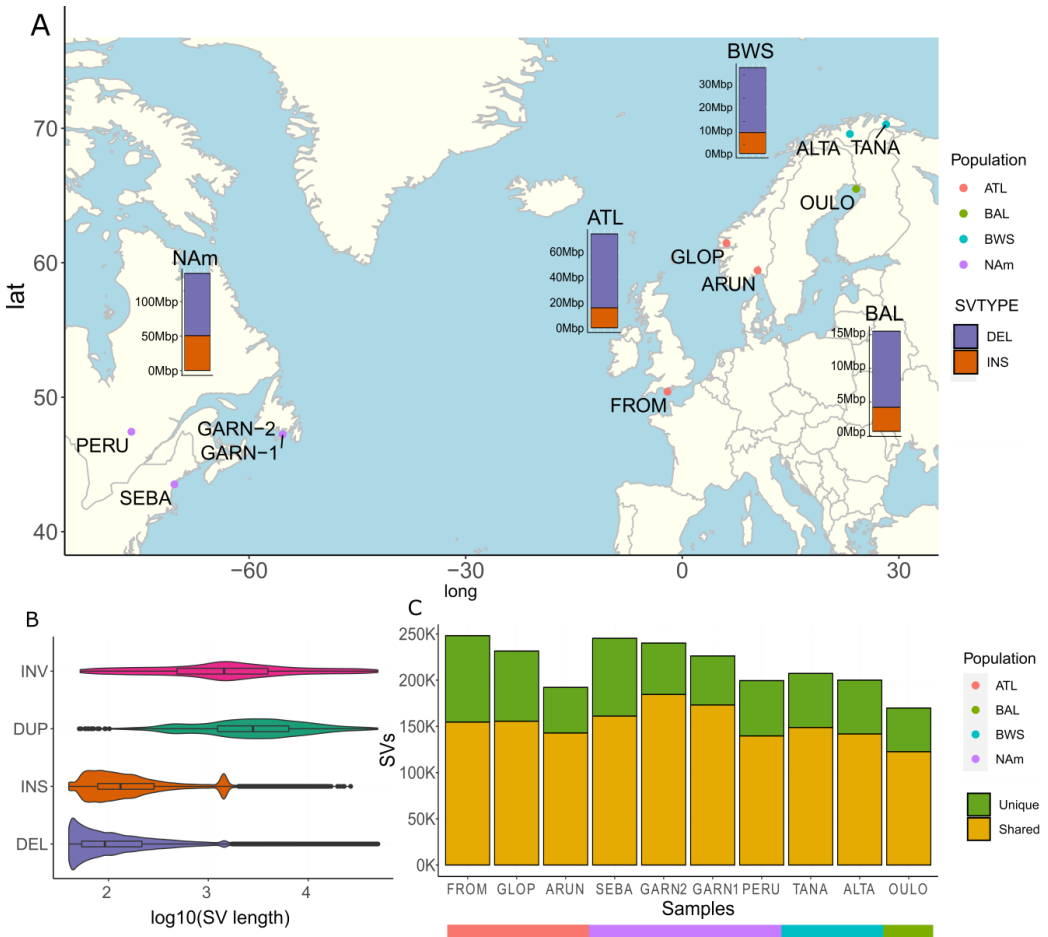
396 Mbp (16% of the genome) displaying a sequence similarity >95%. The majority of the blocks with high sequence similarity correspond to eight larger duplicated regions (2p-5q, 1qb-18qa, 2q-12qa, 3q-6p, 4p-8q, 7q-17qb, 11qa-26 and 16qb-17qa) (Lien *et al.* 2016; Robertson *et al.* 2017) (see Figure S4). These regions turned out to be highly fragmented, rearranged and collapsed between homeologs in the ICSASG\_v2 assembly but are much better resolved in the Ssal\_v3.1 assembly (Figure 2).



**Figure 2.** Syntenic and rearranged regions between Ssal\_v3.1 and ICSASG\_v2 with SV-density (brown) and TR-density (grey) along 29 Atlantic salmon chromosomes (ssa01-ssa29). Pink lines indicate historical telomeres that are translocated into extant intra-chromosomal locations.

## SV-detection reveals previously hidden genomic variation

To catalogue a wide range of SVs in Atlantic salmon we mapped the long-reads of the 10 wild salmon samples to Ssal\_v3.1 using three independent SV-detection software packages. To reduce the number of false positives, SVs identified by just one of the packages were disregarded, leaving a total of 1,061,452 SVs detected independently by multiple pipelines. Deletions ( $n=781,244$ ) and insertions ( $n=275,462$ ) made up most detected SVs, with duplications ( $n=3,340$ ) and inversions ( $n=1,407$ ) contributing modestly to the total SV-landscape (Table S4). The deletions and insertions were relatively short (mean lengths 332 and 389 bp, respectively) compared to inversions and duplications (mean lengths 4,899 and 5,496 bp respectively). The number of SVs detected per individual ranged from 169 to 246 k (Figure 2C; Table S4), with the highest average number of SVs per phylogeographical group found in the NAm group (average 228 k), followed by the ATL (mean 224 k) and BWS (mean 204 k) groups. A large proportion of SVs (632,193) were found in one sample only (Figure 2C; Table S5), reflecting that analyses were based on data from a limited number of individuals sampled from a broad phylogeographic distribution (Figure 3A; Table S1).



**Figure 3.** (A) Sampling sites for Atlantic salmon with bar plots showing the amount of SV-sequence specific to each of the phylogeographical groups; North America (NA<sub>m</sub>), Atlantic (ATL), Barents/White Sea (BWS) and Baltic (BAL). (B) Violin plots showing log transformed size distribution per SV type. (C) Bar plots showing number of SVs detected per sample arranged by phylogeographic groups.

The number of base pairs included in indels (insertions and deletions), also referred to as presence/absence variations (PAVs), totalled 366.61 Mbp across the 10 samples, with an average of 77.4 Mbp of PAV-sequence per sample (range 60.6 to 96.1 Mbp, Table S6). More than 140 Mbp (38.44%) of PAVs overlap genes, for which 2.52 Mbp (0.69%) also overlap protein coding sequences (CDS) (Table S7). The PAVs overlapping CDS, include 13,038 indels and affect 2,725 genes. The effects of indels on gene functions were also modelled using the Variant Effect Predictor pipeline (VEP) (McLaren *et al.* 2016), forecasting high impact consequence for 14,383 indels (Table S8). VEP predicts a slightly higher number of functional indels than CDS overlap alone, as 1,099 intron variants are included, as well as some indels are predicted to have multiple effects. VEP results revealed that the most widespread consequential effect of indels was ‘feature truncation’ (4,462 indels), followed by ‘frameshift variant’ (2,807) and ‘stop lost’ (1,769), see Table S8. In conclusion, both ‘sequence-overlap’ and VEP analyses imply that some indels detected in Atlantic salmon may have significant functional consequences by directly disrupting protein coding regions of genes.

A large proportion (>50%) of Atlantic salmon genes have a retained duplicate copy (ohnolog) after the salmonid specific (Ssa4R) whole genome duplication (Lien, 2016). Many of these duplicated genes are likely functionally redundant (Gillard *et al.* 2021), which is expected to allow for accumulation of large effect deleterious variants. To test if retained duplicated genes more frequently overlap SVs than singleton genes, we tested 11,233 gene pairs (22,466 separate gene sequences) and 9,770 singletons annotated in the Ssal\_v3.1 assembly. In accordance with the results of Bertolotti *et al.* (2020), we found a highly significant enrichment of SVs overlapping any part of duplicated versus singleton genes (Table S9, Fishers exact test: odds ratio = 1.55,  $P < 2.2 \times 10^{-16}$ ). Next, we performed the same test on 1,938 genes where SVs overlapped the CDS. This number is lower than the 2,725 genes identified as SVs overlapping CDS in the section above, as the duplicate/singleton state is not clear for all annotated genes. Again, we found a significant enrichment of SVs in duplicated (1400 genes) compared to singletons (538 genes) (Table S10, Fishers exact test: odds ratio = 1.14,  $P < 0.01$ ). Our results support the findings in Bertolotti *et al.* (2020) and suggest that the functional redundancy of duplicated genes allows for the accumulation of deleterious SVs in duplicated genes.

**SV are enriched in tandem repeats (TRs) but depleted in inactive transposable elements (TEs)**

To annotate the repeat content of the new salmon genome assembly and explore how different classes of repeats contribute to the SV-landscape of Atlantic salmon, we performed both transposable element (TE) and tandem repeat (TR) annotation of Ssal\_v3.1. The total repeat content of Ssal\_v3.1 was estimated to 60.78%, with TEs and TRs accounting for 40.61% and 20.17%, respectively (Table S11). In line with the study of Lien *et al.* (2016), the Tc1-mariner family was found to be the single largest class of TEs (accounting for 11.6% of the genome), but relatively large quantities of LINE-Jockey-like elements (8.3%) and unclassified DNA transposons (5.8%) were also detected (Table S12).

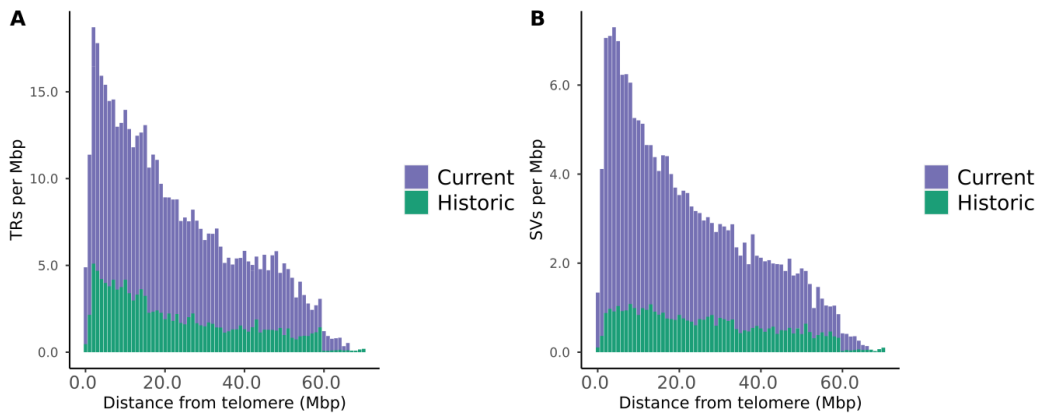
Intersecting TEs and SVs showed an overall depletion of sequences overlapping deletions (24.02%) and insertions (21.15%) compared to the genome wide TE content (40.61%) (Table S11; Table S13). A few TE-families were notably over-represented within SVs indicating recent or current TE activity. This was particularly the case for one Tc1-Mariner element (DTT in Figure S5) bearing close resemblance to the proposedly active DNA transposon reported by Bertolotti *et al.* (2020). This transposon was formerly annotated as a pTSsa2 piggyBac-like DNA transposon due to its sequence similarity to the EF685967.1 element (de Boer *et al.* 2007). It appears as a ~1400 bp fragment in both insertions and deletions (see violin plots in Figure 3B) and overlaps 38,882 deletions, thereby accounting for 4.98% of all deletions in our study. The polymorphic TE element is genome-wide distributed (de Boer *et al.* 2007) and overlaps SV-peaks throughout the genome (see Figure S6).

In contrast to TEs, the correlation between the location of TRs and SVs in the genome was highly significant (Pearson correlation = 0.74,  $P < 2.2 \cdot 10^{-16}$ ). This was true both for both deletions (28.04%) and insertions (35.07%) compared to the overall content of TRs in the genome (20.17%) (see Tables S11 and S13), suggesting that TRs play an important role in forming the SV-landscape of Atlantic salmon.

#### TRs and SVs are enriched in chromosome regions towards extant and historical telomeres

Atlantic salmon typically possess a karyotype with 74 chromosome arms representing an exception to the modal range of 96–104 chromosome arms seen in most extant salmonid fishes (Phillips & Rab 2001). The karyotype is generated by tandem fusions of ancestral chromosomes which, in many cases, translocate historical telomeres into centromeres (the case for ssa12, ssa13 and ssa15) or new intra-chromosomal positions (see Figure S4). This makes Atlantic salmon an attractive system for studying both TRs and SV-landscape in regions towards extant and historical telomeres. To investigate the positional effects of repeats on the SV-landscape in Atlantic salmon, we first plotted the positions of TRs in the extant chromosome structure, revealing an enrichment of TRs towards telomeric ends of chromosomes (Figure 2; Figure S4). The correlation between TR-count and distance to telomeres was highly significant (Pearson correlation = -0.92,  $P < 2.2 \cdot 10^{-16}$ ). Next, we identified 11 historical telomeres being translocated to intra-chromosomal positions within the present Atlantic salmon karyotype, specifically ssa01:119 Mbp, ssa09:55 Mbp, ssa09:106 Mbp, ssa10:59 Mbp, ssa11:59 Mbp, ssa14:51 Mbp, ssa16:59 Mbp, ssa17:47 Mbp, ssa18:49 Mbp, ssa19:32 Mbp and ssa20:43 Mbp (see pink lines in Figure 2). Resembling the results for the extant chromosome structure, the correlation between TR-count and distance to historical telomeres was also highly significant (Pearson correlation = -0.90,  $P < 2.2 \cdot 10^{-16}$ , Figure 4A).

Studies in other species suggest that tandem repeat rich regions towards telomeric ends of chromosomes typically contain more SVs than the genome average (Audano *et al.* 2019; Garrido-Ramos 2017). To investigate this pattern in Atlantic salmon, we tested for correlation between SV-density and distance to both extant and historical telomeres. We found the correlation with SVs and regions towards extant telomeres to be stronger ( $P < 2.2 \cdot 10^{-16}$ , correlation = -0.58) than with historic telomeres (Pearson correlation = -0.24,  $P < 4.5 \cdot 10^{-9}$ , Figure 4B), suggesting that TRs located in regions towards historical telomeres are less variable and hence not as active as those positioned towards extant telomeres.



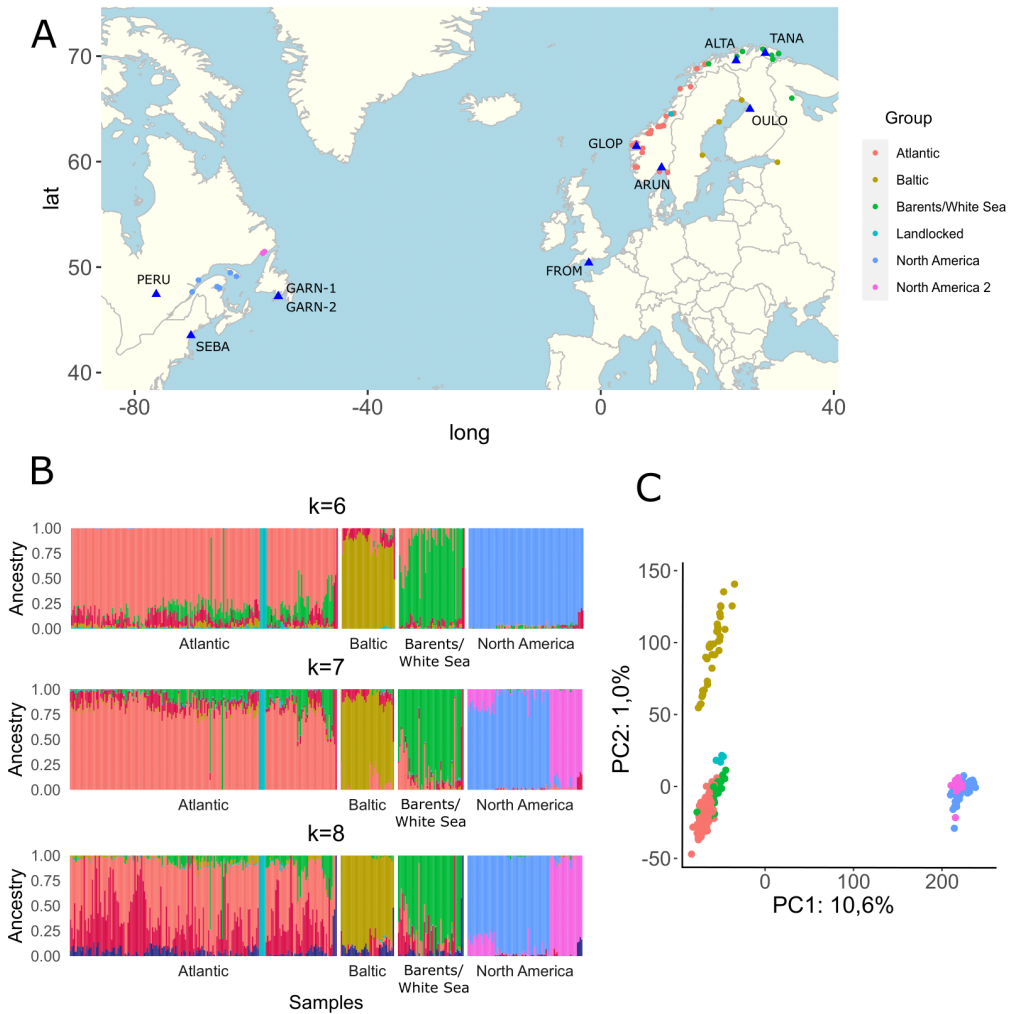
**Figure 4.** Density per Mbp of (A) TRs and (B) SVs and distance from current and historic telomeric regions.

### Pan-genome enable accurate SV-genotyping using short-read data within genome graphs

To transfer the catalogue of SVs discovered using long-read sequencing of 11 individuals into a larger population dataset, we combined high accuracy long-read SV calls and short-read data from 366 salmon in a variation-aware graph structure using the GraphTyper2 software (Eggertsson *et al.* 2019). This approach yielded genotypes for 672,404 SVs, accounting for 63.6% of the 1,056,706 indels initially discovered by long-reads. After filtering the data for minor allele frequency (MAF>0.05) and 70% of missing data (max-missing 0.3), we retained reliable genotypes from 304,407 SVs. The majority of the SVs called by GraphTyper2 (97.6%) were relatively short deletions (297,187 SVs with average size 306 bp), reflecting the enhanced ability of this method to genotype short deletions compared to insertions (Almarri *et al.* 2020; Eggertsson *et al.* 2019).

To confirm data quality, we questioned if the high-confidence SV-genotypes capture expected population genetic structure of wild Atlantic salmon populations. SV genotypes were used in population structure analyses using PCA and NGSadmixmap (Skotte *et al.* 2013). In concordance with previous results (Bertolotti *et al.* 2020; Bourret *et al.* 2013), we found the strongest differentiation between North American and European populations (PC1: 10.6%), followed by Baltic and other European populations (PC2: 1.0%) (Figure 3C). By inspecting the population structure defined by NGSadmixmap (optimal K=7), we confirm the expected phylogeographic grouping of Atlantic, Barents/White Sea and Baltic populations in Europe (Figure 5). In addition, our clustering suggests a separation into two groups in North America, one group with populations BO, CH, JU, LA, MA and PC and another with the populations SP and VF (Data S2).





**Figure 5.** Population structure of 366 Atlantic salmon based on 304,407 SVs genotyped by genome graph analyses (GraphTyper2) making use of short-read sequence data. (A) Map of sampling sites coloured by groups suggested by NGSadmix (optimal  $k=7$ ). Blue triangles correspond to sampling locations of long-read sequenced individuals used to make the initial SV-dataset. (B) NGS-admix plot showing predicted ancestry of  $k=6-8$  suggesting five groups in Europe and two groups in North America. (C) Principal component analysis of 304k genotyped SVs. PC1 separates European and North American samples. PC2 splits the Baltic group from the rest of European samples.

### SVs impact environmental adaptation in Atlantic salmon

To test if SVs contribute to adaptive evolution in natural populations of Atlantic salmon, we performed a genotype-environment association for eight environmental variables (Figure S7 and S8) measuring temperature (annual mean temperature, mean temperature of warmest quarter and mean temperature of coldest quarter, isothermality), precipitation (annual precipitation and precipitation during the wettest quarter), latitude and drainage basin area. These variables capture thermal conditions, precipitation and river size of the spawning habitats used by Atlantic salmon and therefore may exert

selection pressure. Due to the genomic divergence between European and North American lineages, and therefore the likelihood that SVs are not shared between continents, we conducted the analyses separately for each continent.

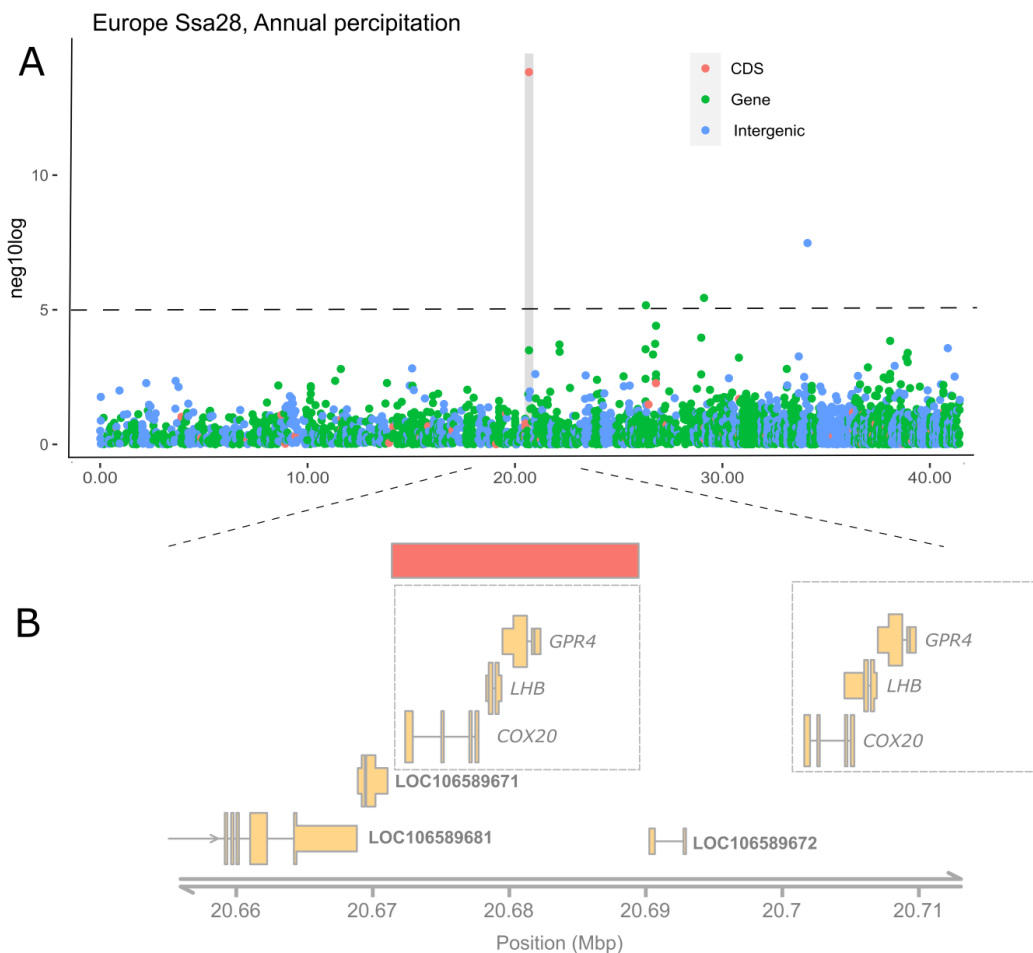
Genotype-environment associations are commonly performed with SNPs as markers, which are more densely and evenly scattered throughout the genome than SVs. In our dataset, the mean distance between SVs was 9.5 kbp, but there was considerable variation ( $SD = 16,15$  kbp). The distribution of distance between SVs is plotted in Figure S9. However, association mapping link environmental factors to SVs with potentially strong effects through disruption of protein coding sequence, changes in gene copy number, or obstruction of regulatory regions with impacts on gene expression (Alonge *et al.* 2020; Hämälä *et al.* 2021). We found thousands of SVs significantly associated with one or more environmental variable (3,136 in Europe and 1,713 in North America,  $P < 0.05$ ), for which 1,582 and 937 overlapped genes in Europe and North America, respectively.

To ascertain if certain biological processes were overrepresented among SVs significantly associated with the environment, we made use of the gene ontology (GO) framework (Ashburner, 2000). GO enrichment tests identified 150 overrepresented biological processes ( $P < 0.05$ ) among the genes linked to environmental associated SVs, with 514 unique genes contributing to the enriched terms (Figure S10, Data S3). Seven of the ten most significantly enriched processes were neurological, suggesting enriched biological processes linked to neuron development. Twenty biological processes were daughter terms of cellular developmental process ( $P < 0.005$ ), including cell differentiation ( $P < 0.005$ ), cell development ( $P < 0.005$ ) and cellular component morphogenesis ( $P < 0.015$ ). Forty unique genes were associated with locomotion ( $P < 0.02$ ) which is not directly related to any other significantly enriched biological processes in our dataset. Locomotion has previously been linked to ecological diversification under relaxed selection in the salmonid fishes *Coregonus* and *Salvelinus* (Schneider *et al.* 2019), which could be relevant for migratory behaviour (McCormick *et al.* 1998), or movement up rivers to hunt prey for juvenile salmon (Godin & Rangeley 1989).

To test for pathway enrichment, we performed KEGG pathway enrichment analysis with the same gene set, revealing 13 significantly enriched pathways ( $P < 0.05$ ) (see Data S4). The five most significantly enriched pathways were adrenergic signalling in cardiomyocytes, focal adhesion, adherent junction, extracellular matrix (ECM) receptor interaction and GnRH signalling pathway. Adrenergic signalling in cardiomyocytes is related to cardiac functions, and is associated with hypoxia tolerance (Cheong *et al.* 2016). Several genes in the adherent junction pathway and other cell shaping regulation pathways have shown to be upregulated in Atlantic salmon infected with nephrocalcinosis and might be part of the inflammatory repair processes (Klykken *et al.* 2022). ECM-coding components has also been shown to be involved in wound healing in Atlantic salmon (Skugor *et al.* 2008). The GnRH signalling pathway is one of the main regulators of reproductive function in vertebrates, including salmon, where it has been shown to regulate the gonadal maturation (Ando & Urano 2005). This pathway includes the Luteinizing Hormone Subunit Beta gene (*LHB*) that is strongly associated with precipitation (see section below).

**An 18 kb deletion overlapping three genes on chromosome 28 is associated with precipitation**  
Significantly associated SVs disrupting CDS will likely affect gene functions and are, therefore, strong candidates for causal variants underlying the association (Guo *et al.* 2020). Our analyses revealed 45

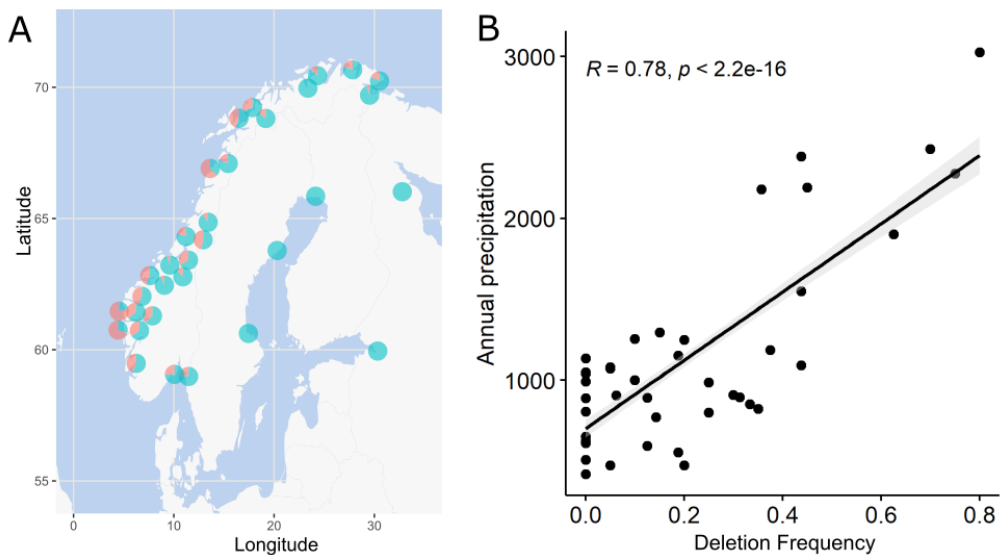
and 39 environmental associated SVs disrupting CDS in Europe and North America, respectively. The most significantly result was found for an 18 kbp deletion on chromosome 28 (Figure 6) associating with environmental factors related to precipitation ( $-\log_{10}(P) = 13.83$ ) in Europe (Figure 6A and S7). Higher deletion frequency per population was significantly correlated with higher precipitation ( $R = 0.78$ ,  $P < 2.2 \times 10^{-16}$ , Figure 7B). The deletion is most frequent in populations in the western part of Norway (Figure 7A) and rare or absent in North American populations. The deletion overlaps a segmental duplication in the region containing three genes; luteinizing hormone subunit beta (*LHB*), G-protein coupled receptor (*GPR4*) and Cytochrome C Oxidase Assembly Factor (*COX20*). The deletion makes the segmental duplication polymorphic with individuals possessing one or two copies of the three genes.



**Figure 6.** An 18 kbp deletion (red) overlapping a segmental duplication containing extra copies of the three genes *LHB*, *GPR4* and *COX20* is significantly associated with annual precipitation in Europe. (A) Associations between SVs and annual precipitation on chromosome 28 (Ssa28). The red square marks a deletion strongly associated

with the environment. (B) Gene models of region with the associated SVs shows that three genes within a segmental duplication (stippled frame to the left) is deleted.

Luteinizing hormone is one of the main pituitary hormones secreted in response to GnRH as part of the brain-pituitary-gonad axis controlling maturation. The *LHB* gene codes for the subunit that confers specificity and is expressed in the pituitary gland of maturing males and female Atlantic salmon, peaking at spermiation and the onset of ovulation (Andersson *et al.* 2013; Mobley *et al.* 2021). Circulating levels of luteinizing hormone peak during spawning in line with its role in the final stages of maturation (Mobley *et al.* 2021). Expression of *LHB* is affected by the external environment, being triggered by decreasing photoperiod (Melo *et al.* 2014) and temperature in Atlantic salmon (King & Pankhurst 2004). The association of the copy number of *LHB* with aspects of the spawning environment is, therefore, consistent with its function in spawning and environmental sensitivity. Both copies of the *LHB* gene on Ssa28 encode full-length proteins, however, with two amino acid differences having potentially functional effects. The first amino acid shift replaces a proline with a histidine in the end of  $\beta$ -loop 3 in the protein. This position is reported to be conserved across fish (Swanson *et al.* 2003). The second amino acid shift (histidine to glutamine) is near the end of the 'seatbelt', a structure that is thought to be important for receptor interactions as well as stability of the heterodimer (Figure S11) (Swanson *et al.* 2003).



**Figure 7.** (A) Frequency of an 18 kbp deletion associated with annual precipitation in European populations. The frequency of reference allele is denoted in blue and the alternative allele (deletion) in red. (B) Scatter plot with Pearson correlation of annual precipitation and frequencies of the deletion in different populations.

The G-protein coupled receptor 4 (*GPR4*) gene is involved in sensing of the acidity of the cellular microenvironment. It is expressed in endothelium and is responsive to protons derived from carbonic and lactic acid, conferring potential functions in respiratory and metabolic acidosis (Hosford *et al.*

2018). *GPR4* has been linked to growth under thermal stress in rainbow trout (Yoshida & Yáñez 2022) suggesting interactions with the external environment.

The Cytochrome C Oxidase Assembly Factor (*COX20*) gene encodes is a transmembrane protein that acts as a chaperon during the assembly of mitochondrial cytochrome c oxidase (*COX*). *COX* is essential for aerobic energy generation (ATP), being the primary site of cellular oxygen consumption (Timón-Gómez *et al.* 2018). These two genes with roles in cellular respiration could be under divergent selection from varying migratory environmental conditions as the return spawning migration is extremely energetically costly (Lennox *et al.* 2018).

### SVs overlapping immune genes associated with environmental factors

A number of environmentally associated SVs disrupt the protein coding sequence of immune genes. For example, two immunoglobulin heavy variable genes (*IGHV*) that are part of the immunoglobulin heavy variable gene group, a key component of the adaptive immune response in jawed vertebrates including humans (Magadan *et al.* 2015; Mikocziova *et al.* 2021; Yasuike *et al.* 2010). A 13.7 kbp deletion encompassing the coding sequence of these genes significantly associates with multiple environmental parameters including latitude ( $-\log_{10}(P) = 6.51$ ) and temperature variables ( $-\log_{10}(P) = 6.57$ ) for annual mean temperature and  $-\log_{10}(P) = 4.36$  for mean temperature of warmest quarter. We find a 27.5 kbp long deletion overlapping the T Cell Receptor Alpha Variable (*TRAV*) gene, which is involved in antigen recognition (Attaf *et al.* 2015) and is significantly associated with latitude ( $-\log_{10}(P) = 7.05$ ). Fc Receptor Like 3 (*FCRL3*), which is linked to regulation of the immune system (Wang *et al.* 2021), overlaps a 1.1 kbp deletion polymorphism that is associated with drainage basin area ( $-\log_{10}(P) = 9.01$ ). A deletion overlapping V-Set Domain Containing T Cell Activation Inhibitor 1 (*VTCN1*) is significantly associated with drainage basin area ( $-\log_{10}(P) = 6.07$ ). *VTCN1* belongs to the B7 costimulatory protein family that is found on the surface of antigen-presenting cells that interact with T-cells to downregulate immune reactions by inhibiting T cell activation, proliferation and cytokine production (Vaishnav *et al.* 2022).

### SV-peaks associated with environmental factors

In addition to the SVs directly overlapping CDS considered above, we identified a number of 'SV-peaks' comprising multiple SVs in linkage disequilibrium, which significantly associate with environmental factors. Eleven of the most striking examples are shown in Figure S12-S13. SVs significantly associated with environmental variables within these peaks are listed in Data S6 for European and Data S7 for North American populations. These peaks did not contain obvious functional SVs overlapping protein coding sequences suggesting that linked variation (other SVs or SNPs), possibly co-localized with regulatory regions (promoters or enhancers), are more likely explanations for the environmental associations. However, predicting causal consequences of potentially regulatory SVs are not straightforward as it generally demands expression (RNA-Seq) and functional annotation (e.g. CHIP-Seq and ATAC-Seq) data from relevant tissues and samples to capture the regulatory landscape across individuals and environments. Unfortunately, currently the availability of relevant functional annotation data for Atlantic salmon limits such predictions about co-localization of SVs and regulatory elements.

## Discussion

### Pan-genomics detects a greatly expanded SV-catalogue for Atlantic salmon

Until recently, most population and landscape genomics studies have relied on a single linear reference genome sequence. However, there is increasing evidence that a large proportion of genomic variation is missed by this approach leading to the more sophisticated concept of the pan-genome (Eizenga *et al.* 2020; Tettelin *et al.* 2005). Here we present a comprehensive pan-genomic resource for Atlantic salmon, comprising 11 highly continuous long-read based assemblies from across the natural distribution. The reported salmon pan-genome was sampled across 4 phylogeographic groups resulting in the detection of a substantial proportion of genomic variation. By analysing the quantity of indels in the salmon pan-genome we confidently detected presence/absence variations (PAVs) accounting for as much as 367 Mbp (14.7% of the chromosome sequences), documenting that the salmon pan-genome size is substantially larger than what can be found in a single reference genome. The salmon pan-genome were used to reliably detect 1,061,452 SVs across the species range affecting an average of 77.4 Mbp per sample or ~3% of the genome. This is 69 fold higher than the number of SVs found by Bertolotti *et al.* (2020) in 492 fish with short-reads, but is generally comparable to other studies involving long-read sequencing of duplicated genomes. For example, the study of 29 accessions from wild and cultivated soybean revealed 776,399 SVs affecting ~16% of the paleopolyploid soybean genome of ~1.0 Gbp (Liu *et al.* 2020). In contrast, plant species with small, compact genomes without a recent polyploidization events, such as *Arabidopsis* (Jiao & Schneeberger 2017), tend to have less PAVs that can complement their reference genomes (Shi *et al.* 2022). There are few comparable long-read based SV studies conducted in fish species, but Mérot *et al.* (2022) found 194,861 SVs between two samples of salmonid Lake whitefish sp. (genome size ~2.68 Gbp). Estimates on SV numbers in human studies tend to be lower, generally in the range of 20-30 k as found in a recent trio study (Chaisson *et al.* (2019) and 3,622 Icelandic genomes (Beyter *et al.* 2021). In another study, including more diverse samples, they report more than 10% added sequence in an African pan-genome compared to a single reference genome (Sherman *et al.* 2019). Together these studies demonstrate the shortcomings of using single linear reference genome to capture the full genomic diversity of a species. The deficiencies become particularly evident for species that have experienced a recent whole genome duplications, as documented by data from polyploid plants (Hurgobin *et al.* 2018; Liu *et al.* 2020) and autotetraploid salmonid genomes like Lake whitefish (Mérot *et al.* 2022) and Atlantic salmon (this study).

The SV-detection strategy in our study is based on long-reads generated from 10 wild salmon sampled across a broad geographical distribution and four phylogeographic groups (ATL, BWS, BAL and NAM), aligned against a single reference genome (Ssal\_v3.1). Most of SVs in our study (59.6%) were found only in one sample, reflecting our SV-detection strategy with a limited number and highly diverging samples. Expanding the sample size is expected to substantially increase number SVs detected in multiple samples but will likely also provide additional SVs. The highest number of SVs per sample and phylogeographical group, were found in North American samples, reflecting the genetic distance between Ssal\_v3.1 (European) and the North American lineage, diverging more than 600,000 years before present (King *et al.* 2007). Mainly because of the genomic divergence, but also due to karyotype differences (Brenna-Hansen *et al.* 2012), Gao *et al.* (2022) argue for developing unique genomic resources for the North American lineage, including the genome sequence USDA\_NASsal\_1.1 (GCA\_021399835.1). This genome, developed from a male from the St. John River aquaculture strain

consists of 3,008 contigs (ctgN50 = 4.21 Mbp) anchored to 27 chromosomes, complements the four North American genome assemblies in our study (see Table S1), including the chromosome-anchored genome constructed from a male sampled in Garnish River on the south coast of Newfoundland and possessing a karyotype with 28 chromosomes.

### Repetitive DNA as a source of structural variation

Repeat elements are shown to contribute substantially to SVs in many species, particularly in plants where it has been investigated in many species (Lisch 2013). For example, in rice and tomato, TEs accounted for >17% of SVs and more than 75% of repeats in SVs, respectively (Alonge *et al.* 2020; Fuentes *et al.* 2019). A study of 26 soybean genomes revealed that ~78.5% of PAVs comes from repetitive sequences (Liu *et al.* 2020), and in grapevine they found that 90% of structural variants are repetitive elements, of which the TEs Gypsy (58.2%) and Copia (23.8%) are the most common (Di Genova *et al.* 2014). In concordance with these studies, we also find that repeats make up the majority of SVs, i.e. 52% of deletion sequence and 56% of insertions. In contrast to plants, we find an overall depletion of TEs in SV sequences compared to the genome wide average, suggesting that TEs are not the main cause of SVs in Atlantic salmon. An exception to this general pattern is a Tc1-Mariner element bearing close resemblance to the proposedly active DNA transposon first reported by de Boer *et al.* (2007). The DNA transposon overlaps a ~1400 bp fragment in both insertions and deletions and accounts for as much as 4.98% (38,882) of all deletions in our study. Aligning our findings with other salmonids, three different classes of TEs are shown to contribute significantly to the SV-landscape in rainbow trout (Liu *et al.* 2021). Like Atlantic salmon, the most marked overlap was found for a Tc1-Mariner but also a Gypsy retrotransposon and an unclassified TE-sequence were identified as substantial contributors. In the Lake whitefish, Mérot *et al.* (2020) found that repeats accounts for as much as 73% of SVs, involving four groups of active transposable elements, including Tc1-Mariner, Line-L2, Gypsy and ERV1.

Strikingly different from TEs, we find a strong and highly significant correlation between TRs and SVs in our study, implying that TRs have contributed substantially to the genomic SV-landscape of Atlantic salmon. Both TRs and SV were found to be highly enriched towards telomeric ends of chromosomes, a pattern also found in other species, including humans (Audano *et al.* 2019). The ancestor of Atlantic salmon experienced a whole genome duplication (Ss4R) event ~89-125 Mya (Gundappa *et al.* 2022), and linkage data (Lien *et al.* 2011) suggest that chromosome pairing and residual tetrasomic inheritance may still occur between duplicated regions, especially in males and between homeologous chromosome arms with high sequence similarity (see Allendorf *et al.* (2015)). We postulate that such an increased possibility for interchange between homeologs has contributed to the enhanced levels of TR-expansions observed towards telomeric ends of chromosomes in Atlantic salmon.

The Atlantic salmon karyotype has been generated by fusions of ancestral chromosomes, which in many cases translocate historical telomeres into centromeres or new intra-chromosomal positions. This provides unique prospects for studying the distribution of TRs and their contribution to the SV-landscape in both contemporary and historical telomeric regions. Our analyses revealed a highly significant expansion of TRs in both extant and historical telomeric regions, suggesting that most of the TR-expansions happened prior to chromosome fusions. Repeating the overlap analyses for SVs, we found that SV-enrichment is more profound in extant than historical telomeric regions (Figure 4),

suggesting that present telomeric regions have retained SVs while the variation in historic telomeres have been lost over time.

### Population scale genotyping of SVs with genome graphs

Our ability to detect different types of SVs has changed dramatically with the development of long-read sequencing technology, achieving high sensitivity and specificity by spanning both SVs and their flanking sequences (Sedlazeck *et al.* 2018a). However, although long-reads are superior to short-reads for SV-detection, they remain prohibitively expensive for population-scale applications due to their high costs and low throughput. As a consequence, population scale studies based on long-read data are rare in humans, but see Beyter *et al.* (2021), and non-existing for most other species. One possible solution to produce population scale data with SVs is to split the discovery and genotyping steps (Huddleston *et al.* 2017) and combine the technologies (i.e. use long-reads for detection and short-reads for genotyping). To achieve this, several variation-aware graph-based tools have been developed; e.g. Graphtyper2 (Eggertsson *et al.* 2019), Variation graph (Garrison *et al.* 2018) and Paragraph (Chen *et al.* 2019).

In this study, we used the Graphtyper2 software package, together with readily available short-read data from wild 366 salmon to genotype 1,056,706 indels initially discovered by long-reads. The analyses yielded genotypes for 672,404 (63.6%) SVs. After filtering the data for minor allele frequency (MAF>0.05) and max-missingness (0.3), we retained reliable genotypes from 304,407 SVs (28.7%). A similar proportion of successfully genotyped SVs was found in 276 humans using Paragraph, with 14,204 of 38,028 or ~37% of SVs (Quan *et al.* (2021)). The majority of the SVs called in our study (97.6%) were relatively short deletions reflecting the enhanced ability of this method to genotype short deletions compared to insertions (Almarri *et al.* 2020; Eggertsson *et al.* 2019). The fact that only 2.4% of the SVs genotyped with short-reads were insertions, compared to 26% in the long-read detection set, exemplifies that insertions are still largely inaccessible using this approach. As previously suggested by (Eggertsson *et al.* 2019), breakpoint inaccuracy is a likely explanation to this. While deletions are relatively straightforward to represent with flanking sequence in the reference, insertions add sequences absent in the reference, complicating breakpoint representation. Somewhat in conflict with this, Eggertsson *et al.* (2019) found more insertions (median 13,353 per sample) than deletions (median 9,474 per sample) using Graphtyper2 in a set of 3,622 Icelanders. A possible explanation to this may be that they employed additional breakpoint refinement steps in their pipeline, which might improve genotyping of insertions.

Despite the fact that many SVs were impossible to genotype using the short-read based genome-graph approach used in our study, we were able to successfully genotype 304,407 SVs (MAF>0.05) in 366 wild Atlantic salmon sampled from a wide phylogeographical range, including SVs that overlapped the CDS of genes. The fact that short-read data were readily available for the wild salmon makes the approach particularly cost-effective. Utilising the SVs in PCA and NGSadmixture analyses revealed population structures echoing results in earlier studies (Bertolotti *et al.* 2020; Bourret *et al.* 2013), increasing our confidence that the data are of high quality.



## Structural variants contribute to environmental adaptation

We identified SVs associated with environmental variables measuring thermal conditions, precipitation, and the river size of spawning habitats for Atlantic salmon. Gene ontology enrichment analysis of SVs associated with environment variables and gene overlap revealed a predominance of neurological biological processes. An increasing number of studies suggest that SVs of all types and sizes may have a large effect on phenotype and consequently major impact on rapid adaptation and population divergence (Hämälä *et al.* 2021; Liu *et al.* 2020). However, the functional effect of the vast majority of SVs is unknown and studies generally lacks evidence on the phenotypic consequences of most SVs that are suggested to have adaptive potential. SVs overlapping protein coding sequences are likely to cause direct functional effects and these are likely to be deleterious if they disrupt gene function (Alonge *et al.* 2020; Chiang *et al.* 2017). Moreover, SVs that cause copy number variation, by duplicating or deleting whole genes, and those impacting regulatory regions may be adaptive and may be important in local adaptation as they are likely to have large effects and, therefore, may underlie balanced polymorphisms (Escaramís *et al.* 2015). For example, in a recent SV-study of natural population of chocolate trees (*Theobroma cacao*) by Hämälä *et al.* (2021) found that most SVs have detrimental effects, but they also found several SVs bearing signals of local adaptation and having positive fitness effects.

Several SVs that were found to be involved in environmental adaptation overlapped immune genes. This result is in line with previous evidence that immune genes are the targets of differential selection among populations and lineages in salmon (Kjærner-Semb *et al.* 2016; Perrier *et al.* 2017) and other species including humans (Yan *et al.* 2021). We find SVs overlapping genes within a immunoglobulin heavy chain (*IgH*) locus, which is a hypervariable region in humans including multiple adaptive SVs (Mikocziova *et al.* 2021; Yan *et al.* 2021), including copy number variation (Watson *et al.* 2013) that is linked to local adaptation in immunity (Yan *et al.* 2021). This region is known to be duplicated, hyperdiverse and tandemly repeating in Atlantic salmon (Yasuike *et al.* 2010) and associated with temperature variation (Perrier *et al.* 2017). Structural variation at this locus is linked to the process of antibody production via somatic variable-(diversity)-joining, V(D)J, recombination and hypermutation. Resolving variation in this complex region has been challenging using short-read methods (Gao *et al.* 2021). It is therefore a notable example of adaptively important variation that can be detected in long-read assemblies coupled with population scale genotyping using short-reads and genome graphs. Variation at this locus is important for environmental adaptation in wild salmon but may also be included in precision breeding and vaccine development to prevent diseases in aquaculture.

SVs exhibiting environmental associations may not themselves be causal but rather tag nearby causal variation in close linkage disequilibrium (LD). Pinpointing the functional variants in 'peaks' of variants in LD is difficult, as the underlying causal variant may not be included in the study, and associations may be caused by complex interactions of several genes and/or variation in regulatory regions with more subtle effects on the gene expression (Alonge *et al.* 2020). Therefore, future investigations of the functional role of SVs on environmental adaptation would benefit from including expression data (RNA-Seq) and functional annotation (e.g. from ChIP-Seq and ATAC-Seq), as well as expanding the number of samples from contrasting environments.

Today, both wild and aquaculture populations of Atlantic salmon are threatened by increased infectious disease and parasites burdens and changes in life-history, including reduced variation in age at maturity

in the wild and increased precocious maturation in aquaculture, all of which are likely to be exacerbated by climate change (Jonsson & Jonsson 2009; Mobley *et al.* 2021; Thorstad *et al.* 2021). Understanding the genetic factors underpinning environmental adaptive variation is essential for predicting the adaptability of natural populations and their future resilience to anthropogenic changes. So far, genomics approaches investigating these issues have been based on the use of SNPs. Although providing fundamental new knowledge regarding trait variation in wild salmon (e.g. (Ayllon *et al.* 2015; Barson *et al.* 2015; Kess *et al.* 2022; Lehnert *et al.* 2019), as well as being instrumental for continuous improvement of traits in aquaculture (Houston *et al.* 2020), it remains clear that much genomic variation remains unexplored by this approach. It is now broadly accepted that inversion polymorphisms can create ‘supergenes’ contributing to adaptive evolution and species diversification (see Wellenreuther *et al.* (2019)). This study document widespread contribution of other types of SVs to environmental adaptation. By combining long read and graph genotyping of short-read sequencing we characterise genomic variation in previously inaccessible genomic regions, unlocking this structural variation for population studies, and revealing that SVs contribute more to adaptive evolution than previously perceived.

## Material and Methods

### Sample collection and DNA isolation

Fresh blood for long read sequencing and liver for Pore-C library preparation were collected from 11 Atlantic salmon individuals representing five different phylogeographical groups from across the Atlantic Ocean (Table S1). Blood samples were collected into EDTA containing vacuum tubes while liver tissue was finely minced. For both sample types, material was swiftly frozen using dry-ice and stored at -80 until DNA extraction. Blood was thawed on ice and high molecular weight DNA was isolated using the Nanobind CBB kit (PacBio, USA). To reduce sample viscosity and make pipetting possible, samples were gently passaged through a 25G blunt-end needle 8-14 times. Fragments smaller than 25kb were progressively depleted using the Short Read Eliminator (PacBio USA). Final DNA concentration was determined using Qubit (ThermoScientific USA), its purity was assessed using NanoDrop, and its integrity and size subjectively assessed using agarose gel electrophoresis.

### Oxford Nanopore long-read sequencing

Libraries were prepared using the Sequencing by Ligation kit from Oxford Nanopore (SQK-LSK109) and were sequenced on PromethION flowcells (R9.4.1) using a PromethION Beta instrument. DNA from Atlantic salmon have a high propensity to cause pore-blockage. Therefore, to increase the sequencing output per flow cell, several (3-5) libraries were prepared and pooled for each individual, flow cells were nuclease flushed when the number of sequencing pores dropped below 15% and library was reloaded. By running multiple flow-cells and performing repeated reloading (3 to 5 times/flow cell) we were able to generate the yields of data reported in Table S2.

### Illumina sequencing

Genomic DNA from the 11 selected individuals was send to a commercial sequencing provider (Novogene, UK) for library preparation and sequencing using an Illumina NovaSeq 6000 and S4 flow-

cell. A minimum of 40X coverage (approximately 120Gb) raw data (PE150) was generated for each sample.

### Chromosome conformation capture sequencing

Three of the samples, representing aquaculture (AQGE), North American (GARN-1) and Barents/White sea (ALTA) populations, were selected for chromatin-contact based scaffolding using Pore-C (Deshpande *et al.* 2022) or Hi-C sequencing (Lieberman-Aiden *et al.* 2009). The Hi-C library was prepared for individual AQGE from liver and kidney tissue using the Qiagen EpiTect Hi-C kit and paired-end sequenced on an Illumina NovaSeq 600 machine for 150 cycles. Hi-C data was mapped to the AQGE genome assembly using the juicer pipeline (v 1.5.7, (Durand *et al.* 2016)). Contigs were scaffolded, visually inspected and rearranged with 3D-DNA according to the contacts (v 180419, (Dudchenko *et al.* 2017)) and JuiceBox (v 1.11.08, (Durand *et al.* 2016)).

Pore-C libraries for ALTA and GARN-1 were generated following the RE-Pore-C protocol using SQK-LSK109 Ligation kit from Oxford Nanopore. Libraries were loaded on PromethION flowcells (R9.4.1) and sequenced using a PromethION beta device. Pore-C data was processed using Pore-C-Snakemake pipeline (v 0.3.0) to generate contact matrices. Identical to Hi-C data processing, contigs from ALTA and GARN-1 were scaffolded, visually inspected and rearranged with 3D-DNA (v 180419, (Dudchenko *et al.* 2017)) and JuiceBox (v 1.11.08, (Durand *et al.* 2016)).

### Genome assemblies

The Atlantic salmon reference genome Ssal\_v3.1 (GCA\_905237065.2) was constructed from a male Norwegian aquaculture salmon (AquaGen) using 70x long-read Oxford Nanopore reads filtered on length (>4k) and quality (q>7) with fastp (v 0.19.5, (Chen *et al.* 2018)). Initially, five *de novo* assemblies were generated with varying sequence overlap (5, 10, 15, 20 and 30kb) using Flye v2.7 and v2.8 (Kolmogorov *et al.* 2019). Contigs from the five assemblies were combined into one assembly by merging contig ends overlapping with 20 kbp or more determined from LASTZ alignments (Harris 2007). The combined assembly was polished with long-reads using PEPPER (v 0.0.6, (Shafin *et al.* 2021)) and Illumina short-reads using pilon (v 1.23, (Walker *et al.* (2014))). The same pipeline was used for the other 10 genome assemblies. Sequence overlap used is listed in Table S3.

### Analysing homeologous (duplicated) blocks within the salmon genome

Repeat masked chromosome sequences for Atlantic salmon (TR + TE libraries from Ssal\_v3.1, see sections below) were aligned against each other using LASTZ (Harris 2007) to disentangle conserved collinear blocks of homeology. LASTZ command line script; --targetcapsule=LZ\_target\_capsule query.fa --nochain --gextend --nogapped --identity=75.0..100.0 --matchcount=100 --format=general --rdotplot=plotoutput.txt. Sequence similarity between homeologous sequences were determined in 1 Mb intervals by averaging local percentage of nucleotide sequence identity using high-scoring segment pair (HSP) from LASTZ alignments (Harris 2007) and presented as a Circos plot in Figure S4.

### Transposable element annotation

A library of TE consensus sequences for Atlantic salmon was already available from the ICSASG\_v2 assembly (Lien *et al.* 2016). However, since the long-read-based assembly is likely to detect additional TE-families, we decided to make a new annotation on the Ssal\_v3.1 assembly. To this end, we used three *de novo* pipelines to generate TE libraries: RepeatModeler2 (Flynn *et al.* 2020), REPET's (Flutre *et al.* 2011) TEdenovo and PASTEC (Hoede *et al.* 2014) suites, and a merged EDTA/DeepTE method (Bell *et al.* 2021). Each of these libraries was reciprocally BLASTed using BLAST+/2.10.1 (Camacho *et al.* 2009), masked using RepeatMasker (Smit *et al.* 2015), and grouped according to the 80-80-80 rule of thumb (i.e., 80% similarity over 80% of the sequence down to at least 80 bp) (Wicker *et al.* 2007). Every extant library entry was compared to the *de novo* libraries and corrected if there was consensus among the automated classifications that the sequence was misclassified. In addition, every satellite DNA or "simple repeat" entry was removed, and finally well-characterised sequences missing in the previously extant library but detected by at least two *de novo* methods had their most reasonable-looking consensus added to the library.

### Tandem repeat annotation

To annotate the tandem repeat content of Ssal\_v3.1 we produced a library of satellite DNA consensus sequences from TAREAN (v 2.3.7) (Novák *et al.* 2017), a part of the RepeatExplorer pipeline (Novák *et al.* 2013). The TAREAN library was merged with output of Tandem Repeat Finder (Benson 1999) filtered to take arrays of at least 10 kb and maximum period size 2 kb after a reciprocal RepeatMask run to filter out any redundancy. This was then annotated using RepeatMasker (Smit *et al.* 2015) on default settings and masked again using our TE library. The reason for this approach is that the previous TE library has some known issues of misannotating satDNA as LTR transposons, and for the purpose of our analyses it is preferable to err on the side of satDNA.

### Genome synteny plot

ICSASG\_v2 and Ssal\_v3.1 were masked with the TR library using RepeatMasker (v 4.1.1, (Smit *et al.* 2015)) and aligned using minimap2 (v 2.18-r1015, (Li 2018)) with the -ax asm5 and --eqx flags. Synteny was computed using syri (v 1.6, (Goel *et al.* 2019)) with default settings, and plotted using plotsr (v 0.5.4, (Goel & Schneeberger 2022)) with the minimum size of a syntenic region to be plotted (-s) set to 50,000.

### Read alignment and SV calling

To detect SVs, we mapped long-reads to Ssal\_v3.1 using Winnowmap2 (v 2.0, (Jain *et al.* 2020)) with the "--MD" flag to better resolve repetitive regions of the genome. Sam-files were sorted and converted into BAM-files with samtools V1.3.1 (Li *et al.* 2009). The SV-detection was performed with three long-read specific SV calling programs: Sniffles (v 1.0.12, (Sedlazeck *et al.* 2018b)), SVIM (v 1.2.0 (Heller & Vingron 2019)) and NanoVar (v 1.3.9, (Tham *et al.* 2020)) with default settings for SVIM and NanoVar. To account for the variable read depth, the minimum number of reads that support a SV to be reported (-s) was set to 1/3 of the median read depth, as suggested by (De Coster *et al.* 2019), calculated using Mosdepth (v 0.2.6, (Pedersen & Quinlan 2018)) when running Sniffles. SVs of type breakpoint (BND) were removed, and other excess information in the VCF files were filtered out using custom R

scripts, available at [https://github.com/kristinastenlokk/long\\_read\\_SV](https://github.com/kristinastenlokk/long_read_SV). We kept SVs detected by at least two callers after merging with Jasmine (v 1.1.0, (Kirsche *et al.* 2021)) including refinement of insertion sequences with Iris “max\_dist\_linear=0.1 min\_dist=50 --default\_zero\_genotype --mutual\_distance min\_support=2 --output\_genotypes --normalize\_type --run\_iris iris\_args=--keep\_long\_variants”.

### Short-read alignment

To expand our dataset, we mapped short-read from Atlantic salmon sampled from a broad phylogeographic distribution. The original dataset presented by Bertolotti *et al.* (2020) is larger, but we removed aquaculture individuals and samples sequenced with the Hiseq2500 sequencing platform as these showed a clear ‘machine effect’. The 356 remaining samples and the 10 samples from our original SV dataset (Data S2) were aligned to Ssal\_v3.1 with bwa-mem2 (v 2.2.1, (Vasimuddin *et al.* 2019)). Duplicate reads were masked with samblaster (v 0.1.26, (Faust & Hall 2014)), and files were sorted and converted into cram-format with samtools (v 1.11, (Li *et al.* 2009)).

### Graph genotyping

By genotyping high confidence SVs detected by long-reads into a population scale short-read sequenced dataset, we can harness the accuracy and recall of long-reads and the numbers of pre-existing short-read data. We mapped a whole genome re-sequenced dataset consisting of 366 Atlantic salmon. Indels between 50 bp and 50 kbp were genotyped using GraphTyper2 (v 2.7.5, (Eggertsson *et al.* 2019)) with default settings. Only SVs tagged with the genotype model “aggregated” were kept, according to the software developer’s recommendations. GraphTyper2 employs a strict coverage filter, as our short-read data was relatively low coverage we instead used map 5% and missingness 30% to quality filter the variant calls with VCFtools (v 0.1.16, (Danecek *et al.* 2011)). VCFtools was also used to calculate allele frequencies.

### Population structure analysis

To check population structure, we performed a PCA using the prcomp function in the stats package (v 4.2.2) in R (v 3.15, (Team 2013)). The population structure was also estimated using NGSadmix (v 32, (Skotte *et al.* 2013)) with K=3-16, with optimal K being 7 (K=5 in Europe and K=2 in North America).

### Genotype-environment association

SVs may have negative fitness effects, especially those in protein coding regions and are, therefore, subject to negative selection, however, SVs can also create adaptive differences that are positively selected. To identify adaptive SVs, we tested for genotype-environment associations (GEA) determined using the Latent Factor Mixed Model (LFMM) approach (Frichot *et al.* 2013) in the R package lea (R v 4.1, (Frichot & François 2015)). LFMM fits a linear mixed model with population structure controlled simultaneously to model estimation using latent factors, where the expected number of genetic clusters (K) is the latent factor, which was estimated using NGSadmix (v 32, (Skotte *et al.* 2013)). Environment associations were tested on the pooled European (n=276) and the North American (n = 80) samples separately because the strong differentiation between these lineages would confound associations and

some SVs were only present, or had MAF>5%, in one group (34,776 unique to Europe and 32,415 unique to Canada). Environment associations were tested for all SVs (N =271,174 and 270,057 for Europe and N. American respectively). False discovery control was employed using the Benjamini-Hochberg (Hochberg & Benjamini 1990) procedure with alpha thresholds of 0.05 and 0.01 across all tests. Environment variables tested related to thermal, precipitation and river size. The individual river parameters were obtained from the WorldClim2 (Frichot & François 2015) and hydrobasins (Lehner & Grill 2013) databases for an arc of 30 translating to 1 km<sup>2</sup> at the river mouth (<https://www.worldclim.org>) to ensure comparable data quality and availability for all rivers. Air temperature has been shown to represent water temperature in Norway except at low temperatures (Otero *et al.* 2014), likely because winter ice cover in some rivers can lead to discrepancies in air and water temperatures. Annual temperature, and additionally the temperature in the coldest and warmest quarters, were selected as they influence the overwinter survival and growth potential respectively.

### Gene ontology and KEGG enrichment

To test for overrepresentation of genes functionally affected by SVs in biological processes we used both gene ontology (GO) and KEGG enrichment analysis. The genes effected by SVs were defined as any gene overlapped by an SV that was significantly associated with an environment by LFMM. The background genes were set to be any genes with at least one SV within 2 kbp distance to compensate for SV density and distribution and therefore the capacity for an association to have been detected if present. For both GO and KEGG we used the R package clusterProfiler (v 4.4.4, (Wu *et al.* 2021)) following the method described here: <https://gitlab.com/sandve-lab/salmon-go-and-kegg-enrichment>.

### Acknowledgement

We thank Sarah Lehnert, Ian Bradbury, Louis Bernatchez, Cooke Aquaculture Inc., AquaGen AS, Craig Primmer, Jamie Stevens, Harald Sægrov, Jenny Jensen and Thronnd Haugen for providing samples for the nanopore sequencing. We acknowledge the use of the Orion computing cluster at the Norwegian University of Life Sciences (NMBU). Storage resources were provided by the Norwegian National Infrastructure for Research Data (NIRD, project NS9055 K). The study was supported by The Research Council of Norway (grant nos. 275310 and 221734).

### References

- Alkan, C., Coe, B. P. & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, 12 (5): 363-376.
- Allendorf, F. W., Bassham, S., Cresko, W. A., Limborg, M. T., Seeb, L. W. & Seeb, J. E. (2015). Effects of crossovers between homeologs on inheritance and population genomics in polyploid-derived salmonid fishes. *Journal of Heredity*, 106 (3): 217-227.
- Almarri, M. A., Bergström, A., Prado-Martinez, J., Yang, F., Fu, B., Dunham, A. S., Chen, Y., Hurles, M. E., Tyler-Smith, C. & Xue, Y. (2020). Population structure, stratification, and introgression of human structural variation. *Cell*, 182 (1): 189-199. e15.

- Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., Suresh, H., Ramakrishnan, S., Maumus, F., Ciren, D., *et al.* (2020). Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell*, 182 (1): 145-161.e23.
- Andersson, E., Schulz, R. W., Male, R., Bogerd, J., Patiña, D., Benedet, S., Norberg, B. & Taranger, G. L. (2013). Pituitary gonadotropin and ovarian gonadotropin receptor transcript levels: Seasonal and photoperiod-induced changes in the reproductive physiology of female Atlantic salmon (*Salmo salar*). *General and Comparative Endocrinology*, 191: 247-258.
- Ando, H. & Urano, A. (2005). Molecular regulation of gonadotropin secretion by gonadotropin releasing hormone in salmonid fishes. *Zoological Science*, 22 (4): 379-389.
- Attaf, M., Legut, M., Cole, D. K. & Sewell, A. K. (2015). The T cell antigen receptor: the Swiss army knife of the immune system. *Clinical & Experimental Immunology*, 181 (1): 1-18.
- Audano, P. A., Sulovari, A., Graves-Lindsay, T. A., Cantsilieris, S., Sorensen, M., Welch, A. E., Dougherty, M. L., Nelson, B. J., Shah, A., Dutcher, S. K., *et al.* (2019). Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell*, 176 (3): 663-675.e19.
- Ayllon, F., Kjærner-Semb, E., Furmanek, T., Wennevik, V., Solberg, M. F., Dahle, G., Taranger, G. L., Glover, K. A., Almén, M. S., Rubin, C. J., *et al.* (2015). The vgl3 Locus Controls Age at Maturity in Wild and Domesticated Atlantic Salmon (*Salmo salar* L.) Males. *PLOS Genetics*, 11 (11): e1005628.
- Barson, N. J., Aykanat, T., Hindar, K., Baranski, M., Bolstad, G. H., Fiske, P., Jacq, C., Jensen, A. J., Johnston, S. E., Karlsson, S., *et al.* (2015). Sex-dependent dominance at a single locus maintains variation in age at maturity in salmon. *Nature*, 528 (7582): 405-408.
- Bell, E. A., Butler, C. L., Oliveira, C., Marburger, S., Yant, L. & Taylor, M. I. (2021). Transposable element annotation in non-model species: The benefits of species-specific repeat libraries using semi-automated EDTA and DeepTE de novo pipelines. *Molecular Ecology Resources*.
- Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research*, 27 (2): 573-580.
- Beroukhi, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J. S., Dobson, J. & Urashima, M. (2010). The landscape of somatic copy-number alteration across human cancers. *Nature*, 463 (7283): 899-905.
- Bertolotti, A. C., Layer, R. M., Gundappa, M. K., Gallagher, M. D., Pehlivanoglu, E., Nome, T., Robledo, D., Kent, M. P., Røsæg, L. L., Holen, M. M., *et al.* (2020). The structural variation landscape in 492 Atlantic salmon genomes. *Nature Communications*, 11 (1): 5176.
- Beyter, D., Ingimundardottir, H., Oddsson, A., Eggertsson, H. P., Bjornsson, E., Jonsson, H., Atlason, B. A., Kristmundsdottir, S., Mehringer, S., Hardarson, M. T., *et al.* (2021). Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nature Genetics*, 53 (6): 779-786.
- Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H. L., Macfarlan, T. S., *et al.* (2018). Ten things you should know about transposable elements. *Genome Biology*, 19 (1): 199.
- Bourret, V., Kent, M. P., Primmer, C. R., Vasemägi, A., Karlsson, S., Hindar, K., McGinnity, P., Verspoor, E., Bernatchez, L. & Lien, S. (2013). SNP-array reveals genome-wide patterns of geographical and potential adaptive divergence across the natural range of Atlantic salmon (*Salmo salar*). *Molecular Ecology*, 22 (3): 532-551.

- Brenna-Hansen, S., Li, J., Kent, M. P., Boulding, E. G., Dominik, S., Davidson, W. S. & Lien, S. (2012). Chromosomal differences between European and North American Atlantic salmon discovered by linkage mapping and supported by fluorescence in situ hybridization analysis. *BMC Genomics*, 13 (1): 432.
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, 10 (1): 421.
- Chaisson, M. J. P., Sanders, A. D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E. J., Rodriguez, O. L., Guo, L., Collins, R. L., *et al.* (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications*, 10 (1): 1784.
- Chen, S., Zhou, Y., Chen, Y. & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34 (17): i884-i890.
- Chen, S., Krusche, P., Dolzhenko, E., Sherman, R. M., Petrovski, R., Schlesinger, F., Kirsche, M., Bentley, D. R., Schatz, M. C., Sedlazeck, F. J., *et al.* (2019). Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biology*, 20 (1): 291.
- Cheong, H. I., Asosingh, K., Stephens, O. R., Queisser, K. A., Xu, W., Willard, B., Hu, B., Dermawan, J. K. T., Stark, G. R. & Prasad, S. V. N. (2016). Hypoxia sensing through  $\beta$ -adrenergic receptors. *JCI insight*, 1 (21).
- Chiang, C., Scott, A. J., Davis, J. R., Tsang, E. K., Li, X., Kim, Y., Hadzic, T., Damani, F. N., Ganel, L., Montgomery, S. B., *et al.* (2017). The impact of structural variation on human gene expression. *Nature Genetics*, 49 (5): 692-699.
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., *et al.* (2011). The variant call format and VCFtools. *Bioinformatics*, 27 (15): 2156-2158.
- de Boer, J. G., Yazawa, R., Davidson, W. S. & Koop, B. F. (2007). Bursts and horizontal evolution of DNA transposons in the speciation of pseudotetraploid salmonids. *BMC Genomics*, 8 (1): 422.
- De Coster, W., De Rijk, P., De Roeck, A., De Pooter, T., D'Hert, S., Strazisar, M., Slegers, K. & Van Broeckhoven, C. (2019). Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. *Genome research*, 29 (7): 1178-1187.
- Deshpande, A. S., Ulahannan, N., Pendleton, M., Dai, X., Ly, L., Behr, J. M., Schwenk, S., Liao, W., Augello, M. A., Tyer, C., *et al.* (2022). Identifying synergistic high-order 3D chromatin conformations from genome-scale nanopore concatemer sequencing. *Nature Biotechnology*, 40 (10): 1488-1499.
- Di Genova, A., Almeida, A. M., Muñoz-Espinoza, C., Vizoso, P., Travisany, D., Moraga, C., Pinto, M., Hinrichsen, P., Orellana, A. & Maass, A. (2014). Whole genome comparison between table and wine grapes reveals a comprehensive catalog of structural variants. *BMC Plant Biology*, 14 (1): 7.
- Dobzhansky, T. (1947). Adaptive changes induced by natural selection in wild populations of *Drosophila*. *Evolution*: 1-16.
- Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., Shamim, M. S., Machol, I., Lander, E. S., Aiden, A. P., *et al.* (2017). De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*, 356 (6333): 92-95.



- Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S. & Aiden, E. L. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell systems*, 3 (1): 95-98.
- Eggertsson, H. P., Kristmundsdottir, S., Beyter, D., Jonsson, H., Skuladottir, A., Hardarson, M. T., Gudbjartsson, D. F., Stefansson, K., Halldorsson, B. V. & Melsted, P. (2019). GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nature Communications*, 10 (1): 5402.
- Eizenga, J. M., Novak, A. M., Sibbesen, J. A., Heumos, S., Ghaffaari, A., Hickey, G., Chang, X., Seaman, J. D., Rounthwaite, R., Ebler, J., *et al.* (2020). Pangenome Graphs. *Annual Review of Genomics and Human Genetics*, 21 (1): 139-162.
- Escaramís, G., Docampo, E. & Rabionet, R. (2015). A decade of structural variants: description, history and methods to detect structural variation. *Briefings in Functional Genomics*, 14 (5): 305-314.
- Eslami Rasekh, M., Hernández, Y., Drinan, S. D., Fuxman Bass, J. I. & Benson, G. (2021). Genome-wide characterization of human minisatellite VNTRs: population-specific alleles and gene expression differences. *Nucleic Acids Research*, 49 (8): 4308-4324.
- Farnoud, F., Schwartz, M. & Bruck, J. (2019). Estimation of duplication history under a stochastic model for tandem repeats. *BMC Bioinformatics*, 20 (1): 64.
- Faust, G. G. & Hall, I. M. (2014). SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics*, 30 (17): 2503-2505.
- Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H. (2011). Considering Transposable Element Diversification in De Novo Annotation Approaches. *PLOS ONE*, 6 (1): e16526.
- Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C. & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*, 117 (17): 9451.
- Frichot, E., Schoville, S. D., Bouchard, G. & François, O. (2013). Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular biology and evolution*, 30 (7): 1687-1699.
- Frichot, E. & François, O. (2015). LEA: An R package for landscape and ecological association studies. *Methods in Ecology and Evolution*, 6 (8): 925-929.
- Fuentes, R. R., Chebotarov, D., Duitama, J., Smith, S., De la Hoz, J. F., Mohiyuddin, M., Wing, R. A., McNally, K. L., Tatarinova, T. & Grigoriev, A. (2019). Structural variants in 3000 rice genomes. *Genome research*, 29 (5): 870-880.
- Gao, G., Magadan, S., Waldbieser, G. C., Youngblood, R. C., Wheeler, P. A., Scheffler, B. E., Thorgaard, G. H. & Palti, Y. (2021). A long reads-based de-novo assembly of the genome of the Arlee homozygous line reveals chromosomal rearrangements in rainbow trout. *G3 Genes/Genomes/Genetics*, 11 (4): jkab052.
- Gao, Y., Ma, L. & Liu, G. E. (2022). Initial Analysis of Structural Variation Detections in Cattle Using Long-Read Sequencing Methods. *Genes*, 13 (5).
- Garrido-Ramos, M. A. (2017). Satellite DNA: An Evolving Topic. *Genes*, 8 (9): 230.
- Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., Jones, W., Garg, S., Markello, C., Lin, M. F., *et al.* (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, 36 (9): 875-879.

- Gillard, G. B., Grønvold, L., Røsæg, L. L., Holen, M. M., Monsen, Ø., Koop, B. F., Rondeau, E. B., Gundappa, M. K., Mendoza, J., Macqueen, D. J., *et al.* (2021). Comparative regulomics supports pervasive selection on gene dosage following whole genome duplication. *Genome Biology*, 22 (1): 103.
- Godin, J.-G. J. & Rangeley, R. W. (1989). Living in the fast lane: effects of cost of locomotion on foraging behaviour in juvenile Atlantic salmon. *Animal Behaviour*, 37: 943-954.
- Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. (2019). SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biology*, 20 (1): 277.
- Goel, M. & Schneeberger, K. (2022). plotsr: visualizing structural similarities and rearrangements between multiple genomes. *Bioinformatics*, 38 (10): 2922-2926.
- Gundappa, M. K., To, T.-H., Grønvold, L., Martin, S. A. M., Lien, S., Geist, J., Hazlerigg, D., Sandve, S. R. & Macqueen, D. J. (2022). Genome-Wide Reconstruction of Rediploidization Following Autopolyploidization across One Hundred Million Years of Salmonid Evolution. *Molecular Biology and Evolution*, 39 (1): msab310.
- Guo, J., Cao, K., Deng, C., Li, Y., Zhu, G., Fang, W., Chen, C., Wang, X., Wu, J., Guan, L., *et al.* (2020). An integrated peach genome structural variation map uncovers genes associated with fruit traits. *Genome Biology*, 21 (1): 258.
- Hämälä, T., Wafula, E., Guiltinan, M., Ralph, P., dePamphilis, C. & Tiffin, P. (2021). Genomic structural variants constrain and facilitate adaptation in natural populations of *Theobroma cacao*, the chocolate tree. *Proceedings of the National Academy of Sciences*, 118: e2102914118.
- Harris, R. S. (2007). *Improved pairwise alignment of genomic DNA*: The Pennsylvania State University.
- Heller, D. & Vingron, M. (2019). SVIM: structural variant identification using mapped long reads. *Bioinformatics*, 35 (17): 2907-2915.
- Ho, S. S., Urban, A. E. & Mills, R. E. (2020). Structural variation in the sequencing era. *Nature Reviews Genetics*, 21 (3): 171-189.
- Hochberg, Y. & Benjamini, Y. (1990). More powerful procedures for multiple significance testing. *Statistics in medicine*, 9 (7): 811-818.
- Hoede, C., Arnoux, S., Moisset, M., Chaumier, T., Inizan, O., Jamilloux, V. & Quesneville, H. (2014). PASTEC: An Automatic Transposable Element Classification Tool. *PLOS ONE*, 9 (5): e91929.
- Hosford, P., Mosienko, V., Kishi, K., Jurisic, G., Seuwen, K., Kinzel, B., Ludwig, M., Wells, J., Christie, I. & Koolen, L. (2018). CNS distribution, signalling properties and central effects of G-protein coupled receptor 4. *Neuropharmacology*, 138: 381-392.
- Houston, R. D., Bean, T. P., Macqueen, D. J., Gundappa, M. K., Jin, Y. H., Jenkins, T. L., Selly, S. L. C., Martin, S. A. M., Stevens, J. R., Santos, E. M., *et al.* (2020). Harnessing genomics to fast-track genetic improvement in aquaculture. *Nature Reviews Genetics*, 21 (7): 389-409.
- Huddleston, J., Chaisson, M. J., Steinberg, K. M., Warren, W., Hoekzema, K., Gordon, D., Graves-Lindsay, T. A., Munson, K. M., Kronenberg, Z. N. & Vives, L. (2017). Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome research*, 27 (5): 677-685.
- Hurgobin, B., Golicz, A. A., Bayer, P. E., Chan, C.-K. K., Tirnaz, S., Dolatabadian, A., Schiessl, S. V., Samans, B., Montenegro, J. D., Parkin, I. A. P., *et al.* (2018). Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid *Brassica napus*. *Plant Biotechnology Journal*, 16 (7): 1265-1274.

- Jain, C., Rhie, A., Hansen, N., Koren, S. & Phillippy, A. M. (2020). A long read mapping method for highly repetitive reference sequences. *bioRxiv*: 2020.11.01.363887.
- Jeffreys, A. J., Neil, D. L. & Neumann, R. (1998). Repeat instability at human minisatellites arising from meiotic recombination. *The EMBO Journal*, 17 (14): 4147-4157.
- Jiao, W.-B. & Schneeberger, K. (2017). The impact of third generation genomic technologies on plant genome assembly. *Current Opinion in Plant Biology*, 36: 64-70.
- Jonsson, B. & Jonsson, N. (2009). A review of the likely effects of climate change on anadromous Atlantic salmon *Salmo salar* and brown trout *Salmo trutta*, with particular reference to water temperature and flow. *Journal of Fish Biology*, 75 (10): 2381-2447.
- Kent, T. V., Uzunović, J. & Wright, S. I. (2017). Coevolution between transposable elements and recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372 (1736): 20160458.
- Kess, T., Lehnert, S. J., Bentzen, P., Duffy, S., Messmer, A., Dempson, J. B., Newport, J., Whidden, C., Robertson, M. J. & Chaput, G. (2022). Parallel genomic basis of age at maturity across spatial scales in Atlantic Salmon. *bioRxiv*.
- King, H. & Pankhurst, N. (2004). Effect of maintenance at elevated temperatures on ovulation and luteinizing hormone releasing hormone analogue responsiveness of female Atlantic salmon (*Salmo salar*) in Tasmania. *Aquaculture*, 233 (1-4): 583-597.
- King, T. L., Verspoor, E., Spidle, A. P., Gross, R., Phillips, R. B., Koljonen, M. L., Sanchez, J. A. & Morrison, C. L. (2007). Biodiversity and Population Structure. In *The Atlantic Salmon*, pp. 117-166.
- Kirsche, M., Prabhu, G., Sherman, R., Ni, B., Aganezov, S. & Schatz, M. C. (2021). Jasmine: Population-scale structural variant comparison and analysis. *bioRxiv*: 2021.05.27.445886.
- Kjærner-Semb, E., Ayllon, F., Furmanek, T., Wennevik, V., Dahle, G., Niemelä, E., Ozerov, M., Vähä, J.-P., Glover, K. A., Rubin, C. J., *et al.* (2016). Atlantic salmon populations reveal adaptive divergence of immune related genes - a duplicated genome under selection. *BMC Genomics*, 17 (1): 610.
- Klykken, C., Boissonnot, L., Reed, A. K., Whatmore, P., Attramadal, K. & Olsen, R. E. (2022). Gene expression patterns in Atlantic salmon (*Salmo salar*) with severe nephrocalcinosis. *Journal of Fish Diseases*, 45 (11): 1645-1658.
- Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, 37 (5): 540-546.
- Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M. & Kamatani, Y. (2019). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biology*, 20 (1): 117.
- Lam, M. E. & Borch, T. (2011). Cultural valuing of fishery resources by the Norwegian Saami. In vol. 361 *Globalisation and ecological integrity in science and international law*, pp. 361-376: Cambridge Scholars Publishing in association with GSE Research.
- Lappalainen, T., Scott, A. J., Brandt, M. & Hall, I. M. (2019). Genomic Analysis in the Age of Human Genome Sequencing. *Cell*, 177 (1): 70-84.
- Lehner, B. & Grill, G. (2013). Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. *Hydrological Processes*, 27 (15): 2171-2186.

- Lehnert, S. J., Kess, T., Bentzen, P., Kent, M. P., Lien, S., Gilbey, J., Clément, M., Jeffery, N. W., Waples, R. S. & Bradbury, I. R. (2019). Genomic signatures and correlates of widespread population declines in salmon. *Nature Communications*, 10 (1): 2996.
- Lennox, R. J., Eliason, E. J., Havn, T. B., Johansen, M. R., Thorstad, E. B., Cooke, S. J., Diserud, O. H., Whoriskey, F. G., Farrell, A. P. & Uglem, I. (2018). Bioenergetic consequences of warming rivers to adult Atlantic salmon *Salmo salar* during their spawning migration. *Freshwater Biology*, 63 (11): 1381-1393.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25 (16): 2078-2079.
- Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34 (18): 3094-3100.
- Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragozcy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., *et al.* (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, 326 (5950): 289-293.
- Lien, S., Gidskehaug, L., Moen, T., Hayes, B. J., Berg, P. R., Davidson, W. S., Omholt, S. W. & Kent, M. P. (2011). A dense SNP-based linkage map for Atlantic salmon (*Salmo salar*) reveals extended chromosome homeologies and striking differences in sex-specific recombination patterns. *BMC Genomics*, 12 (1): 615.
- Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Kent, M. P., Nome, T., Hvidsten, T. R., Leong, J. S., Minkley, D. R., Zimin, A., *et al.* (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature*, 533 (7602): 200-205.
- Lisch, D. (2013). How important are transposons for plant evolution? *Nature Reviews Genetics*, 14 (1): 49-61.
- Liu, S., Gao, G., Layer, R. M., Thorgaard, G. H., Wiens, G. D., Leeds, T. D., Martin, K. E. & Palti, Y. (2021). Identification of High-Confidence Structural Variants in Domesticated Rainbow Trout Using Whole-Genome Sequencing. *Frontiers in Genetics*, 12.
- Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G.-A., Zhang, H., Liu, Z., Shi, M., *et al.* (2020). Pan-Genome of Wild and Cultivated Soybeans. *Cell*, 182 (1): 162-176.e13.
- Lu, T.-Y., Munson, K. M., Lewis, A. P., Zhu, Q., Tallon, L. J., Devine, S. E., Lee, C., Eichler, E. E., Chaisson, M. J. P. & The Human Genome Structural Variation, C. (2021). Profiling variable-number tandem repeat variation across populations using repeat-pangenome graphs. *Nature Communications*, 12 (1): 4250.
- Magadan, S., Sunyer, O. J. & Boudinot, P. (2015). Unique Features of Fish Immune Repertoires: Particularities of Adaptive Immunity Within the Largest Group of Vertebrates. In Hsu, E. & Du Pasquier, L. (eds) *Pathogen-Host Interactions: Antigenic Variation v. Somatic Adaptations*, pp. 235-264. Cham: Springer International Publishing.
- Manni, M., Berkeley, M. R., Seppely, M., Simão, F. A. & Zdobnov, E. M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*, 38 (10): 4647-4654.

- McCormick, S. D., Hansen, L. P., Quinn, T. P. & Saunders, R. L. (1998). Movement, migration, and smolting of Atlantic salmon (*Salmo salar*). *Canadian Journal of Fisheries and Aquatic Sciences*, 55 (S1): 77-92.
- McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P. & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, 17 (1): 122.
- Melo, M. C., Andersson, E., Fjellidal, P. G., Bogerd, J., França, L. R., Taranger, G. L. & Schulz, R. W. (2014). Salinity and photoperiod modulate pubertal development in Atlantic. *J Endocrinol*, 220: 1-15.
- Mérot, C., Oomen, R. A., Tigano, A. & Wellenreuther, M. (2020). A roadmap for understanding the evolutionary significance of structural genomic variation. *Trends in Ecology & Evolution*, 35 (7): 561-572.
- Mérot, C., Stenløkk, K. S. R., Venney, C., Laporte, M., Moser, M., Normandeau, E., Árnýasi, M., Kent, M., Rougeux, C., Flynn, J. M., *et al.* (2022). Genome assembly, structural variants, and genetic differentiation between lake whitefish young species pairs (*Coregonus* sp.) with long and short reads. *Molecular Ecology*.
- Mikocziova, I., Greiff, V. & Sollid, L. M. (2021). Immunoglobulin germline gene variation and its impact on human disease. *Genes & Immunity*, 22 (4): 205-217.
- Mobley, K. B., Aykanat, T., Czorlich, Y., House, A., Kurko, J., Miettinen, A., Moustakas-Verho, J., Salgado, A., Sinclair-Waters, M., Verta, J.-P., *et al.* (2021). Maturation in Atlantic salmon (*Salmo salar*, Salmonidae): a synthesis of ecological, genetic, and molecular processes. *Reviews in Fish Biology and Fisheries*, 31 (3): 523-571.
- Mun, S., Kim, S., Lee, W., Kang, K., Meyer, T. J., Han, B.-G., Han, K. & Kim, H.-S. (2021). A study of transposable element-associated structural variations (TASVs) using a de novo-assembled Korean genome. *Experimental & Molecular Medicine*, 53 (4): 615-630.
- Novák, P., Neumann, P., Pech, J., Steinhaisl, J. & Macas, J. (2013). RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, 29 (6): 792-793.
- Novák, P., Ávila Robledillo, L., Koblížková, A., Vrbová, I., Neumann, P. & Macas, J. (2017). TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Research*, 45 (12): e111-e111.
- OECD. (2017). *OECD Review of Fisheries*: OECD publishing.
- Otero, J., L'Abée-Lund, J. H., Castro-Santos, T., Leonardsson, K., Storvik, G. O., Jonsson, B., Dempson, B., Russell, I. C., Jensen, A. J., Baglinière, J.-L., *et al.* (2014). Basin-scale phenology and effects of climate variability on global timing of initial seaward migration of Atlantic salmon (*Salmo salar*). *Global Change Biology*, 20 (1): 61-75.
- Pâques, F., Leung, W.-Y. & Haber, J. E. (1998). Expansions and contractions in a tandem repeat induced by double-strand break repair. *Molecular and cellular biology*, 18 (4): 2045-2054.
- Pearse, D. E., Barson, N. J., Nome, T., Gao, G., Campbell, M. A., Abadía-Cardoso, A., Anderson, E. C., Rundio, D. E., Williams, T. H., Naish, K. A., *et al.* (2019). Sex-dependent dominance maintains migration supergene in rainbow trout. *Nature Ecology & Evolution*, 3 (12): 1731-1742.
- Pedersen, B. S. & Quinlan, A. R. (2018). Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics (Oxford, England)*, 34 (5): 867-868.

- Perrier, C., Ferchaud, A.-L., Sirois, P., Thibault, I. & Bernatchez, L. (2017). Do genetic drift and accumulation of deleterious mutations preclude adaptation? Empirical investigation using RADseq in a northern lacustrine fish. *Molecular Ecology*, 26 (22): 6317-6335.
- Phillips, R. & Rab, P. (2001). Chromosome evolution in the Salmonidae (Pisces): an update. *Biological Reviews*, 76 (1): 1-25.
- Quan, C., Li, Y., Liu, X., Wang, Y., Ping, J., Lu, Y. & Zhou, G. (2021). Characterization of structural variation in Tibetans reveals new evidence of high-altitude adaptation and introgression. *Genome Biology*, 22 (1): 159.
- Robertson, F. M., Gundappa, M. K., Grammes, F., Hvidsten, T. R., Redmond, A. K., Lien, S., Martin, S. A. M., Holland, P. W. H., Sandve, S. R. & Macqueen, D. J. (2017). Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification. *Genome Biology*, 18 (1): 111.
- Rougemont, Q. & Bernatchez, L. (2018). The demographic history of Atlantic salmon (*Salmo salar*) across its distribution range reconstructed from approximate Bayesian computations\*. *Evolution*, 72 (6): 1261-1277.
- Schneider, K., Adams, C. E. & Elmer, K. R. (2019). Parallel selection on ecologically relevant gene functions in the transcriptomes of highly diversifying salmonids. *BMC genomics*, 20 (1): 1-23.
- Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A. & Kendall, J. (2007). Strong association of de novo copy number mutations with autism. *Science*, 316 (5823): 445-449.
- Sedlazeck, F. J., Lee, H., Darby, C. A. & Schatz, M. C. (2018a). Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics*, 19 (6): 329-346.
- Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A. & Schatz, M. C. (2018b). Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, 15 (6): 461-468.
- Shafin, K., Pesout, T., Chang, P.-C., Nattestad, M., Kolesnikov, A., Goel, S., Baid, G., Kolmogorov, M., Eizenga, J. M., Miga, K. H., *et al.* (2021). Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nature Methods*, 18 (11): 1322-1332.
- Sherman, R. M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., Boorgula, M. P., Chavan, S., Vergara, C., Ortega, V. E., *et al.* (2019). Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nature Genetics*, 51 (1): 30-35.
- Shi, J., Tian, Z., Lai, J. & Huang, X. (2022). Plant pan-genomics and its applications. *Molecular Plant*.
- Sinclair-Waters, M., Nome, T., Wang, J., Lien, S., Kent, M. P., Sægvog, H., Florø-Larsen, B., Bolstad, G. H., Primmer, C. R. & Barson, N. J. (2022). Dissecting the loci underlying maturation timing in Atlantic salmon using haplotype and multi-SNP based association methods. *Heredity*, 129 (6): 356-365.
- Skotte, L., Korneliussen, T. S. & Albrechtsen, A. (2013). Estimating Individual Admixture Proportions from Next Generation Sequencing Data. *Genetics*, 195 (3): 693-702.
- Skugor, S., Glover, K. A., Nilsen, F. & Krasnov, A. (2008). Local and systemic gene expression responses of Atlantic salmon (*Salmo salar* L.) to infection with the salmon louse (*Lepeophtheirus salmonis*). *BMC genomics*, 9 (1): 1-18.
- Smit, A., Hubley, R. & Green, P. (2015). *RepeatMasker Open-4.0. 2013–2015*.

- Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., *et al.* (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, 526 (7571): 75-81.
- Sulovari, A., Li, R., Audano, P. A., Porubsky, D., Vollger, M. R., Logsdon, G. A., Consortium, H. G. S. V., Warren, W. C., Pollen, A. A., Chaisson, M. J. P., *et al.* (2019). Human-specific tandem repeat expansion and differential gene expression during primate evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 116 (46): 23243-23253.
- Swanson, P., Dickey, J. T. & Campbell, B. (2003). Biochemistry and physiology of fish gonadotropins. *Fish Physiology and biochemistry*, 28 (1): 53-59.
- Talkowski, M. E., Rosenfeld, J. A., Blumenthal, I., Pillalamarri, V., Chiang, C., Heilbut, A., Ernst, C., Hanscom, C., Rossin, E. & Lindgren, A. M. (2012). Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. *Cell*, 149 (3): 525-537.
- Taylor, J. S. & Breden, F. (2000). Slipped-strand mispairing at noncontiguous repeats in *Poecilia reticulata*: a model for minisatellite birth. *Genetics*, 155 (3): 1313-1320.
- Team, R. C. (2013). R: A language and environment for statistical computing.
- Tettelin, H., Maignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L. & Durkin, A. S. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences*, 102 (39): 13950-13955.
- Tham, C. Y., Tirado-Magallanes, R., Goh, Y., Fullwood, M. J., Koh, B. T. H., Wang, W., Ng, C. H., Chng, W. J., Thiery, A., Tenen, D. G., *et al.* (2020). NanoVar: accurate characterization of patients' genomic structural variants using low-depth nanopore sequencing. *Genome Biology*, 21 (1): 56.
- Thorstad, E. B., Bliss, D., Breau, C., Damon-Randall, K., Sundt-Hansen, L. E., Hatfield, E. M. C., Horsburgh, G., Hansen, H., Maoiléidigh, N. Ó., Sheehan, T., *et al.* (2021). Atlantic salmon in a rapidly changing environment—Facing the challenges of reduced marine survival and climate change. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 31 (9): 2654-2665.
- Timón-Gómez, A., Nývltová, E., Abriata, L. A., Vila, A. J., Hosler, J. & Barrientos, A. (2018). *Mitochondrial cytochrome c oxidase biogenesis: Recent developments*. Seminars in cell & developmental biology: Elsevier. 163-178 pp.
- Vaishnav, J., Khan, F., Yadav, M., Parmar, N., Buch, H., Jadeja, S. D., Dwivedi, M. & Begum, R. (2022). V-set domain containing T-cell activation inhibitor-1 (VTCN1): A potential target for the treatment of autoimmune diseases. *Immunobiology*, 227 (6): 152274.
- Vasimuddin, M., Misra, S., Li, H. & Aluru, S. (2019). *Efficient architecture-aware acceleration of BWA-MEM for multicore systems*. 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS): IEEE. 314-324 pp.
- Villoutreix, R., de Carvalho, C. F., Soria-Carrasco, V., Lindtke, D., De-la-Mora, M., Muschick, M., Feder, J. L., Parchman, T. L., Gompert, Z. & Nosil, P. (2020). Large-scale mutation in the evolution of a gene complex for cryptic coloration. *Science*, 369 (6502): 460-466.
- Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., *et al.* (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLOS ONE*, 9 (11): e112963.

- Wang, J., Belosevic, M. & Stafford, J. L. (2021). Identification of distinct LRC-and Fc receptor complex-like chromosomal regions in fish supports that teleost leukocyte immune-type receptors are distant relatives of mammalian Fc receptor-like molecules. *Immunogenetics*, 73 (1): 93-109.
- Watson, Corey T., Steinberg, Karyn M., Huddleston, J., Warren, Rene L., Malig, M., Schein, J., Willsey, A. J., Joy, Jeffrey B., Scott, Jamie K., Graves, T. A., *et al.* (2013). Complete Haplotype Sequence of the Human Immunoglobulin Heavy-Chain Variable, Diversity, and Joining Genes and Characterization of Allelic and Copy-Number Variation. *The American Journal of Human Genetics*, 92 (4): 530-546.
- Watson, K. B., Lehnert, S. J., Bentzen, P., Kess, T., Einfeldt, A., Duffy, S., Perriman, B., Lien, S., Kent, M. & Bradbury, I. R. (2022). Environmentally associated chromosomal structural variation influences fine-scale population structure of Atlantic Salmon (*Salmo salar*). *Molecular Ecology*, 31 (4): 1057-1075.
- Weissensteiner, M. H., Bunikis, I., Catalán, A., Francoijs, K.-J., Knief, U., Heim, W., Peona, V., Pophaly, S. D., Sedlazeck, F. J., Suh, A., *et al.* (2020). Discovery and population genomics of structural variation in a songbird genus. *Nature Communications*, 11 (1): 3403.
- Wellenreuther, M., Mérot, C., Berdan, E. & Bernatchez, L. (2019). Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification. *Molecular Ecology*, 28 (6): 1203-1209.
- Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., *et al.* (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, 8 (12): 973-982.
- Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., *et al.* (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*, 2 (3).
- Yan, S. M., Sherman, R. M., Taylor, D. J., Nair, D. R., Bortvin, A. N., Schatz, M. C. & McCoy, R. C. (2021). Local adaptation and archaic introgression shape global diversity at human structural variant loci. *eLife*, 10: e67615.
- Yasuike, M., De Boer, J., von Schalburg, K. R., Cooper, G. A., McKinnel, L., Messmer, A., So, S., Davidson, W. S. & Koop, B. F. (2010). Evolution of duplicated IgH loci in Atlantic salmon, *Salmo salar*. *BMC genomics*, 11 (1): 1-16.
- Yoshida, G. M. & Yáñez, J. M. (2022). Increased accuracy of genomic predictions for growth under chronic thermal stress in rainbow trout by prioritizing variants from GWAS using imputed sequence data. *Evolutionary applications*, 15 (4): 537-552.



## Supplementary tables and figures

**Table S1:** Metadata for 11 Atlantic salmon samples used for assemblies and long-read based SV-detection. AQGE = aquaculture, ATL = Atlantic, BWS = Barents/White Sea, BAL = Baltic, NAM = North America.

Name	River name	Phylo-geographic group	Country	Gender	Population type	Lat, long
AQGE	-	AQU	Norway	Male	Aquaculture	-
GLOP	Gloppenelva	ATL	Norway	Male	Anadromous	61.46N, 6.12E
ARUN	Årungselta	ATL	Norway	Male	Anadromous	59.43N, 10.43E
ALTA	Altaelva	BWS	Norway	Male	Anadromous	69.58N, 23.22E
TANA	Tanaelva	BWS	Norway	Male	Anadromous	70.29N, 28.23E
FROM	River Frome	ATL	UK	Male	Anadromous	50.41N, 2.05W
OULO	Oulujoki	BAL	Finland	Male	Anadromous	65.49N, 24.09E
PERU	Lac Perugia	NAM	Canada	Male	Landlocked	47.43N, 76.30W
SEBA	Sebago Lake	NAM	USA	Female	Landlocked	43.52N, 70.34W
GARN-1	Garnish River	NAM	Canada	Male	Anadromous	47.23N, 55.35W
GARN-2	Garnish River	NAM	Canada	Male	Anadromous	47.23N, 55.35W

**Table S2:** Data amount and read N50 for long-reads used to call SVs. Median read depth is calculated after mapping per 100 bp.

Name	Amount (Gbp)	Read N50 (kbp)	Median depth
AQGE	191.6	30.7	72
GLOP	73.9	35.1	26
ARUN	63.0	49.1	22
ALTA	75.8	47.0	26
TANA	63.4	22.0	22.38
FROM	123.1	56.6	42
TORN	48.8	35.6	16
PERU	54.2	32.6	18
SEBA	63.5	34.3	21
GARN-1	62.7	50.4	15.07
GARN-2	68.3	51.4	18

**Table S3:** Statistics for Atlantic salmon assemblies.

Name	Total sequence length	Anchored sequence	Number of contigs	Contig N50	BUSCO (%)	Overlap used for Flye (kbp)
ICSASG_v2	2,966,890,203	2,240,204,991	368,060	57,618	94.5	
AQGE	2,756,584,103	2,499,322,922	4,222	28,058,890	98.1	5, 10, 15, 20, 30
GLOP	2,638,948,061	-	5,962	10,290,042	97.9	10
ARUN	2,679,716,821	-	5,496	8,712,972	97.9	10
ALTA	2,623,493,511	2,459,683,978	4,068	19,218,354	98.4	7, 10, 15
TANA	2,530,437,290	-	8,427	3,255,277	97.6	8
FROM	2,693,730,146	-	3,156	28,320,961	96.4	5, 10, 15, 20
OULO	2,638,488,729	-	4,417	9,391,664	98.0	10
PERU	2,543,457,543	-	4,091	8,181,635	98.1	7
SEBA	2,691,559,704	-	6,090	3,912,010	97.8	14
GARN-1	2,624,413,291	2,489,676,464	2,983	27,893,759	98.0	7, 10, 15, 30
GARN-2	2,724,867,716	-	6,982	5,216,392	97.7	10

**Table S4:** SV statistics for the full list of SVs and per sample. "All" refers to the dataset merged across samples.

Name	Total number of SVs	Number of deletions	Number of insertions	Number of duplications	Number of inversions
ALL	1,061,452	781,244	275,462	3,340	1,407
ALTA	199,993	135,753	63,574	349	317
ARUN	192,163	139,067	52,397	397	302
FROM	248,043	175,827	71,533	363	320
GARN-1	226,146	151,114	73,748	794	491
GARN-2	240,081	167,973	70,935	739	434
GLOP	231,463	179,187	51,709	265	302
OULO	169,781	118,888	50,164	396	333
PERU	199,447	158,002	40,671	679	95
SEBA	245,250	195,460	48,937	744	109
TANA	207,239	152,932	53,609	362	336

**Table S5:** Number of indels shared between samples.

Shared between n samples	n indels
1	632,193
2	165,474
3	93,093
4	65,333
5	36,218
6	24,341
7	16,347
8	11,753
9	7,401
10	4,553

**Table S6:** Total number of base pairs covered by insertions and deletions. All refers to the dataset merged across samples.

Name	Total base pairs in indels	Base pairs in deletions	Base pairs in insertion
ALL	366,610,121	259,412,339	107,197,782
ALTA	71,485,934	49,926,307	21,559,627
ARUN	64,958,459	45,478,512	19,479,947
FROM	81,105,325	58,944,012	22,161,313
GARN-1	96,136,834	61,667,181	34,469,653
GARN-2	94,310,010	61,807,572	32,502,438
GLOP	74,116,108	56,837,064	17,279,044
OULO	60,612,730	42,582,888	18,029,842
PERU	76,764,373	60,320,104	16,444,269
SEBA	88,626,268	69,236,518	19,389,750
TANA	66,231,274	48,198,099	18,033,175

**Table S7:** Overlap between SVs and CDS and genes in number of occurrences and by base pairs.

	n SVs overlapping	Proportion of n SVs (%)	Overlap in bp	Proportion of SV sequence (%)
CDS	13,038	1.23	2,519,944	0.69
Gene	627,588	59.39	140,911,410	38.44

**Table S8:** Number of functional consequences estimated by Variant Effect Predictor.

Consequence	n overlapping SVs
Feature truncation	4,462
Frameshift variant	2,807
Stop lost	1,769
3' UTR variant	1,749
Coding sequence variant	1,652
Intron variant	1,099
Transcript ablation	435
5' UTR variant	219
Start lost	94
Start retained variant	94
Inframe deletion	3
Total	14,383

**Table S9:** Duplicate and singleton genes in the Atlantic salmon overlapping SVs used for Fisher's exact test ( $P < 2.2 \cdot 10^{-16}$ ).

	Duplicate genes	Singleton genes	Total
<b>SV overlap</b>	17,640	6,853	24,493
<b>No SV overlap</b>	4,826	2,917	7,743
<b>Total</b>	22,466	9,770	32,236

**Table S10:** Duplicate and singleton genes in the Atlantic salmon with coding sequence (CDS) overlapping SVs used for Fisher's exact test ( $P < 0.0116$ ).

	Duplicate genes	Singleton genes	Total
<b>SV overlap</b>	1,400	538	1,938
<b>No SV overlap</b>	21066	9,232	30,298
<b>Total</b>	22,466	9,770	32,236

**Table S11: Repeat annotations.** Number of repeat entries and proportion of genome of transposable elements and tandem repeats.

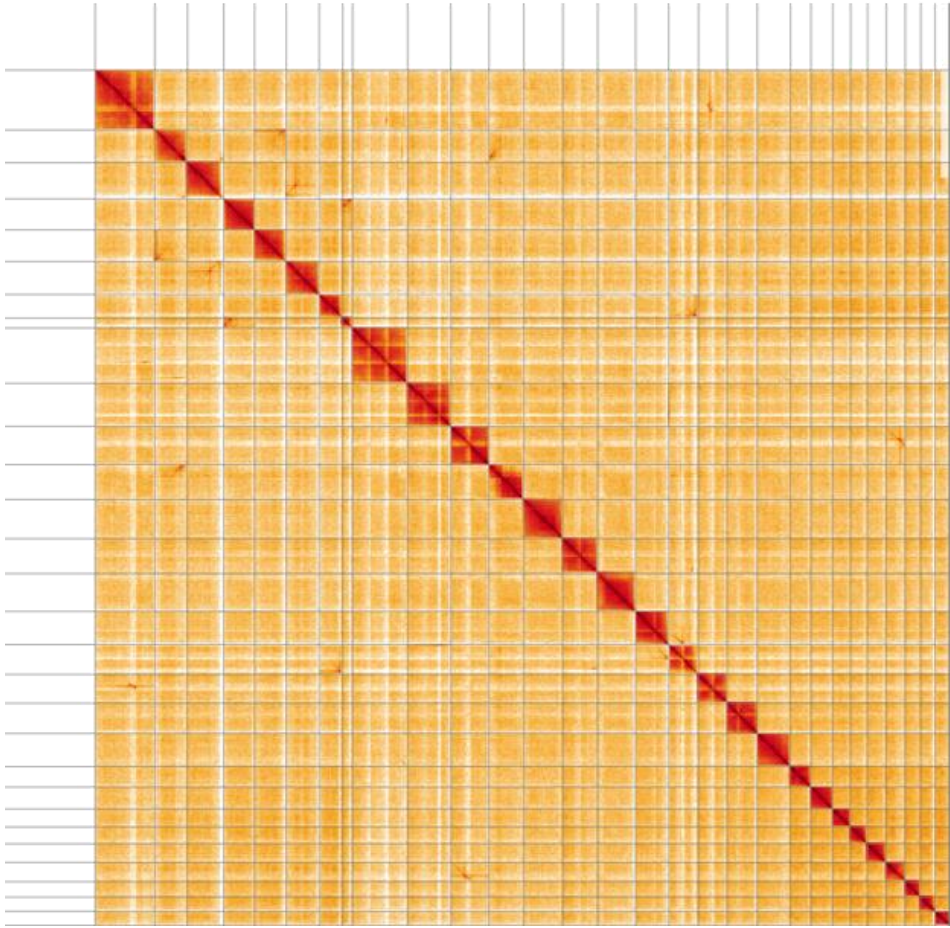
	Number of elements	Proportion of genome (%)
Transposable elements	1,832,285	40.61
Tandem repeats	2,334,554	20.17

**Table S12:** Distribution of TEs by family, incl. summaries. E.g. “DNA transposons” is the sum of all categorised and uncategorised DNA transposons in the genome.

Type	n	Total bp	Proportion of genome (%)
Retroelements	989,281	385,493,162	15.42
SINEs	184,238	26,950,020	1.08
LINEs	654,096	287,294,018	11.49
R1/LOA/Jockey	486,952	207,565,864	8.30
RTE/Bov-B	26,139	17,346,465	0.69
L1/CIN4	24,902	10,906,299	0.44
LTR elements	150,947	7,1249,124	2.85
BEL/Pao	758	480,521	0.02
Gypsy/DIRS1	63,674	46,459,553	1.86
Retroviral	6,315	1,712,719	0.07
DNA transposons	1,526,496	479,546,375	19.19
hobo-Activator	135,832	44,458,628	1.78
Tc1-IS630-Pogo	805,085	289,286,954	11.57
PiggyBac	26,804	7,201,568	0.29
Tourist/Harbinger	1,652	854,835	0.03
Unclassified	570,559	133,701,458	5.35

**Table S13:** Overlap between SV and repeat DNA annotations. TE and TR overlap with deletions and insertions measured by number of overlaps, overlap in base pairs and proportion of SVs overlapping repeats. \*Overlapping base pairs of insertions equals the number of entries as the insertion coordinates are recorded as a single base pair relative to the reference.

	Number of overlapping variants		Total overlap in base pairs		Proportion of SV sequence overlapping repeat elements (%)	
	Deletions	Insertions	Deletions	Insertions	Deletions	Insertions
Transposable elements	189,274	58,262	62,326,703	58,262*	24.02	21.15*
Tandem repeats	442,147	96,598	72,738,822	96,598*	28.04	35.07*



**Figure S1:** Pore-C contact map for GARN-1.

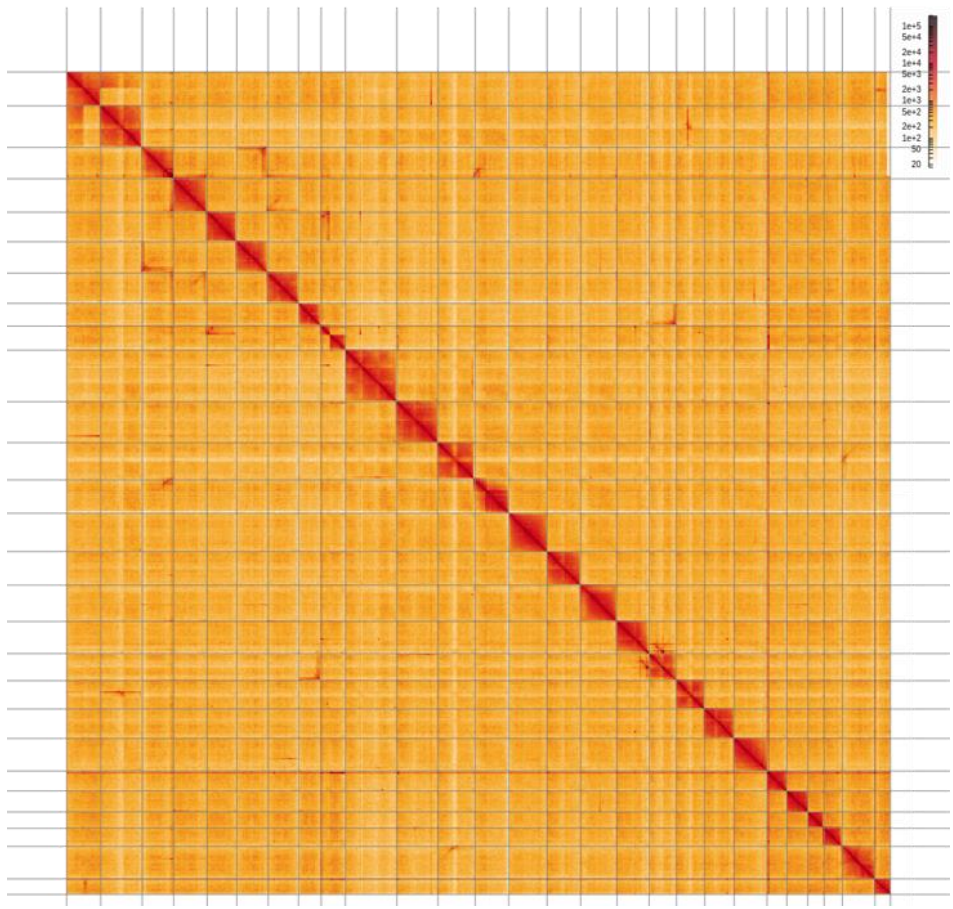
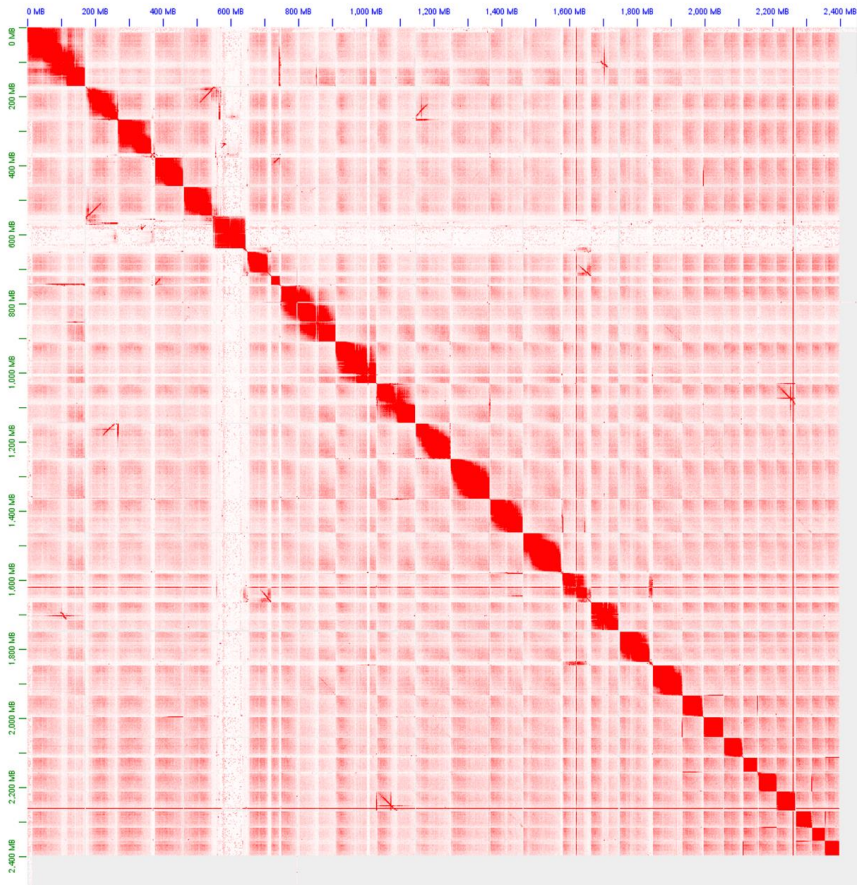
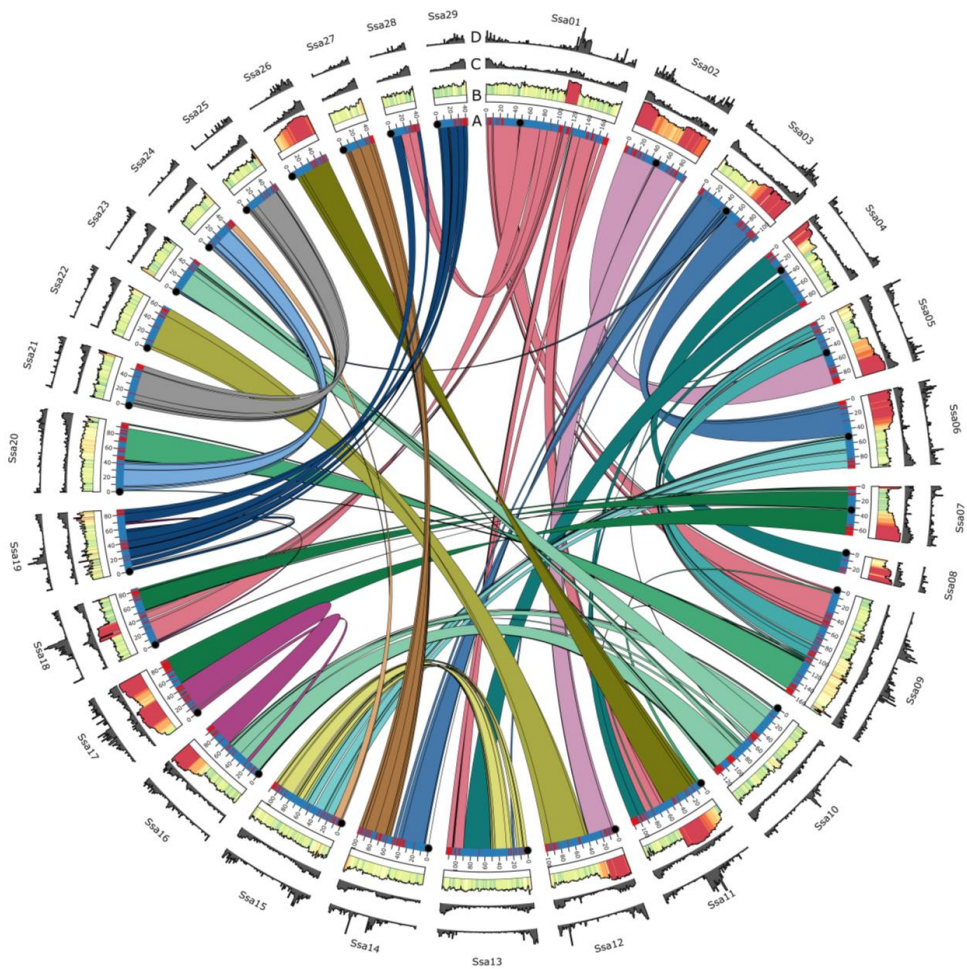


Figure S2: Pore-C contact map for ALTA.



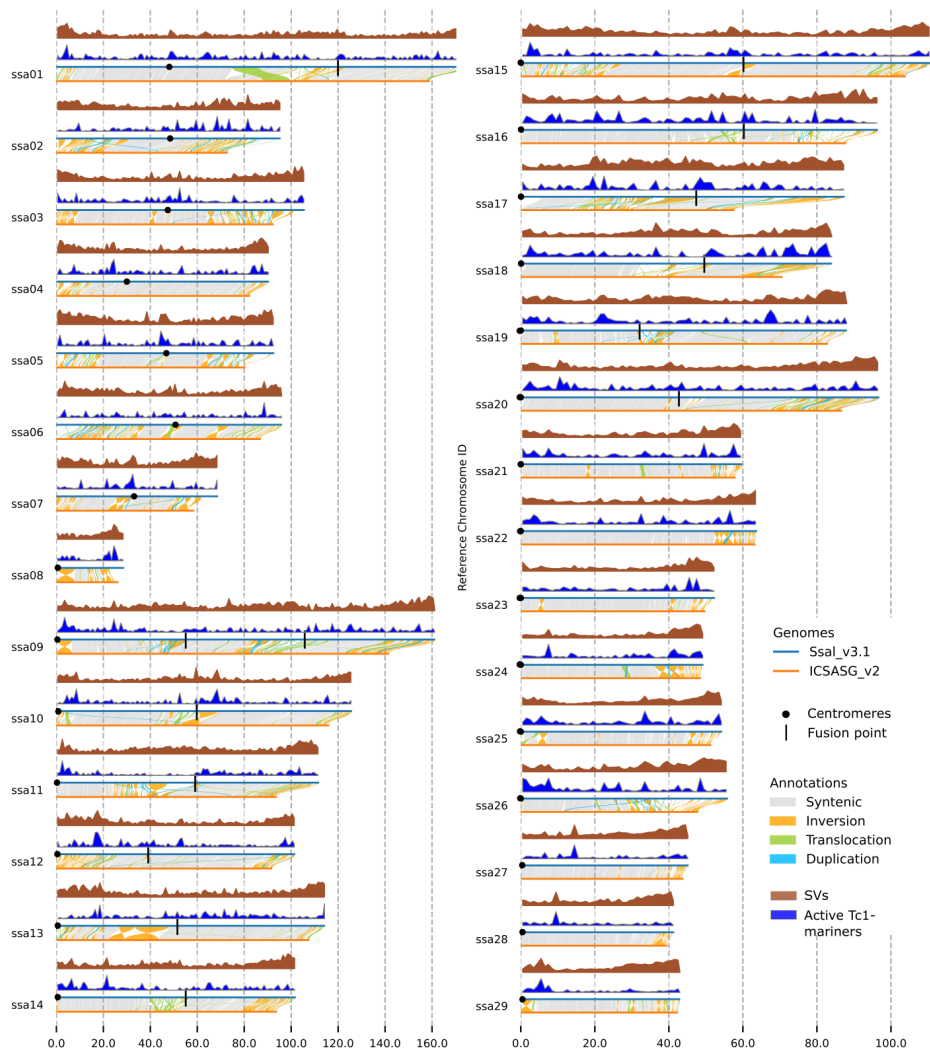
**Figure S3:** Hi-C contact map for AQGE.



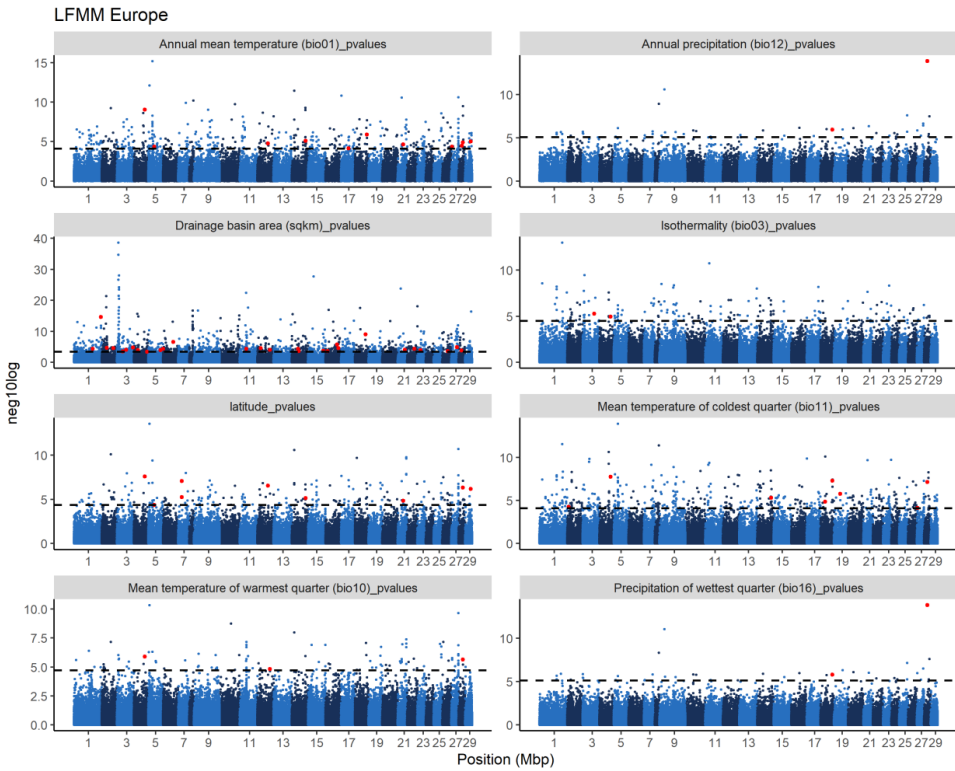


**Figure S4. Circos plot links showing homeologous regions in the Atlantic salmon genome. (A)** regions with average mapping depth of less than 1 (red) per Mbp between ICSASG\_v2 and Ssa\_v3.1. Black circles represent the centromere position. **(B)** Sequence similarity between homologous blocks in the genome ranging from 80% (green) to 100% (red) **(C)** SVs per Mbp. **(D)** TRs per Mbp.

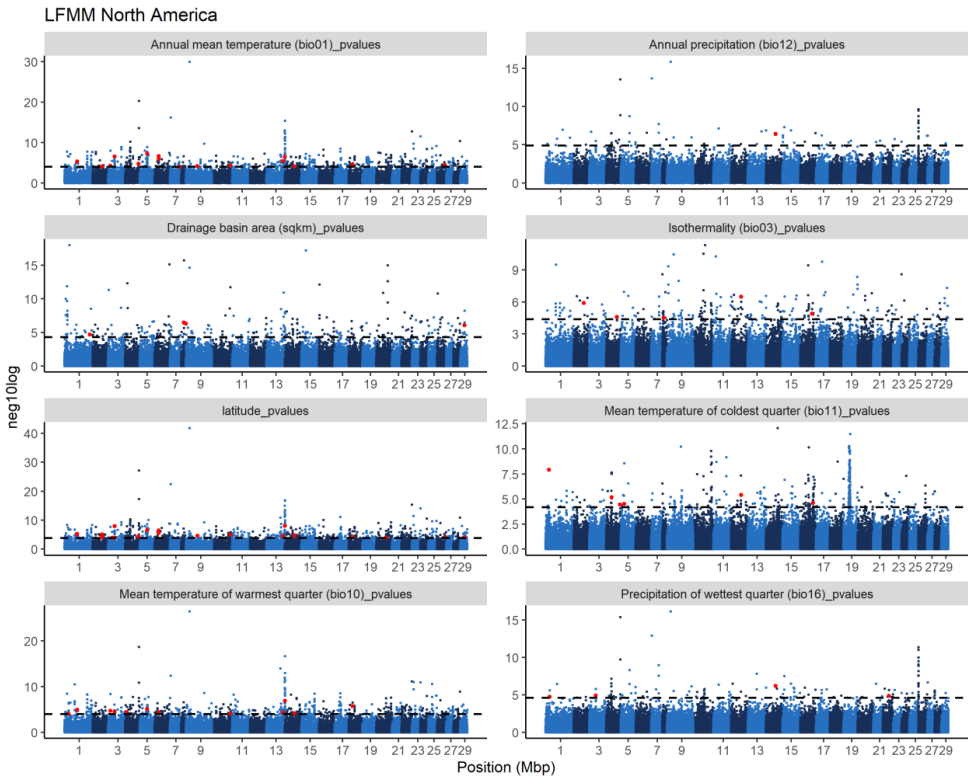




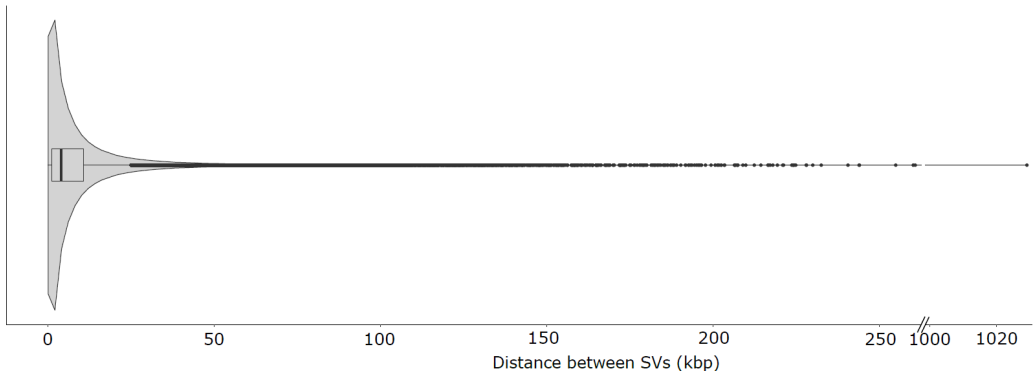
**Table S6:** Syntenic and rearranged regions between Ssal\_v3.1 and ICSASG\_v2 with SV (brown) and active or recent Tc-1 mariners (TE) (blue) density for the 29 chromosomes representing the European karyotype of Atlantic salmon.



**Figure S7:** Genotype-environment associations for European samples (n=286). Blue points represent SVs and red points represent SVs overlapping coding sequence of a gene. The striped line indicated the significance threshold ( $-\log_{10}(P) > 0.05$ ).



**Figure S8:** Genotype-environment associations for North American samples (n=80). Blue points represent SVs and points dots represent SVs overlapping coding sequence of a gene. The striped line indicated the significance threshold ( $-\log_{10}(P) > 0.05$ ).



**Figure S9:** Violin plot showing distance between SVs (in kbp) used in genotype-environment association study. The range 250-1000 kbp is cut out.



**Figure S10:** Biological processes enriched in gene ontology analysis of genes overlapped by SVs significantly associated with environmental variables.

```

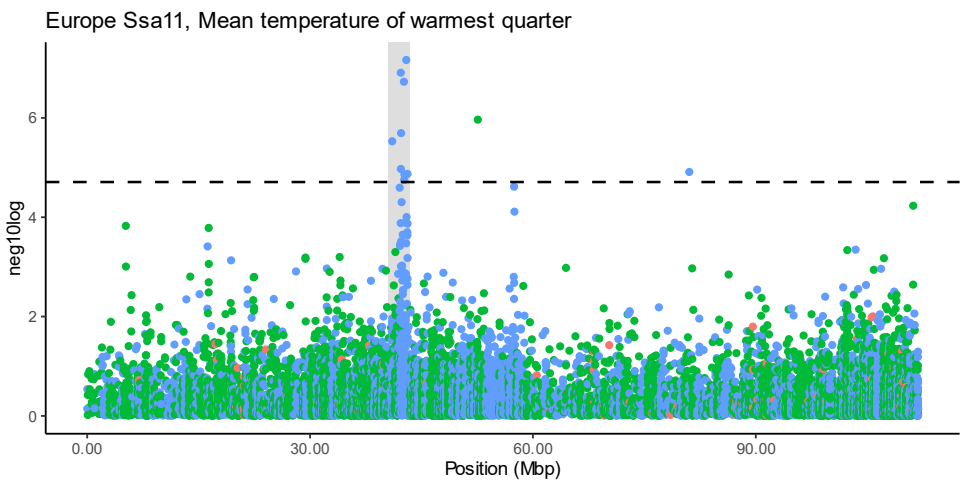
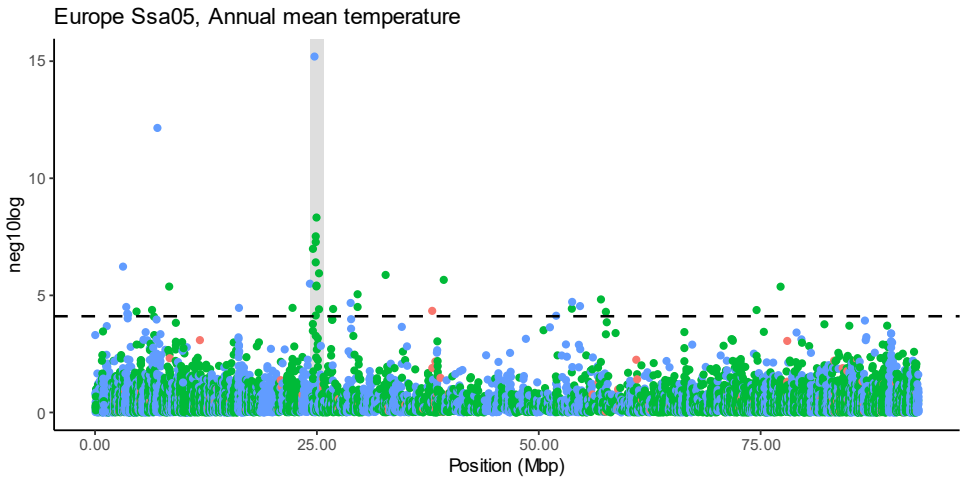
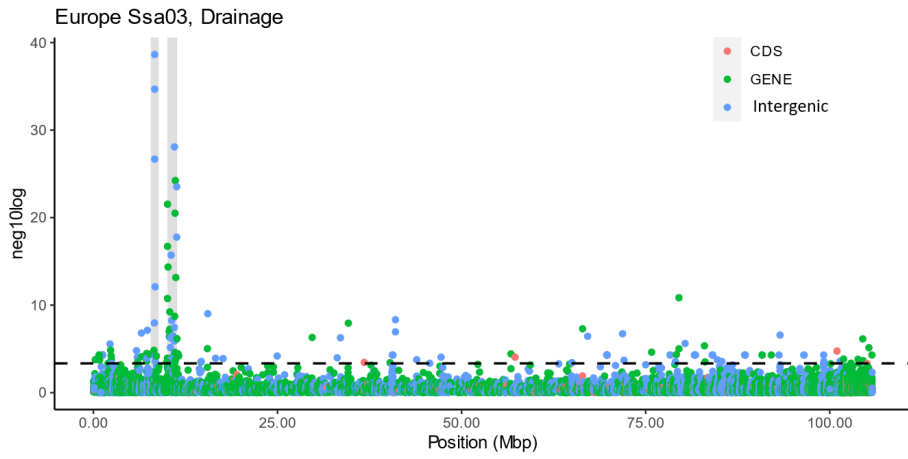
NP_001167142.1 MLGLHVGTLISLFLCILLEPVEGSLMQPCQPINQTVSLEKEGCPTCLVIQTPICSGHCVT
ref|XP_04556600 MLGLHVGTLISLFLCILLEPVEGSLMQPCQPINQTVSLEKEGCPTCLVIQTPICSGHCVT
*****

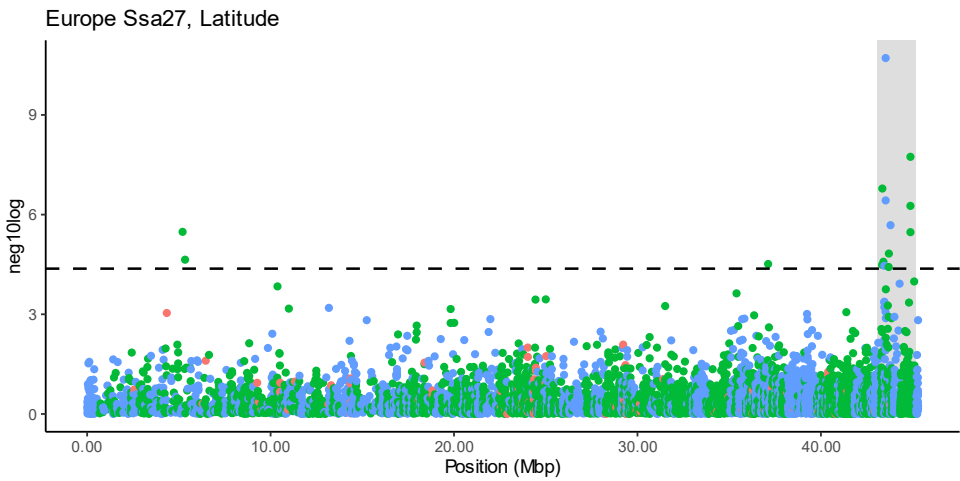
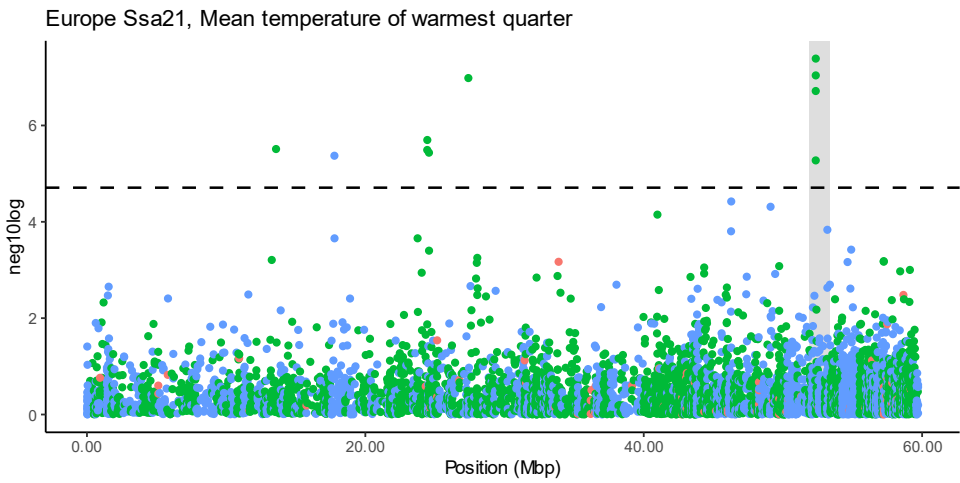
NP_001167142.1 KEPVFKSPFSTVYQHVCYTRDVRVYETIRLPDCPPWVDHVVITYPVALS CDCSLCNMDTSDC
ref|XP_04556600 KEPVFKSPFSTVYQHVCYTRDVRVYETIRLPDCPPWVDHVVITYPVALS CDCSLCNMDTSDC
*****

NP_001167142.1 TIESLQPDFCITRALMDGNMW
ref|XP_04556600 TIESLQPDFCITRALMDGNMW
*****

```

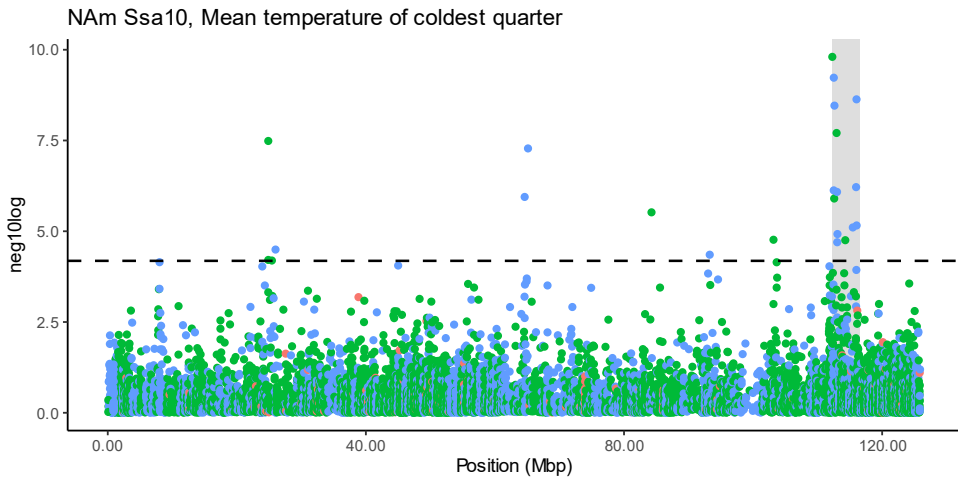
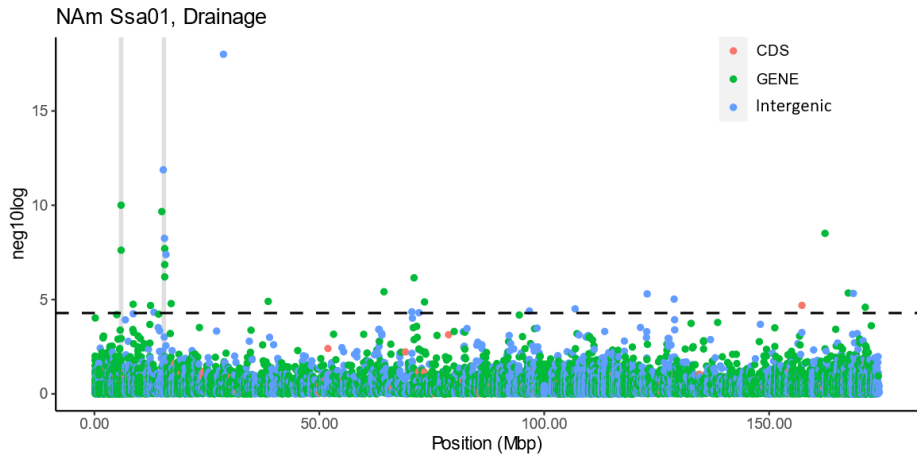
**Figure S11:** Alignment of protein sequences of the two *LHB* duplicates on chromosome 28. Important sites for interaction with receptor with amino acid shifts are marked with yellow. The first amino acid marked in yellow is at the end of loop 3 and the second marking is a amino acid shift just after the “seatbelt” (marked with an orange line).



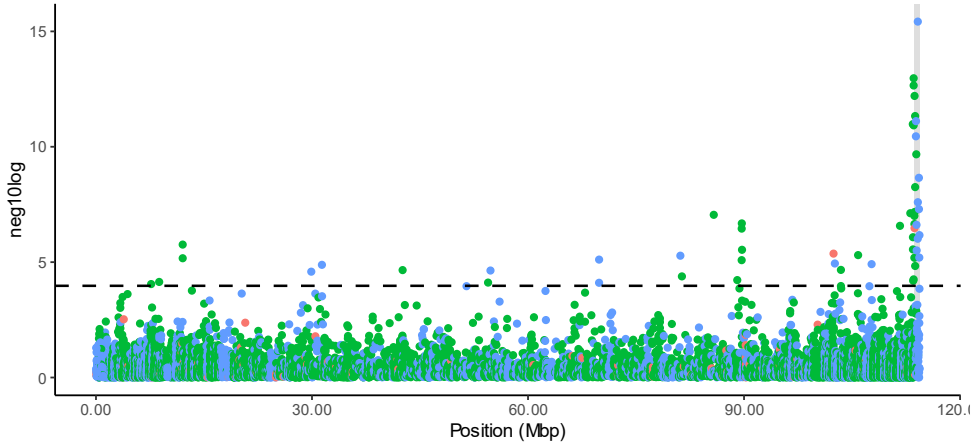


**Figure S12:** Genotype-environment association for chromosomes with distinct peaks (shaded in grey) for European samples. SVs overlapping genes are coloured green, coding sequence red and no functional overlap blue. Significance threshold is indicated with a dashed line.

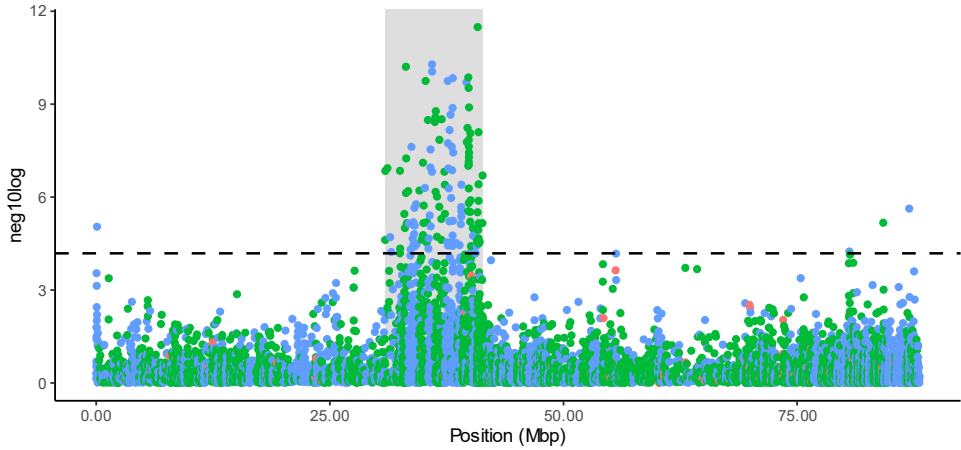




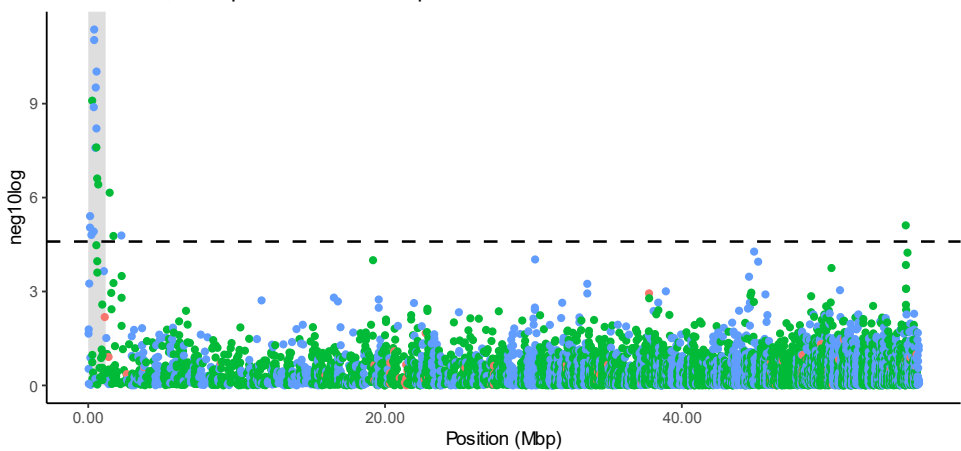
NAm Ssa13, Annual mean temperature

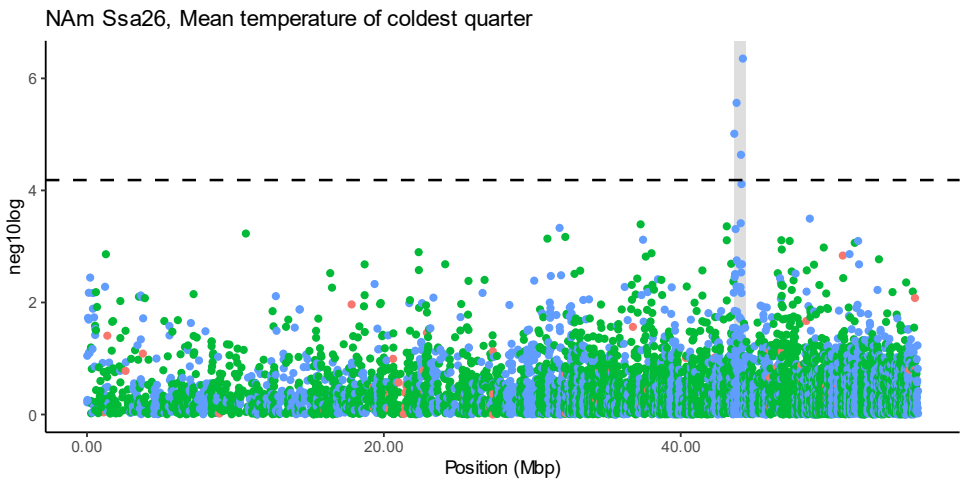


NAm Ssa19, Mean temperature of coldest quarter (bio11)\_pvalues



NAm Ssa26, Precipitation of wettest quarter





**Figure S13:** Genotype-environment association for chromosomes with distinct peaks (shaded in grey) for North American samples. SVs overlapping genes are coloured green, coding sequence red and no functional overlap blue. Significance threshold is indicated with a dashed line.



# PAPER II

# 1 The emergence of supergenes from inversions in Atlantic salmon.

2  
3 Kristina Stenløkk, Marie Saitou, Live Rud-Johansen, Torfinn Nome, Michel Moser, Mariann Árnýasi, Matthew  
4 Kent, Nicola Jane Barson\* and Sigbjørn Lien\*.

5  
6 \*contributed equally as senior and corresponding author

7  
8 Centre for Integrative Genetics (CIGENE) and Department of Animal and Aquacultural Sciences,  
9 Faculty of Biosciences,  
10 Norwegian University of Life Sciences

## 11 12 13 **Abstract**

14 Supergenes link allelic combinations into non-recombining units known to play an essential role in  
15 maintaining adaptive genetic variation. However, because supergenes can be maintained over millions of  
16 years by balancing selection and typically exhibit strong recombination suppression, both the underlying  
17 functional variants and how the supergenes are formed are largely unknown. Particularly, questions remain  
18 over the importance of inversion breakpoint sequences and whether supergenes capture preexisting  
19 adaptive variation or accumulate this following recombination suppression. To investigate the process of  
20 supergene formation, we identified inversion polymorphisms in Atlantic salmon by assembling eleven  
21 genomes with nanopore long-read sequencing technology. A genome assembly from the sister species,  
22 brown trout, was used to determine the standard state of the inversions. We found evidence for adaptive  
23 variation through genotype-environment associations, but not for the accumulation of deleterious  
24 mutations. One young 3Mb inversion segregating in North American populations, has captured adaptive  
25 variation that is still segregating within the standard arrangement of the inversion, while some adaptive  
26 variation has accumulated after the inversion. This inversion and two others had breakpoints disrupting  
27 genes. Three multigene inversions with matched repeat structures at the breakpoints did not show any  
28 supergene signatures, suggesting that shared breakpoint repeats may obstruct supergene formation.

## 29 30 **Keywords**

31 Inversion, supergene, Atlantic salmon, long-read sequencing, adaptive variation, population differentiation

## 32 33 **Introduction**

34 Supergenes are clusters of linked alleles which segregate as if they were a single locus and that determine  
35 alternate phenotypes in balanced polymorphisms [1]. They can evolve in regions of suppressed  
36 recombination caused by structural variation, including inversions, insertions and deletions where the  
37 linkage among favourable combinations of alleles is increased [2]. Chromosomal inversions are increasingly  
38 recognized as important for adaptation across taxa [1, 3-8] and often underlie supergenes. However,  
39 suppressed recombination makes them particularly vulnerable to the accumulation of recessive deleterious  
40 mutations and these may be key to supergene persistence through increased heterozygote fitness [2, 4].  
41 Furthermore, breakpoints mutations themselves can have direct phenotypic effects [5, 7, 8] making it unclear  
42 which processes lead to the development of supergenes. The suppressed recombination that is central to  
43 supergene development hinders the dissection of the events leading to their formation. Many iconic  
44 supergenes are relatively old, (e.g. >1MY [3, 5]) obscuring the sequence of events leading to their  
45 development. Despite increasing numbers of supergenes being detected, the early stages of supergene  
46 formation remain poorly understood.

47  
48 Inversion supergenes can be maintained by various forms of balancing selection. For example, sexual  
49 antagonism [3, 9], negative frequency dependent selection [5, 8], temporally and spatially varying  
50 selection [3, 4] and heterozygote advantage [4]. These different forms of balancing selection can combine to  
51 promote supergene persistence [4] and their importance can change over the lifespan of the inversion [10].  
52 Spatial variation in selection with migration can lead to a form of balancing selection, migration-selection  
53 balance, which can maintain inversion polymorphisms [11]. This is an attractive model as it explains both the

54 initial invasion, and subsequent maintenance as a polymorphism using a widespread process. When an  
55 inversion first occurs, and is rare, it is vulnerable to being lost by genetic drift or selection if it causes  
56 deleterious effects. The capture of locally adapted variation has been shown to increase the probability of  
57 an inversion increasing in frequency, however there is little empirical evidence to support its occurrence [12]  
58 (but see [6, 13]). Recombination suppression in inversions complicates the identification of the variants  
59 underlying their effects and the age of many inversions make it difficult to determine whether adaptive  
60 variation was present before the inversion occurred or accumulated afterwards. Likewise, the capture and  
61 accumulation of deleterious mutations can make heterozygotes more fit because they do not express this  
62 recessive genetic load leading to associative overdominance that prevents the inversion from becoming fixed  
63 [4, 11]. Capture of recessive deleterious variants within the inversion would not have an effect while the  
64 inversion is rare, because homozygotes would be very rare, but may influence its persistence as a  
65 polymorphism once it rises in frequency when the cost of reduced fitness of homozygotes is experienced [4].  
66

67 Although suppressed recombination causing tight linkage among adaptive variants located within inversions  
68 is thought to be central to their potential to develop into supergenes, inversions can also cause large  
69 effect mutations at their breakpoints. Breakpoint mutations can disrupt the coding sequence of genes (e.g.  
70 [5, 8]) or cause large deletions [2, 7]. These mutations can directly drive the phenotypic effects and be the  
71 target of selection themselves, or they can influence the evolutionary dynamics of the inversion, e.g. through  
72 recessive lethality of inversion homozygotes [5, 8]. Alternatively, selection may act on an adaptive breakpoint  
73 in combination with adaptive variants within the inversion, or in regions of reduced recombination extending  
74 beyond the breakpoints [7], to which they are linked. Unlike the variants contained within the inversion,  
75 breakpoint mutations occur concurrently with the inversion and so naturally segregate perfectly with it.  
76 However, because inversion breakpoints are often highly repetitive, they have been difficult to assemble and  
77 characterize using short-read sequencing and, consequently, our understanding of their contribution to  
78 supergene formation is incomplete [7].  
79

80 Atlantic salmon (*Salmo salar*) is an anadromous fish that spends its juvenile period in freshwater before  
81 undergoing a marine feeding migration and then returning to freshwater to spawn. Natal homing promotes  
82 local adaptation among heterogenous riverine environments [14]. Imperfect homing can lead to gene flow  
83 among these locally adapted populations that can promote the recruitment of large-effect loci [15]. These  
84 conditions are likely to favor inversion polymorphisms as shown in other salmonids (e.g. rainbow trout [3]).  
85 Salmonids experienced a whole-genome duplication (WGD) event 85-106 million years ago from which many  
86 duplicate genes are retained [16]. Large chromosomal rearrangements have been an important evolutionary  
87 mechanism during rediploidization following the WGD [16], but the present polymorphic inversion landscape  
88 in Atlantic salmon remains poorly characterized.  
89

90 Despite phenomenal advances in the detection and characterization of inversions, challenges remain  
91 regarding characterizing of inversion breakpoints, especially those containing inverted repeats or segmental  
92 duplications [7]. Here we identified polymorphic inversions using newly available long-read assemblies from  
93 11 Atlantic salmon sampled across the species range, representing all four phylogeographic groups; North  
94 American (NA), Baltic (BAL), Barents/White Sea (BWS) and Atlantic (ATL) [17]. We systematically searched  
95 the Atlantic salmon genome for inversions using assembly-based detection and investigated their potential  
96 for supergene formation. Long-read sequencing of its sister species, brown trout (*Salmo trutta*), was used to  
97 determine the standard arrangement of the inversions and characterize breakpoints. We use these  
98 inversions to investigate the importance of the capture of preexisting variation for supergene emergence.  
99

## 100 **Material and Methods**

### 101 *Nanopore long-read sequencing and building of genome assemblies*

102 The Atlantic salmon reference genome (GCA\_905237065.2) was built from 70x genome coverage with long-read  
103 Oxford Nanopore reads generated from a Norwegian aquaculture salmon (AQGE; Table S1). Long-read  
104 libraries were prepared using the SQK-LSK109 kit following the Genomic DNA by ligation protocol and  
105 sequenced on a PromethION sequencer. Initially, five *de novo* assemblies were generated with varying  
106 sequence overlaps (5, 10, 15, 20 and 30kb) using Flye v2.7 and v2.8 [18]. Contigs from the five assemblies  
107 were combined into one assembly by merging contig ends overlapping with >20kb or more determined from

108 LASTZ alignments [19]. The combined assembly was polished with long-reads using PEPPER (v0.0.6) [20] and  
109 Illumina short-reads using pilon (v1.23) [21]. Hi-C data was used to build chromosome sequences. Except for  
110 Hi-C, the assembly pipeline described above was used to create 10 additional genome assemblies for Atlantic  
111 salmon, as well as the brown trout assembly used for determining the standard arrangement of the  
112 inversions (Table S1).

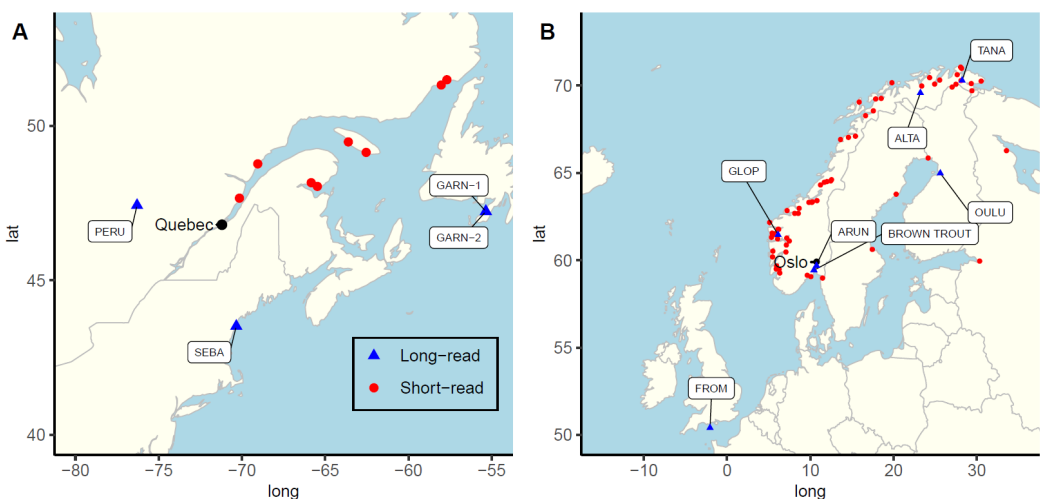
113  
114 *Inversion detection*

115 We detected inversions in the long-read sequenced samples with both read-mapping and assembly  
116 comparisons. For the read-based SV-calling pipeline, see electronic supplementary material. Custom scripts  
117 can be found at [https://github.com/kristinastenlokk/long\\_read\\_SV](https://github.com/kristinastenlokk/long_read_SV). Assembly alignments were made with  
118 Minimap2 v2.23 [22] and were used to verify the candidate inversions from the read-based SV-detection  
119 pipeline, limited to inversions  $\geq 10$ kb that are visible in assembly alignments, and to detect additional large  
120 inversions that read-based methods have low power to detect. This provided a set of 11 high confidence  
121 inversions (for details see Data S1). For inspection of repeat blocks in inversion breakpoints, we created self-  
122 alignments with LASTZ v1.0.4 [19], (Figure S1). These 11 inversions were genotyped by manual inspection of  
123 plots where contigs and nanopore reads were mapped to inversion breakpoints using Minimap2 v2.23 [22].  
124 To validate heterozygous inversions in the reference AQGE, ultra-long reads were created with PromethION  
125 using the Ultra-Long DNA Sequencing Kit and protocol (SQK-ULK001, v.ULK\_9124\_v110\_revA\_24Mar2021).  
126 Figure S2B demonstrates how a AQGE ultra-long read mapped to the AQGE assembly reveals that both  
127 orientations of the inversion are present and resolving that AQGE is heterozygous for the inversion. Figure  
128 S2C shows mapping of a contig from the OULU assembly spanning the repeat structure of the upstream  
129 inversion breakpoint of chr9inv, validating the alternative state of the inversion in this sample.

130  
131 *Illumina short-read mapping and variant calling*

132 For the short-read mapping and variant calling we used Illumina data from the whole genome re-sequencing  
133 of 482 Atlantic salmon sampled from a broad phylogeographic distribution [23] (Figure 1, Data S2). The  
134 Illumina reads were mapped to the Atlantic salmon genome (Ssal\_v3.1; GCA\_905237065.2) using the bcbio-  
135 nextgen v1.2.3 pipeline [24] with the bwa-mem aligner v.0.7.17 [25]. Aligned reads were sorted with  
136 Samtools v1.9 [26] and duplicate reads were marked with Sambamba v0.7.1 [27]. Genomic variation was  
137 identified using Google's DeepVariant pipeline v1.1.2 with default parameters [28] and the individual  
138 genotypes were merged using Glnexus v1.2.2 with the 'DeepVariantWGS' configuration [29]. Variants were  
139 then filtered for depth  $>4$  and  $<40$ , genotype quality  $>10$  followed by missingness  $<30\%$  in vcftools (v0.1.16).

140  
141



142  
143 **Figure 1: Map of sampling sites.** Location of wild Atlantic salmon and brown trout samples used in study. Blue triangles designate  
144 nanopore long-read sequenced samples used for inversion detection and red dots indicate populations sampled and sequenced with  
145 Illumina short-read in North America (A) and Europe (B).



146

147 *Identifying tag-SNPs and inversions type in population samples*

148 To identify SNPs ‘tagging’ the standard and inverted haplotypes we used Illumina short-read genotype data  
149 for the same individuals as were genotyped using long-read data (Table S1). This comparison allowed us to  
150 validate the ability of the short-read data to correctly call inversion genotypes as determined from long-read  
151 assembly comparisons (see above). Haplotypes were determined separately for Europe and North America  
152 as there is strong genetic structure between the continents that is not linked to the inversion. No haplotype  
153 structure was determined for chr16inv because no short-read SNPs were called in this short (~77kb)  
154 inversion. To phase inversions variants and SNPs we used Princess v0.01 [30] with default parameters. After  
155 phasing, unphased loci were removed, and SNPs and inversion variants were refined and merged using  
156 Jasmine v1.1.0 [31]. Exemplified by chr18inv, a total of 146 SNPs, which perfectly match inversion types in  
157 the 11 long-read samples, were defined as tag-SNPs and used to genotype eight populations in North America  
158 using the short-read data (Figure 2A). Scripts can be found at  
159 [https://github.com/mariesaitou/supergenes\\_inversions](https://github.com/mariesaitou/supergenes_inversions).

160

161 *Inversion dating*

162 For inversions with inversion-linked haplotype structure and length greater than 100kb (~1cM) (Data S1), we  
163 dated the inversions using the split function in smc++ v1.15.2 [32]. We used alternate homozygotes as  
164 defined from the haplotype analysis (Figure 2B, Figure S3), i.e. standard and inverted homozygotes, as  
165 populations and SNPs within the inverted region. These conditions were only met for chr18inv in North  
166 America (n=41 standard and 21 inverted homozygotes). A mutation rate of  $1.06 \times 10^{-8}$  was inferred by  
167 comparing sequence divergence between long-read sequenced individuals from Europe and North America  
168 and an estimated divergence time of 0.5MY [33].

169

170 *Genotype-Environment associations*

171 To test if inversions were associated with adaptive variation, we tested for genotype-environment  
172 associations (GEA) determined using the Latent Factor Mixed Model (LFMM) approach [34] in the R package  
173 lea (R v4.1)[35]. LFMM fits a linear mixed-model with population structure controlled simultaneously to  
174 model estimation using latent factors, where the expected number of genetic clusters (K) is the latent factor,  
175 which was estimated using admixture (v1.23) [36]. Environment associations were tested on the pooled  
176 European (n=402) and the North American (n=80) samples separately because the strong differentiation  
177 between these lineages would confound associations and some inversions were only polymorphic in one  
178 group. Environment associations were tested for all SNPs on chromosomes containing an inversion (8  
179 chromosomes, 1.07-1.23 million SNPs). False discovery control was employed using the Benjamini-Hochberg  
180 procedure with alpha thresholds of 0.05 and 0.01 across all tests. Variants were phased and imputed using  
181 Beagle v5.2 [37] (burnin 3, interactions 12, phase states 280) and then filtered for minor allele frequency  
182 >5%. Environment variables tested related to thermal, precipitation and river size conditions in the spawning  
183 and juvenile habitat expected to exert selection pressures on salmon. The individual river parameters were  
184 obtained from the WorldClim database for an arc of 30 translating to 1 square km at the river mouth  
185 (<https://www.worldclim.org>) to ensure comparable data quality and availability for all rivers. Air temperature  
186 has been shown to represent water temperature in Norway except at low temperatures [38], likely because  
187 winter ice cover in some rivers can lead to discrepancies in air and water temperatures. Annual temperature,  
188 and additionally the temperature in the coldest and warmest quarters, were selected as these influence the  
189 overwinter survival and growth potential respectively. Inversions were inferred to have adaptive potential  
190 where they overlap with multiple variant associations suggesting that the inversion has the potential to link  
191 different adaptive variants and is capable of becoming a supergene. The frequency of associated loci was  
192 calculated for inversion homozygotes, to avoid any influence of phasing errors, by summing the allele count  
193 and dividing by twice the number of homozygous individuals for each arrangement.

194

195 *Mutation load*

196 To test for the accumulation of deleterious mutations we predicted missense variants with snpEff v5.0e [39]  
197 on variant calls (filtered with vcfTools v0.1.16 on minor allele count = 2). PROVEAN scores (PROVEAN v1.1.5

198 [40]) were computed to assess the impact of the detected missense variants for each protein using the  
199 Ensembl Rapid Release annotation of GCA\_905237065.2. PROVEAN scores  $\geq |2.5|$  were defined as  
200 deleterious. We compared the density of deleterious mutations within inversions to the genome wide level  
201 by dividing the number of significant PROVEAN scores by the number of genes per megabase (Mb) to obtain  
202 the mutation load per gene and Mb. We used the Wilcoxon Rank Sum test to test for significant enrichment  
203 inside inversions.

204

#### 205 *Detection of indels within inversions*

206 Indels were called using the long-read based detection pipeline (electronic supplementary material,  
207 Methods). Insertions and deletions were filtered based on length using a common minimum cut-off of 50bp  
208 and a maximum of 100kb, as earlier studies have shown that it is challenging to reliably call longer insertions  
209 [30]. To assess indel enrichment within inversions, we compared the indel density inside inversions by indel  
210 densities within corresponding homeologous regions in the Atlantic salmon genome. Indel density was  
211 calculated as the number of indels per sequence length.

212

## 213 **Results and Discussion**

### 214 *Detection and characterization of inversions*

215 Long-read data from 11 Atlantic salmon sampled across the species' range were used to systematically  
216 identify inversions, allowing us to detect and compare inversions that had not formed supergenes to those  
217 with supergene characteristics. Read-based methods for structural variant detection had low precision  
218 regarding the position and size of the inversions, indicating a high number of false positives. These  
219 inconsistencies are likely because of the complex breakpoint repeat structures as the only large inversion  
220 detected by these methods had simple non-repetitive breakpoints (chr18inv). In contrast, assembly-based  
221 methods were much more reliable for detecting and genotyping inversions. Assembly methods detected a  
222 modest but reliable set of 11 inversions, with five inversions being larger than 1.5 Mb and containing multiple  
223 genes (summarized in Data S1). All inversions detected by the method were observed in more than one  
224 individual corroborating that the inversions are real and polymorphic. The increasing availability of multiple  
225 assemblies (pangenomes) will facilitate the detection of inversions by this method in more species.

226 Further, alignment of chromosome sequences in the Atlantic salmon reference (AQGE; GCA\_905237065.2)  
227 with syntenic regions in the sister species brown trout, shows that for chr4inv, chr11inv3, chr16inv, chr18inv,  
228 chr22inv and chr26inv the reference has the standard configuration, whereas for chr3inv, chr9inv, chr10inv,  
229 chr11inv1 and chr11inv2 it has the inverted orientation (Figure S4).

230

### 231 *Characterization of inversion breakpoints*

232 Inversion breakpoints can have functional impacts, e.g. by disrupting coding genes, and can impact the  
233 evolution of inversions, but can be difficult to sequence through as they are often highly repetitive. To  
234 characterize inversion breakpoints, we analyzed both nanopore reads and multiple *de novo* assemblies of  
235 the 11 Atlantic salmon (Table S1). Five inversions (chr3inv, chr9inv, chr11inv3, chr22inv and chr26inv) are  
236 flanked by complex tandem repeats, four of which (chr3inv, chr9inv, chr22inv and chr26inv), have similar  
237 tandem motifs at both breakpoints (see Figure S1). Shared repeat expansions on either end of these  
238 inversions may indicate recurrence and may make the development of a supergene less likely by permitting  
239 recombination among haplotypes. chr18inv is the only large, multigene inversion with no obvious repeat  
240 structures at the inversion breakpoints (Figure S1-H). For the large inversions with matched tandem repeats  
241 at both breakpoints we were unable to detect extended LD and the development of divergent haplotypes,  
242 suggesting they are younger or recurrent inversions that may be unlikely to become supergenes. While the  
243 small chr3inv did have haplotype structure, this did not reflect the inversion genotype and so probably  
244 reflects the small size of the region.

245

246 Three of the inversions have possible functional impacts through gene-disrupting breakpoints. The upstream  
247 breakpoint of chr18inv breaks in intron 1 of *MRC2-like* (Figure 3C), making the gene likely to become non-  
248 functional. Mannose receptor genes have immune-related functions and have been shown to be upregulated  
249 following bacterial infection in fish [41]. Two copies of *MRC2-like* are found nearby that may compensate for  
250 the breakpoint mutation, preventing negative fitness effects (Figure S5). Chr22inv disrupts genes at both  
251 breakpoints, breaking *TGM2-like* and *VRK3* at the upstream (Figure S6A) and *DNASE1L3* at the downstream

252 breakpoint (Figure S6B). *TGM2* is involved in cell death, pro-inflammatory response [42] and is associated  
253 with the environment in Arctic Charr [43]. *VRK3* is also involved in apoptosis and inflammatory processes  
254 [44]. *DNASE1L3* is known to mediate degradation of DNA during apoptosis [45]. We observed individuals  
255 homozygous for the gene breaks for chr18inv and chr22inv, implying that they are not lethal, as observed for  
256 some inversion supergenes (e.g. [5, 8]). Finally, the downstream breakpoint of chr26inv disrupts  
257 *BAT1/DDX39B* (Figure S7), a helicase involved in RNA metabolism and inflammatory disease [46].  
258 Duplications are present at both the upstream (~300kb; pos. 52,003,636-52,306,925) and downstream  
259 (~100kb; pos. 53,816,770-53,913,337) breakpoints of chr26inv (Figure S1 J), however, no protein-coding  
260 genes are duplicated and so the functional consequences are unclear. Negative effects of breakpoint-induced  
261 gene disruptions may prevent these inversions from successfully spreading. However, many genes in the  
262 Atlantic salmon genome have functional duplicates originating from the salmonid whole genome duplication,  
263 which may compensate for eventual functional consequences of some of these gene disruptions.

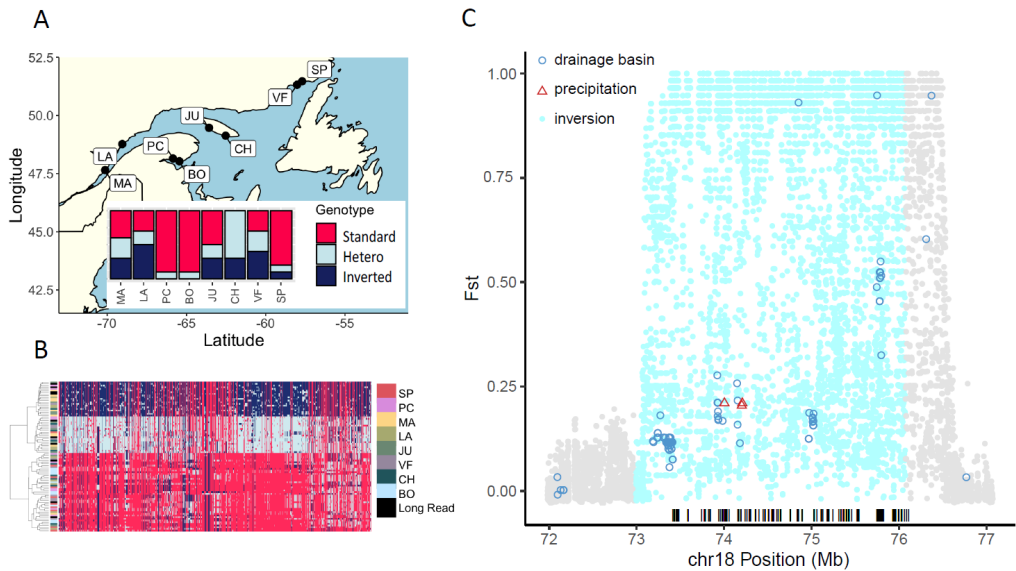
#### 264 265 *Accumulation of deleterious mutations and indel enrichment*

266 Recombination suppression makes inversions vulnerable to the accumulation of deleterious mutations,  
267 which could be important in determining their fate. If recessive deleterious mutations accumulate it can  
268 result in associative overdominance, where heterozygous individuals are more fit [2, 4, 10]. None of the  
269 inversions showed significant enrichment of deleterious mutations (Wilcoxon Rank Sum test;  $P > 0.05$ ) (Figure  
270 S8). For three of the inversions (chr3inv, chr18inv and chr26inv) there is a >2x enrichment of small indels  
271 compared to their corresponding homeologous region in the salmon genome (Table S2). One 260 bp deletion,  
272 fixed in the standard chr18inv arrangement in North American populations, overlaps the 3'-end of *P2RY5*  
273 (*P2Y* purinoceptor 5; ENSSSAG00000044266), indicating that it may be of functional importance (Figure S9).  
274 However, for chr18inv there is no evidence for a deleterious impact of inversion homozygosity, since both  
275 haplotypes were frequent in our population samples. These results suggest that it may be too early in the  
276 evolution of chr18inv for sufficient deleterious mutations to have accumulated to influence the maintenance  
277 of the inversion.

#### 278 279 *Haplotype structure within inversions*

280 A key aspect of supergene formation and invasion is reduced recombination leading to strong linkage  
281 disequilibrium (LD) and divergent haplotypes. Only six inversions have haplotype structures extending across  
282 the inversion: chr3inv, chr11inv2, chr11inv3 and chr18inv in North America, and chr4inv and chr11inv1 in  
283 Europe (Figure S3). However, only for chr4inv and chr18inv did this structure match the inversion genotype  
284 from long-read analyses (Data S3). This suggests that the other inversions are either recurrent, so little  
285 structure has developed, or rare such that the haplotypes are dominated by one configuration. The large  
286 multigene chr18inv inversion is frequent (0.38) across eight North American populations (Figure 2A) and the  
287 short chr4inv was also frequent in Europe (0.31). Consistent with the clear haplotype structure,  $F_{ST}$  between  
288 alternative homozygotes for chr18inv was strongly elevated across the inversion (Figure 2C). The elevation  
289 of  $F_{ST}$  extends beyond the downstream breakpoint, suggesting that recombination is also suppressed for  
290 ~490kb downstream of the inversion. We found no indication in the long-read assemblies for further linked  
291 structural variants that could explain this extended recombination suppression, but such may be present in  
292 other individuals.

293  
294



295  
 296 **Figure 2. Chr18inv haplotype structure and  $F_{ST}$  between inverted and standard arrangements.** A. Location of samples and estimated  
 297 chr18inv haplotype frequencies in different Canadian rivers based on short-read sequence data. BO: Bonaventure, CH,  
 298 De\_la\_Chaloupe, JU: Jupiter, LA: Laval, MA: Malbaie(Charlevoix), PC: Petite\_riviere\_Cascapedia, SP: Saint-Paul, VF, Du\_Vieux\_Fort.  
 299 B. Red: reference homozygous type, Light blue: heterozygous type, Navy blue: inversion homozygous type. B. Haplotype structure  
 300 within the chr18inv region (The 1000th to the 2000th variants were selected from the 4828 variants to reduce the computational load  
 301 for the effective visualization) based on short-read sequence data in North American populations. Individuals with long-read  
 302 sequences are highlighted in black in the left bar. The haplotype structure of the entire inversion is described in Figure S3. C.  $F_{ST}$  (Weir  
 303 and Cockerham) between homozygotes with alternative inversion orientations showing environment associated SNPs with  $p < 0.05$ ,  
 304 black bars under plot show the positions of tag-SNPs used to genotype the inversion.  
 305

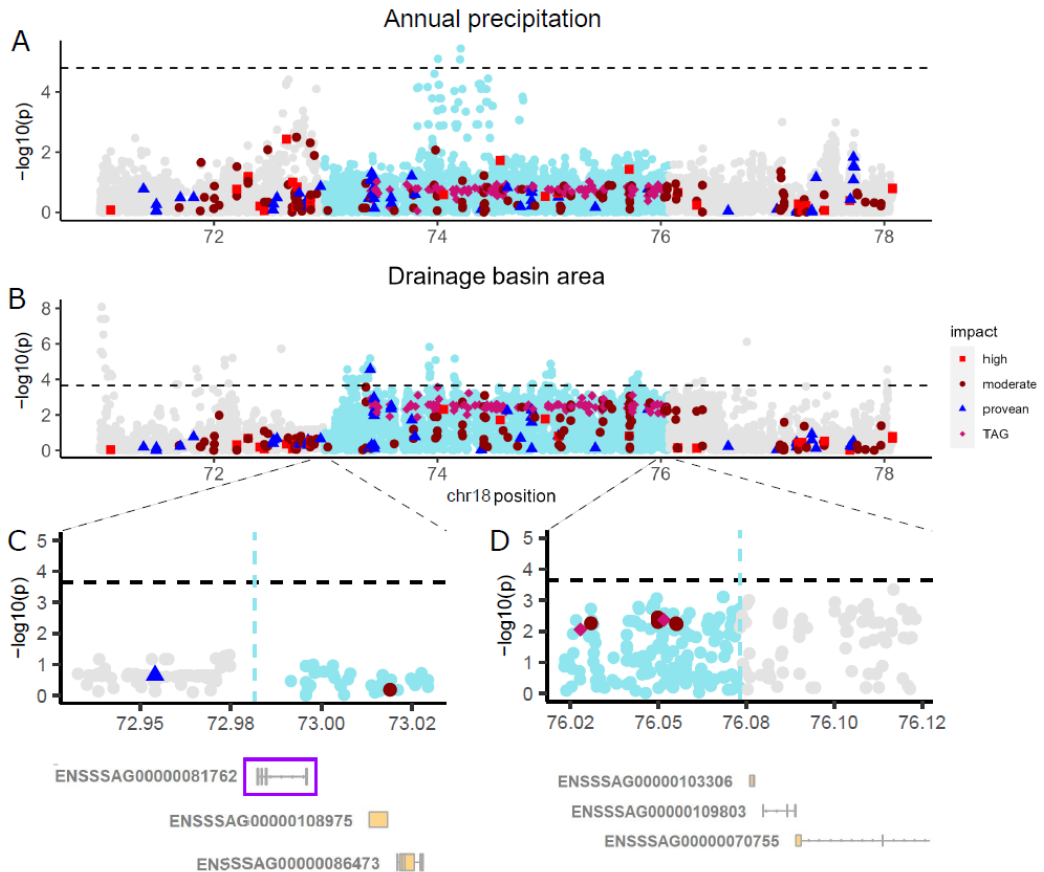
### 306 *Dating of the inversions*

307 Only one inversion could be dated because it had inversion-linked haplotype structure and was >100kb, i.e.  
 308 >~1cM. The chr18inv inversion was estimated to have split from the standard arrangement ~5000  
 309 generations, ~15,000 years ago (Figure S10) making this a young inversion, originating about the time of the  
 310 last glacial retreat.

### 312 *Genotype-environment associations*

313 Local adaptation with gene flow, as occurs in Atlantic salmon, has been suggested as a driver for the  
 314 establishment of inversions because recombination suppression within the inversion can protect locally co-  
 315 adapted variants from being broken apart by the influx of migrant variation [10, 12]. To become a supergene  
 316 the region of suppressed recombination should link together multiple adaptive variants that behave as a  
 317 single haplotype[1]. Larger inversions are expected to capture more genes and locally adapted alleles, which  
 318 may help to explain their greater likelihood of being recruited as supergenes [10]. Consistent with this  
 319 prediction, among 11 inversions only four large (>1.5Mb) multigene inversions (Data S1) overlapped with  
 320 environment association peaks, three of which overlapped with multiple environments. Chr9inv and chr26inv  
 321 have associations with two different environments in Europe and chr9inv with three in North America (Figure  
 322 S11), which were weak to moderately correlated ( $r^2 = 0.01-0.43$  Table S3). Chr18inv is only polymorphic in  
 323 North America where multiple associations were found with two environmental variables, annual mean  
 324 precipitation (LFMM  $p < 0.05$ ) and drainage basin area (LFMM  $p < 0.05$ ) (see Figure 3, Figure S11a, S11c),  
 325 which are weakly correlated ( $r^2 = 0.17$ , Table S3). None of the associations overlapped the breakpoints,  
 326 suggesting these are not involved in environmental adaptation to these variables. Further work will be  
 327 required to determine if the gene disrupting breakpoint is adaptive, or just tolerated. These results suggest  
 328 that the potential for large inversions to capture and link adaptive clusters is common, in line with

329 expectations [47]. However, only chr18inv had both environment associations and a strong inversion-linked  
 330 haplotype structure (Figure 11c, Figure S3), indicative of supergene formation, suggesting the presence of  
 331 pre-existing adaptive variation is not sufficient alone or favorable allele combinations were not captured in  
 332 the three other inversions.  
 333  
 334  
 335



336  
 337 **Figure 3. Genotype-environment associations for chr18inv.** Associations with **A** annual precipitation and **B** drainage basin area,  
 338 functional- and tag-SNPs are highlighted for the inversion (blue) and 2Mb flanking (grey). Significant associations, dashed horizontal  
 339 line, indicates significance level  $p < 0.05$ . Red squares (high) and dark red points (moderate) show functional impact estimated by  
 340 SNPeff on protein-coding genes. Blue triangles show significant deleterious mutations estimated by PROVEAN, whereas pink  
 341 diamonds represent tag-SNPs for chr18inv. Zoom-in of breakpoints **C** and **D** show one gene (*MRC2-like*, ENSSSAG00000081762)  
 342 overlapping breakpoint at  $\sim 72.98$ Mb (purple frame).  
 343

### 344 Capture and accumulation of adaptive variation

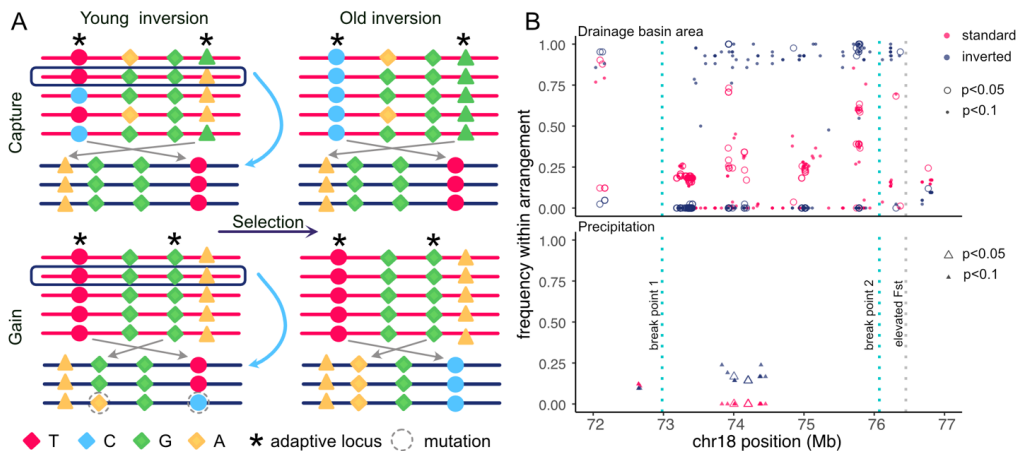
345 Whether environment associations arise from the capture of pre-existing variation and, therefore, are  
 346 important in establishing the inversion, or accumulate over time after inversions have occurred, is still  
 347 unclear [6, 12, 13, 47]. When an inversion is first formed it is expected that the inverted arrangement will be  
 348 invariant, having captured a single standard haplotype (Figure 4A). In contrast, initially the standard  
 349 arrangement will still carry any allelic variation that was previously segregating in the population, including  
 350 that captured by the inversion [10]. Over time, this variation will be lost by drift and selection in the standard  
 351 arrangement and the inverted arrangement haplotypes will gain variation via new mutations [10] (Figure 4A).  
 352 For the inversion to be maintained the linkage among adaptive variants within the inverted arrangement  
 353 should confer higher fitness than the same variants within the recombining standard arrangement. Only  
 354 three significant variants were found to be strongly differentiated (LFMM  $p < 0.05$  and  $F_{st} > 0.8$ ) across

355 chr18inv (Figure 2C), one of which is located outside of the inverted region, but within the area of suppressed  
 356 recombination downstream of the distal breakpoint. The inverted arrangement is fixed or nearly fixed for all  
 357 alleles associated with drainage basin area at  $p < 0.05$ , while in the standard arrangement these alleles have  
 358 intermediate frequencies (Figure 4B). This pattern explains the elevated but moderate  $F_{st}$ s for most adaptive  
 359 variants (Figure 2C) and is consistent with retention of pre-existing adaptive polymorphisms in a young  
 360 inversion. The pattern is less strong for weakly associated SNPs,  $p < 0.1$ , where three SNPs had intermediate  
 361 frequencies in the inverted arrangement. The pattern is reversed for precipitation associated SNPs, all of  
 362 which are fixed in the standard arrangement but are variable in the inverted arrangement, suggesting that  
 363 sufficient time has elapsed since the inversion event to allow the generation of new adaptive variation in the  
 364 inverted arrangement (Figure 4B). These patterns suggest that the inversion has captured previously  
 365 segregating adaptive polymorphisms, linking them within the inversion, but selection and drift have not yet  
 366 removed the pre-existing variation within the standard arrangement. However, at least some adaptive  
 367 variation has emerged within the inverted arrangement.

369 The inversion remains polymorphic in all populations (Figure 2A), co-existence of both arrangements is  
 370 expected if the migration rate is not so high that it leads to swamping. The maximum benefit of an inversion  
 371 is expected when migration is just below this critical level [12]. If the spatial heterogeneity occurs over small  
 372 scales, or environmental variation is continuous we also expect within population inversion polymorphisms  
 373 to persist. All these factors are likely to contribute to the maintenance of within population polymorphism  
 374 here. Lee et al. (2017) [6], also found support for capture in a young inversion (~2.1-8.8ka) in a relative of  
 375 *Arabidopsis*. However, because high levels of self-fertilization would reduce the benefit of recombination  
 376 suppression, invasion of a supergene by this mechanism was difficult to reconcile with model expectations  
 377 [12]. Here we find evidence for capture and accumulation of adaptive variation in an outcrossed species  
 378 where populations are connected by gene flow. Both capture of pre-existing and subsequent accumulation  
 379 of adaptive variation is also suggested for a butterfly mimicry supergene in an analysis presented by Jay et al in  
 380 this special issue [13].

381

382



383

384

385

386

387

388

389

390

391

392

393

394

395

396

**Figure 4. Capture and accumulation of adaptive variation by inversions.** **A.** When an inversion first occurs the inverted arrangement captures a single invariant haplotype, while the standard arrangement will still have multiple variable haplotypes including pre-existing adaptive variation (upper left). Following the inversion event, the inverted arrangement may accumulate new adaptive mutations that are not present in the standard arrangement (lower left). Over time selection for alternatively coadapted allelic combinations results in the fixation of adaptive variation in the standard arrangement (upper right) and selective sweeps and fixation of adaptive new mutations in the inverted arrangement (lower right) making the origin of the variation hard to determine. **B.** Frequency of environment-associated SNPs (LFMM open symbols  $p < 0.05$ , closed symbols  $p < 0.1$  circles: drainage basin, triangles: precipitation) in standard and inverted homozygotes pooled across eight North American populations. Drainage basin associated SNPs are almost all fixed or nearly fixed in the inverted arrangement, especially for  $p < 0.05$ , as expected for a young inversion, whereas most variants segregate at intermediate frequencies in the standard arrangement. Only three SNPs were nearly fixed for alternative alleles between standard and inverted arrangements, a pattern expected to occur in older inversions. In contrast, precipitation associated SNPs are fixed in the standard arrangement, suggesting variation in the inverted arrangement has accumulated since the inversion occurred.

397  
398  
399  
400  
401  
402  
403  
404  
405  
406  
407  
408  
409  
410  
411  
412  
413  
414  
415  
416  
417  
418  
419  
420  
421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431  
432  
433  
434  
435  
436  
437  
438  
439  
440  
441  
442  
443  
444  
445  
446  
447  
448  
449

## Conclusions

Our genome wide survey of inversions in Atlantic salmon detected 11 highly reliable inversions. Of these, two showed evidence of inversion driven haplotype formation. Only large multigene inversions overlapped with adaptive variants as detected by GEA, and among these only chr18inv also had inversion-linked haplotype structure. For chr18inv we found evidence that the adaptive variants linked to the inverted haplotype also segregate as ancestral polymorphisms as they are still present in the standard arrangement haplotypes. Additionally, adaptive variation has accumulated within the inverted haplotype since its formation. These findings support that both the capture of preexisting variation and subsequent accumulation of variation has been important in forming this emerging supergene. Three of the 11 inversions had breakpoints that disrupted genes. For chr18inv, the disruption could be compensated for by local duplicates. Our results suggest that multiple processes contribute to the formation of supergenes from inversions, e.g. both capture and accumulation of adaptive variation and tolerated breakpoint mutations, but do not support an early role for deleterious mutation load.

## Data availability

Illumina whole-genome sequencing data of 482 individuals are available in projects (PRJEB38061). The long-read genome assemblies have been submitted to ENA, project accessions listed in Table S1. Environmental information for the environment associations is contained in Data S2. The authors declare that all data supporting the findings of this study are available within the paper and its electronic supplementary material [48].

Authors' contributions. K.S.: data curation, formal analysis, investigation, methodology, visualization, writing—original draft, writing—review and editing; M.S.: formal analysis, investigation, visualization, writing—original draft, writing—review and editing; L.R.-J.: formal analysis; T.N.: formal analysis, methodology, writing—review and editing; M.M.: formal analysis, methodology, writing—review and editing; M.Á.: formal analysis, methodology, writing—review and editing; M.K.: formal analysis, methodology, writing—review and editing; N.J.B.: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, supervision, visualization, writing— original draft, writing—review and editing; S.L.: conceptualization, data curation, formal analysis, methodology, project administration, resources, supervision, writing—original draft, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

## Acknowledgements

The study was supported by The Research Council of Norway (grant nos. 275310 and 221734). We thank Sarah Lehnert, Ian Bradbury, Louis Bernatchez, Cooke Aquaculture Inc., AquaGen AS, Craig Primmer, Jamie Stevens, Harald Sægrov, Jenny Jensen and Thrond Haugen for providing samples for the nanopore sequencing. We acknowledge the use of the Orion computing cluster at the Norwegian University of Life Sciences (NMBU). Storage resources were provided by the Norwegian National Infrastructure for Research Data (NIRD, project NS9055K). We thank two anonymous reviewers and the editor for thoughtful comments that improved the manuscript.

## References

- [1] Thompson, M. J. & Jiggins, C. D. 2014 Supergenes and their role in evolution. *Heredity* (Edinb) 113, 1-8. (doi:10. 1038/hdy.2014.20).
- [2] Gutiérrez-Valencia, J., Hughes, P. W., Berdan, E. L. & Slotte, T. 2021 The Genomic Architecture and Evolutionary Fates of Supergenes. *Genome Biology and Evolution* 13, evab057. (DOI:https://doi.org/10.1093/gbe/evab057).
- [3] Pearse, D. E., Barson, N. J., Nome, T., Gao, G., Campbell, M. A., Abadía-Cardoso, A., Anderson, E. C., Rundio, D. E., Williams, T. H., Naish, K. A., et al. 2019 Sex-dependent dominance maintains migration

450 supergene in rainbow trout. *Nature Ecology & Evolution* 3, 1731-1742.  
451 (DOI:<https://doi.org/10.1038/s41559-019-1044-6>).

452 [4] Jay, P., Chouteau, M., Whibley, A., Bastide, H., Parrinello, H., Llaurens, V. & Joron, M. 2021 Mutation  
453 load at a mimicry supergene sheds new light on the evolution of inversion polymorphisms. *Nature Genetics*  
454 53, 288-293. (DOI:<https://doi.org/10.1038/s41588-020-00771-1>).

455 [5] Lamichaney, S., Fan, G., Widemo, F., Gunnarsson, U., Thalmann, D. S., Hoepfner, M. P., Kerje, S.,  
456 Gustafson, U., Shi, C., Zhang, H., et al. 2016 Structural genomic changes underlie alternative reproductive  
457 strategies in the ruff (*Philomachus pugnax*). *Nature Genetics* 48, 84-88.  
458 (DOI:<https://doi.org/10.1038/ng.3430>).

459 [6] Lee, C.-R., Wang, B., Mojica, J. P., Mandáková, T., Prasad, K. V. S. K., Goicoechea, J. L., Perera, N.,  
460 Hellsten, U., Hundley, H. N., Johnson, J., et al. 2017 Young inversion with multiple linked QTLs under  
461 selection in a hybrid zone. *Nature Ecology & Evolution* 1, 0119. (DOI:<https://doi.org/10.1038/s41559-017-0119-0>).

462 [7] Villoutreix, R., Ayala, D., Joron, M., Gompert, Z., Feder, J. L. & Nasil, P. J. M. E. 2021 Inversion  
463 breakpoints and the evolution of supergenes. *Molecular Ecology*.  
464 (DOI:<https://doi.org/10.1111/mec.15907>).

465 [8] Küpper, C., Stocks, M., Risse, J. E., dos Remedios, N., Farrell, L. L., McRae, S. B., Morgan, T. C.,  
466 Karlionova, N., Pinchuk, P., Verkuil, Y. I., et al. 2016 A supergene determines highly divergent male  
467 reproductive morphs in the ruff. *Nature Genetics* 48, 79-83. (DOI:<https://doi.org/10.1038/ng.3443>).

468 [9] Giraldo-Deck, L. M., Loveland, J. L., Goymann, W., Tschirren, B., Burke, T., Kempenaers, B., Lank, D. B. &  
469 Küpper, C. 2022 Intralocus conflicts associated with a supergene. *Nature Communications* 13, 1384.  
470 (DOI:<https://doi.org/10.1038/s41467-022-29033-w>).

471 [10] Faria, R., Johannesson, K., Butlin, R. K. & Westram, A. M. 2019 Evolving Inversions. *Trends in Ecology &*  
472 *Evolution* 34, 239-248. (DOI:<https://doi.org/10.1016/j.tree.2018.12.005>).

473 [11] Kirkpatrick, M. & Barton, N. 2006 Chromosome Inversions, Local Adaptation and Speciation. *Genetics*  
474 173, 419-434. (DOI:<https://doi.org/10.1534/genetics.117.300572>).

475 [12] Charlesworth, B. & Barton, N. H. 2018 The Spread of an Inversion with Migration and Selection.  
476 *Genetics* 208, 377-382. (DOI:<https://doi.org/10.1534/genetics.117.300426>).

477 [13] Jay P, Leroy M, Le Poul Y, Whibley A, Arias M, Chouteau M, Joron M. 2022 Association mapping of  
478 colour variations in a butterfly provides evidences that a supergene locks together a cluster of adaptive loci.  
479 *Phil. Trans. R. Soc. B* 377, 20210193. (doi:10.1098/rstb.2021.0193)

480 [14] Hutchings, J. A. & Jones, M. E. B. 1998 Life history variation and growth rate thresholds for maturity in  
481 Atlantic salmon, *Salmo salar*. *Canadian Journal of Fisheries and Aquatic Sciences* 55, 22-47.  
482 (DOI:<https://doi.org/10.1139/d98-004>).

483 [15] Barson, N. J., Aykanat, T., Hindar, K., Baranski, M., Bolstad, G. H., Fiske, P., Jacq, C., Jensen, A. J.,  
484 Johnston, S. E., Karlsson, S., et al. 2015 Sex-dependent dominance at a single locus maintains variation in  
485 age at maturity in salmon. *Nature* 528, 405-408. (DOI:<http://dx.doi.org/10.5061/dryad.23h4q>).

486 [16] Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Kent, M. P., Nome, T., Hvidsten, T. R., Leong, J. S.,  
487 Minkley, D. R., Zimin, A., et al. 2016 The Atlantic salmon genome provides insights into rediploidization.  
488 *Nature* 533, 200-205. (DOI:<https://doi.org/10.1038/nature17164>).

489 [17] Bourret, V., Kent, M. P., Primmer, C. R., Vasemägi, A., Karlsson, S., Hindar, K., McGinnity, P., Verspoor,  
490 E., Bernatchez, L. & Lien, S. 2013 SNP-array reveals genome-wide patterns of geographical and potential  
491 adaptive divergence across the natural range of Atlantic salmon (*Salmo salar*). *Molecular Ecology* 22, 532-  
492 551. (DOI:<https://doi.org/10.1111/mec.12003>).

493 [18] Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. J. N. b. 2019 Assembly of long, error-prone reads  
494 using repeat graphs. *Nature biotechnology* 37, 540-546. (DOI:<https://doi.org/10.1038/s41587-019-0072-8>).

495 [19] Harris, R. S. 2007 Improved pairwise alignment of genomic DNA, The Pennsylvania State University.

496 [20] Shafin, K., Pesout, T., Chang, P.-C., Nattestad, M., Kolesnikov, A., Goel, S., Baid, G., Kolmogorov, M.,  
497 Eizenga, J. M. & Miga, K. H. J. N. m. 2021 Haplotype-aware variant calling with PEPPER-Margin-DeepVariant  
498 enables high accuracy in nanopore long-reads. *Nature methods* 18, 1322-1332.  
499 (DOI:<https://doi.org/10.1038/s41592-021-01299-w>).

500 [21] Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q.,  
501 Wortman, J. & Young, S. K. J. P. o. 2014 Pilon: an integrated tool for comprehensive microbial variant



503 detection and genome assembly improvement. *PLOS one* 9, e112963.  
504 (DOI:<https://doi.org/10.1371/journal.pone.0112963>).

505 [22] Li, H. 2018 Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094-3100.  
506 (DOI:<https://doi.org/10.1093/bioinformatics/bty191>).

507 [23] Bertolotti, A. C., Layer, R. M., Gundappa, M. K., Gallagher, M. D., Pehlivanoglu, E., Nome, T., Robledo,  
508 D., Kent, M. P., Røssæg, L. L. & Holen, M. M. J. N. c. 2020 The structural variation landscape in 492 Atlantic  
509 salmon genomes. *Nature communications* 11, 1-16. (DOI:<https://doi.org/10.1038/s41467-020-18972-x>).

510 [24] Chapman, B., Kirchner, R., Pantano, L., Naumenko, S., De Smet, M., Beltrame, L., Khotiainsteva, T.,  
511 Sytchev, I., Guimera, R. V., Kern, J., et al. 2021 bcbio/bcbio-nextgen: (v1.2.9). (Zenodo).

512 [25] Li, H. 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*.

513 [26] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R.  
514 J. B. 2009 The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078-2079.  
515 (DOI:<https://doi.org/10.1093/bioinformatics/btp352>).

516 [27] Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. J. B. 2015 Sambamba: fast processing of  
517 NGS alignment formats. *Bioinformatics* 31, 2032-2034.  
518 (DOI:<https://doi.org/10.1093/bioinformatics/btv098>).

519 [28] Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J.,  
520 Nguyen, N. & Afshar, P. T. J. N. b. 2018 A universal SNP and small-indel variant caller using deep neural  
521 networks. *Nature biotechnology* 36, 983-987. (DOI:<https://doi.org/10.1038/nbt.4235>).

522 [29] Yun, T., Li, H., Chang, P.-C., Lin, M. F., Carroll, A. & McLean, C. Y. J. B. 2020 Accurate, scalable cohort  
523 variant calls using DeepVariant and GLnexus. *Bioinformatics* 36, 5582-5589.  
524 (DOI:<https://doi.org/10.1093/bioinformatics/btaa1081>).

525 [30] Mahmoud, M., Doddapaneni, H., Timp, W. & Sedlazeck, F. J. J. G. b. 2021 PRINCESS: comprehensive  
526 detection of haplotype resolved SNVs, SVs, and methylation. *Genome biology* 22, 1-17.  
527 (DOI:<https://doi.org/10.1186/s13059-021-02486-w>).

528 [31] Kirsche, M., Prabhu, G., Sherman, R., Ni, B., Aganezov, S. & Schatz, M. C. J. B. 2021 Jasmine:  
529 Population-scale structural variant comparison and analysis. *bioRxiv*.  
530 (DOI:<https://doi.org/10.1101/2021.05.27.445886>).

531 [32] Terhorst, J., Kamm, J. A. & Song, Y. S. 2017 Robust and scalable inference of population history from  
532 hundreds of unphased whole genomes. *Nature Genetics* 49, 303-309.  
533 (DOI:<https://doi.org/10.1038/ng.3748>).

534 [33] King, T. L., Verspoor, E., Spidle, A. P., Gross, R., Phillips, R. B., Koljonen, M. L., Sanchez, J. A. & Morrison,  
535 C. L. 2007 Biodiversity and Population Structure. *The Atlantic Salmon*, 117-166.  
536 (DOI:<https://doi.org/10.1002/9780470995846.ch5>).

537 [34] Frichot, E., Schoville, S. D., Bouchard, G. & François, O. 2013 Testing for associations between loci and  
538 environmental gradients using latent factor mixed models. *Molecular biology and evolution* 30, 1687-1699.  
539 (DOI:<https://doi.org/10.1093/molbev/mst063>).

540 [35] Frichot, E. & François, O. 2015 LEA: An R package for landscape and ecological association studies.  
541 *Methods in Ecology and Evolution* 6, 925-929. (DOI:<https://doi.org/10.1111/2041-210X.12382>).

542 [36] Alexander, D. H., Novembre, J. & Lange, K. 2009 Fast model-based estimation of ancestry in unrelated  
543 individuals. *Genome research* 19, 1655-1664. (DOI:<https://doi.org/10.1101/gr.094052.109>).

544 [37] Browning, B. L., Tian, X., Zhou, Y. & Browning, S. R. J. T. A. J. o. H. G. 2021 Fast two-stage phasing of  
545 large-scale sequence data. *The American Journal of Human Genetics* 108, 1880-1890.  
546 (DOI:<https://doi.org/10.1016/j.ajhg.2021.08.005>).

547 [38] Otero, J., L'Abée-Lund, J. H., Castro-Santos, T., Leonardsson, K., Storvik, G. O., Jonsson, B., Dempson, B.,  
548 Russell, I. C., Jensen, A. J., Baglinière, J.-L., et al. 2014 Basin-scale phenology and effects of climate  
549 variability on global timing of initial seaward migration of Atlantic salmon (*Salmo salar*). *Global Change*  
550 *Biology* 20, 61-75. (DOI:<https://doi.org/10.1111/gcb.12363>).

551 [39] Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X. & Ruden, D. M. J.  
552 F. 2012 A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff:  
553 SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* 6, 80-92.  
554 (DOI:<https://doi.org/10.4161/fly.19695>).

555 [40] Choi, Y. & Chan, A. P. J. B. 2015 PROVEAN web server: a tool to predict the functional effect of amino  
556 acid substitutions and indels. *Bioinformatics* 31, 2745-2747.  
557 (DOI:<https://doi.org/10.1093/bioinformatics/btv195>).

558 [41] Dong, X., Li, J., He, J., Liu, W., Jiang, L., Ye, Y. & Wu, C. 2016 Anti-infective mannose receptor immune  
559 mechanism in large yellow croaker (*Larimichthys crocea*). *Fish & Shellfish Immunology* 54, 257-265.  
560 (DOI:<https://doi.org/10.1016/j.fsi.2016.04.006>).

561 [42] Wentzel, A. S., Petit, J., van Veen, W. G., Fink, I. R., Scheer, M. H., Piazzon, M. C., Forlenza, M., Spaink,  
562 H. P. & Wiegertjes, G. F. 2020 Transcriptome sequencing supports a conservation of macrophage  
563 polarization in fish. *Scientific Reports* 10, 13470. (DOI:<https://doi.org/10.1038/s41598-020-70248-y>).

564 [43] Layton, K. K. S., Snelgrove, P. V. R., Dempson, J. B., Kess, T., Lehnert, S. J., Bentzen, P., Duffy, S. J.,  
565 Messmer, A. M., Stanley, R. R. E., DiBacco, C., et al. 2021 Genomic evidence of past and future climate-  
566 linked loss in a migratory Arctic fish. *Nature Climate Change* 11, 158-165.  
567 (DOI:<https://doi.org/10.1038/s41558-020-00959-7>).

568 [44] Liu, P.-f., Du, Y., Meng, L., Li, X., Yang, D. & Liu, Y. 2019 Phosphoproteomic analyses of kidneys of  
569 Atlantic salmon infected with *Aeromonas salmonicida*. *Scientific Reports* 9, 2101.  
570 (DOI:<https://doi.org/10.1038/s41598-019-38890-3>).

571 [45] Shi, G., Abbott, K. N., Wu, W., Salter, R. D. & Keyel, P. A. 2017 Dnase1L3 Regulates Inflammasome-  
572 Dependent Cytokine Secretion. 8. (DOI:<https://doi.org/10.3389/fimmu.2017.00522>).

573 [46] Szymura, S. J., Bernal, G. M., Wu, L., Zhang, Z., Crawley, C. D., Voce, D. J., Campbell, P.-A., Ranoa, D. E.,  
574 Weichselbaum, R. R. & Yamini, B. 2020 DDX39B interacts with the pattern recognition receptor pathway to  
575 inhibit NF- $\kappa$ B and sensitize to alkylating chemotherapy. *BMC Biology* 18, 32.  
576 (DOI:<https://doi.org/10.1186/s12915-020-0764-z>).

577 [47] Schaal S., Haller B., Lotterhos K. 2022 Inversion Invasions: when the genetic basis of local adaptation is  
578 concentrated within inversions in the face of gene flow. *Philosophical Transactions of the Royal Society B X*,  
579 X. (DOI:<https://doi.org/10.1098/rstb.2021.0200>)

580 [48]. Stenløkk K, Saitou M, Rud-Johansen L, Nome T, Moser M, Árnýasi M, Kent M, Barson NJ, Lien S. 2022  
581 The emergence of supergenes from inversions in Atlantic salmon. *Figshare*. (doi:10.6084/m9.  
582 figshare.c.5983514)

583

## Supplementary Material

**Table S1. Metadata for 11 Atlantic salmon samples sequenced with nanopore long-read technology.** Wild salmon were sampled to represent the four main phylogeographic groups; North American (NAm), Baltic (BAL), Barents/White Sea (BWS) and Atlantic (ATL). The aquaculture sample (AQGE) was sampled from the AquaGen strain originating mainly from the ATL group.

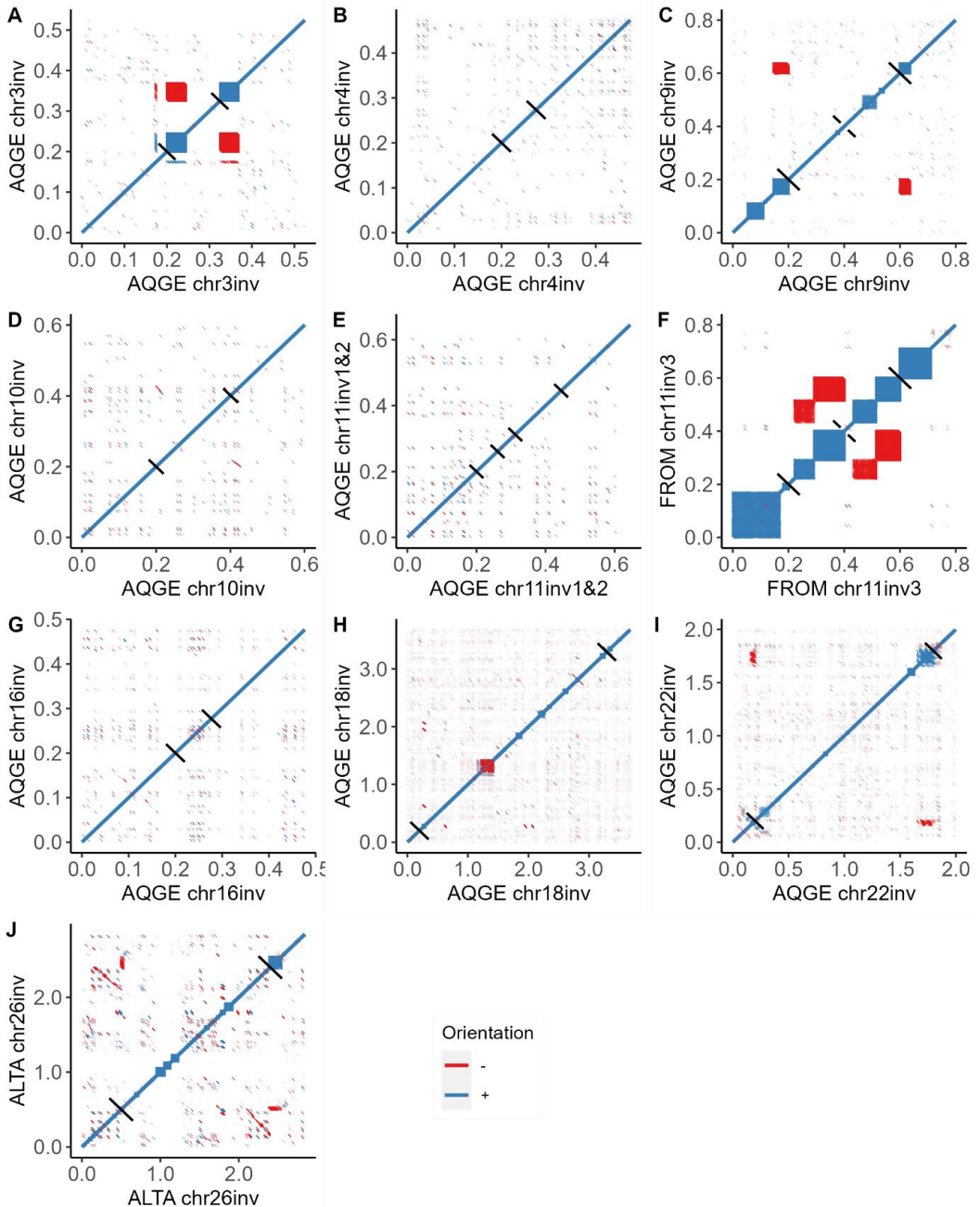
Name	Species	River name	Phylo. group	Country	Gender	Pop. type	Lat, Long	ENA Project accession
AQGE	Atlantic salmon	-	-	Norway	Male	Aquaculture	-	PRJEB43080
GLOP	Atlantic salmon	Gloppenelva	ATL	Norway	Male	Anadromous	61.46N, 6.12E	PRJEB50984
ARUN	Atlantic salmon	Årungselva	ATL	Norway	Male	Anadromous	59.43N, 10.43E	PRJEB50985
ALTA	Atlantic salmon	Altaelva	BWS	Norway	Male	Anadromous	69.58N, 23.22E	PRJEB50986
TANA	Atlantic salmon	Tanaelva	BWS	Norway	Male	Anadromous	70.29N, 28.23E	PRJEB50987
FROM	Atlantic salmon	River Frome	ATL	UK	Male	Anadromous	50.41N, 2.05W	PRJEB50988
OULO	Atlantic salmon	Oulujoki	BAL	Finland	Male	Anadromous	64.98N, 25.61E	PRJEB50989
PERU	Atlantic salmon	Lac Perugia	NAm	Canada	Male	Landlocked	47.43N, 76.30W	PRJEB50990
SEBA	Atlantic salmon	Sebago Lake	NAm	USA	Female	Landlocked	43.52N, 70.34W	PRJEB50991
GARN-1	Atlantic salmon	Garnish River	NAm	Canada	Male	Anadromous	47.23N, 55.35W	PRJEB49548
GARN-2	Atlantic salmon	Garnish River	NAm	Canada	Male	Anadromous	47.23N, 55.35W	PRJEB50992
ARUN	Brown trout	Årungselva	ATL	Norway	Male	Anadromous	59.43N, 10.43E	PRJEB50994

**Table S2. Indel enrichment in inversion regions.** The enrichment was considered as differences between inversion region and homologous region in the genome divided by length of the regions counted and as; i) number of indels and ii) base pairs in indels.

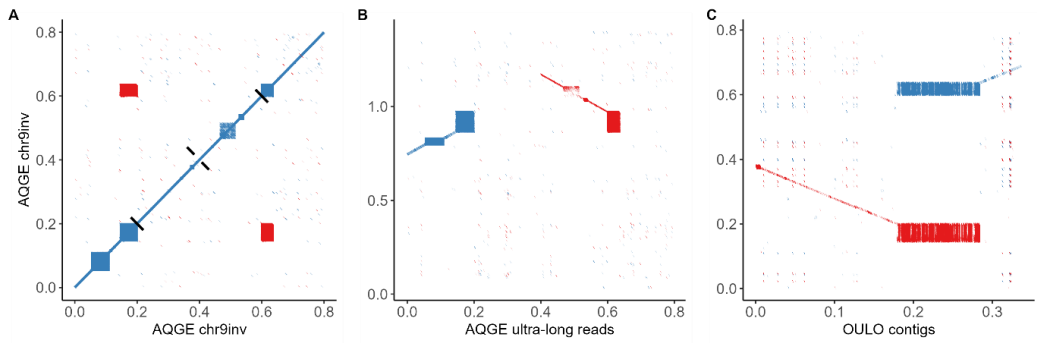
Inversion	Homeologous region	North America		Europe	
		i) $\Delta$ number of indels	ii) $\Delta$ base pairs in indels	i) $\Delta$ number of indels	ii) $\Delta$ base pairs in indels
chr3inv	chr5:21,053,100-21,112,740	2.0	1.3	4.1	1.2
chr4inv	chr13:76,443,570-76,538,610	1.0	1.4	0.0	0.3
chr09inv	chr5:10,657,170-17,305,100	1.0	1.1	1.1	1.0
chr10inv	chr16:3,100,000-3,300,000	0.4	0.8	1.6	0.7
chr11inv1	No homeologous region	-	-	-	-
chr11inv2	No homeologous region	-	-	-	-
chr11inv3	chr26:53,135,590-54,987,680	0.6	0.5	0.6	0.5
chr16inv	No homeologous region	-	-	-	-
chr18inv	chr7:27,565,350-29,960,450	3.4	5.0	2.7	2.6
chr22inv	chr12:92,302,090-93,554,530	0.7	1.0	1.2	0.6
chr26inv	chr11:53,066,290-56,862,090	2.5	4.3	1.9	1.9

**Table S3. Environment correlation matrix.** Correlation ( $r^2$ ) among environmental variables in European (above diagonal) and North American (below diagonal) populations. The correlation is high between mean annual temperature and temperature in the coldest quarter in Europe and the warmest quarter in North America. Mean annual temperature and latitude are highly correlated in both groups.

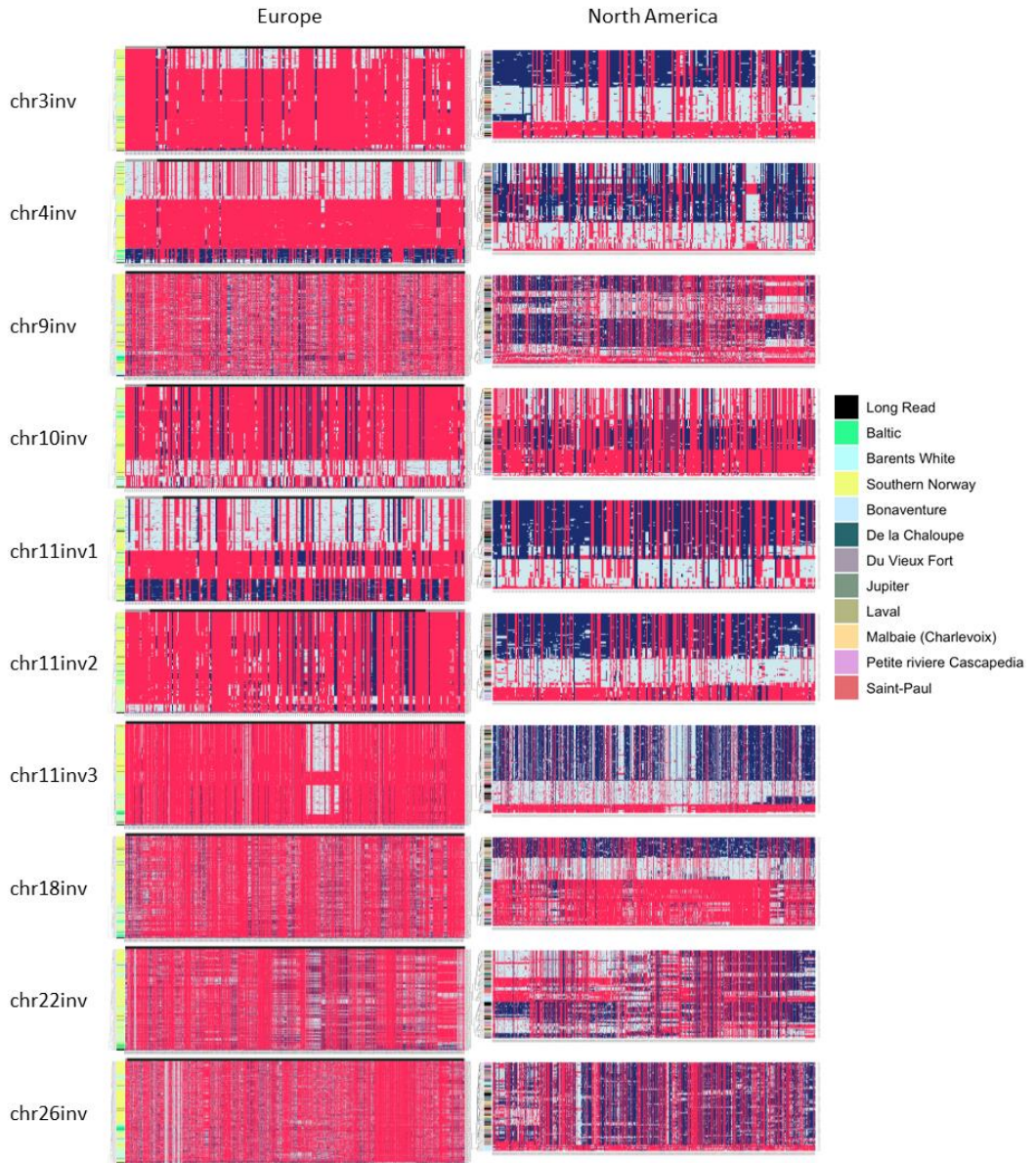
	Lat	Basin size	Temp (mean)	Temp (warmQ)	Temp (coldQ)	IsoTherm	Precip
Lat		-0.14	-0.9	-0.83	-0.75	-0.6	-0.59
Basin size	0.41		-0.01	0.38	-0.18	-0.29	-0.19
Temp (mean)	-0.97	-0.36		0.75	0.95	0.66	0.7
Temp (warmQ)	-0.94	-0.32	0.93		0.51	0.25	0.37
Temp (coldQ)	-0.37	-0.21	0.5	0.14		0.72	0.74
IsoTherm	0.23	0.01	-0.2	-0.45	0.43		0.58
Precip	0.61	0.17	-0.45	-0.44	-0.07	-0.13	



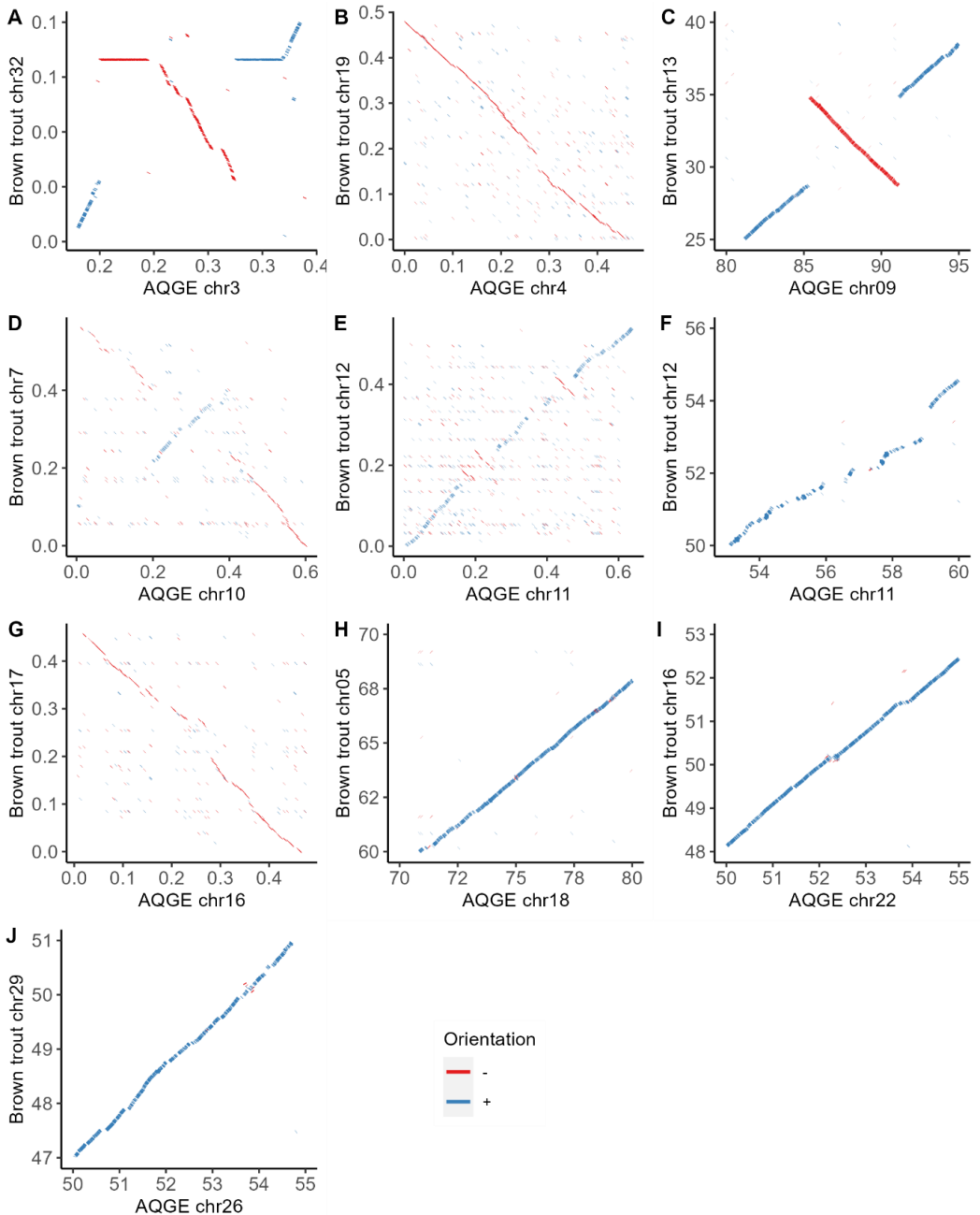
**Figure S1: Tandem repeat structures in inversion breakpoints.** Self-alignments of 11 inversion sequences in Atlantic salmon visualizing inversion breakpoints. Diagonal solid lines indicate breakpoint coordinates for **A** (chr3inv), **C** (chr9inv), **F** (chr11inv3), **I** (chr22inv) and **J** (chr26inv) with repeat structures at the breakpoints, and **B** (Chr4inv), **D** (chr10inv), **E** (chr11inv1 and chr11inv2), **G** (chr16inv) and **H** (chr18inv) without obvious tandem repeat structures at inversion breakpoints. **C** (Chr9inv) and **F** (chr11inv3) display 200kbp up- and downstream of the breakpoints with border between the sequences marked with a dashed line.



**Figure S2. Confirmation of inversion structure of chr9inv using long-reads and contigs spanning the upstream breakpoint.** A. Self-alignment of chr9inv in sample AQGE used to generate the Atlantic salmon reference sequence GCA\_905237065.2. B. Ultra-long reads spanning the upstream breakpoint chr9inv used to determine that AQGE is heterozygous for the inversion. C. Contigs spanning the upstream inversion breakpoint in the OULO sample, validating the alternative state of the inversion in this sample.



**Figure S3. Visualization of haplotype structures within inversion regions.** SNP genotypes were clustered by the hierarchical clustering methods per sample. Navy: alternative homozygous, white: heterozygous, and red: reference homozygous SNPs. Individuals used to construct long-read assemblies are highlighted in black in the left bar, these were used to test if the structure reflects inversion orientation. Variants in 5kb up/downstream and the inversion were used (more than 5% minor allele frequency). Colors indicate the different phylogeographical lineages, Atlantic, White/Barents Sea, Baltic and North American. Chr4inv in European populations and chr18inv in North American populations are the only haplotype structures following the inversion pattern. No SNPs could be called for chr16iv.

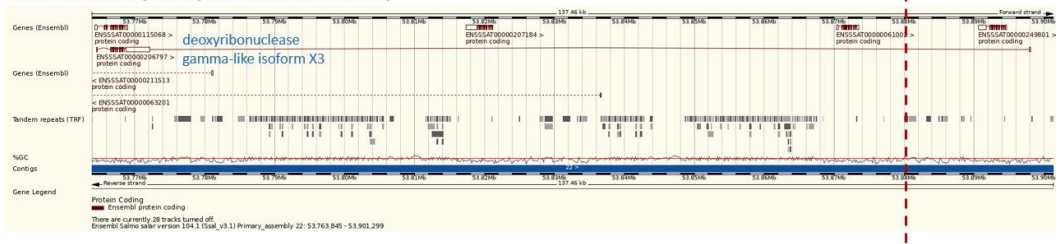


**Figure S4. Comparisons of Atlantic salmon inversions with brown trout to determine ancestral state.** Comparison of inversions in Atlantic salmon reference AQGE with syntenic regions in brown trout. Smaller inversions (<1Mbp) were aligned with LASTZ (A, B, D, E and G) and larger with Minimap2 (C, F, H, I and J). Chr3inv (A), chr9inv (C), chr10inv (D) and chr11inv1&2 (E) show inverted orientation, while the remaining regions show ancestral orientation of the inversions. Colour coding shows alignment orientation (red – and blue +).



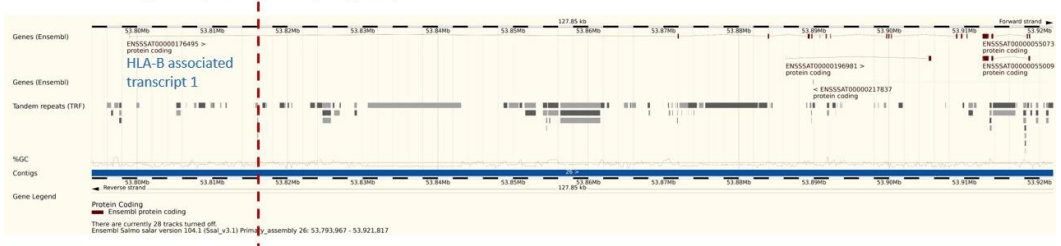


### chr22inv breakpoint (chr22inv:53,888,689)

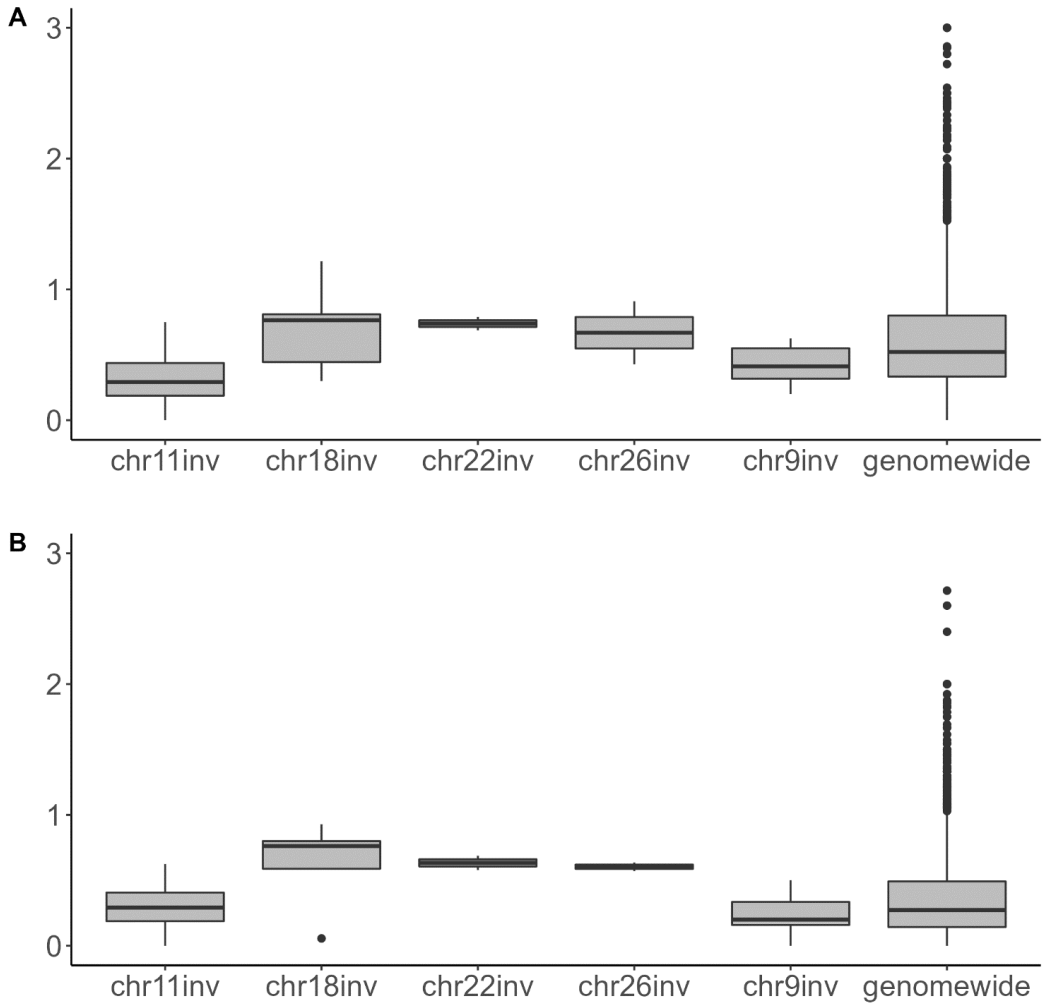


**Figure S6b. Gene annotation in chr22inv downstream breakpoint.** Ensembl Rapid Release annotation (Ssal\_v3.1 version 104.1) for the downstream breakpoint of chr22inv, indicated by vertical red stapled line, disrupt the gene ENSSSAG0000090871 annotated as deoxyribonuclease gamma-like isoform X3 (*DNASE1L3*).

### chr26inv breakpoint (chr26inv:53,816,770)

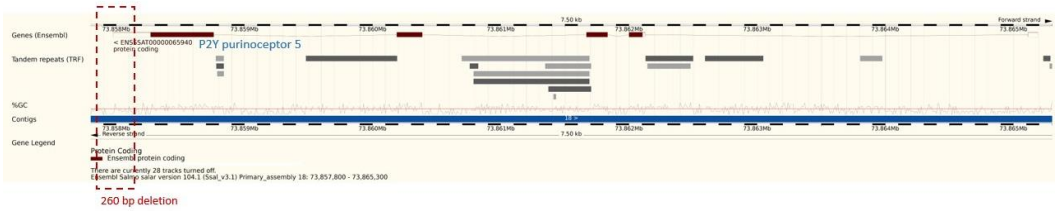


**Figure S7. Gene annotation in chr26inv downstream breakpoint.** Ensembl Rapid Release annotation (Ssal\_v3.1 version 104.1) for the downstream breakpoint of chr26inv, indicated by vertical red stapled line, disrupt the gene ENSSSAG00000107384 annotated as HLA-B associated transcript 1 (*BAT1*).

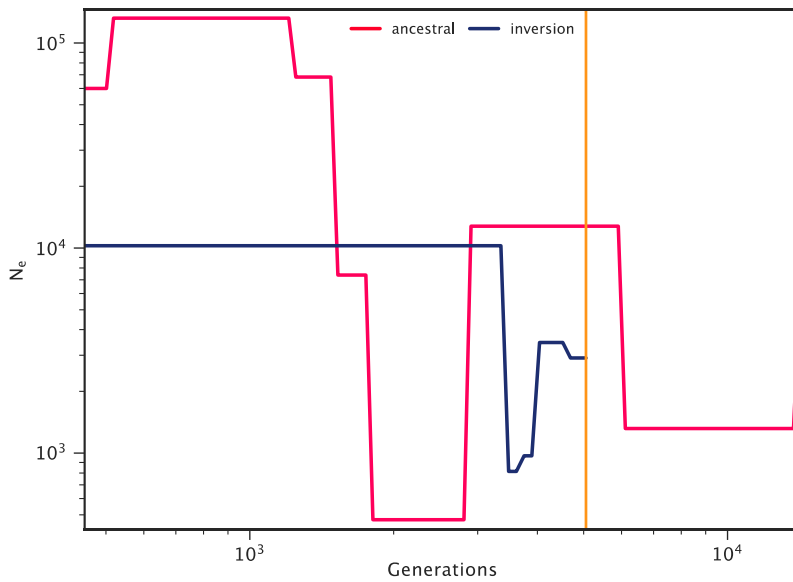


**Figure S8. Accumulation of deleterious mutations in inversions.** Boxplot of the number of deleterious mutations per megabase [0,3] in inversions containing genes compared to the rest of the genome for **A** variants detected in European populations, and **B** variants detected in North American populations.

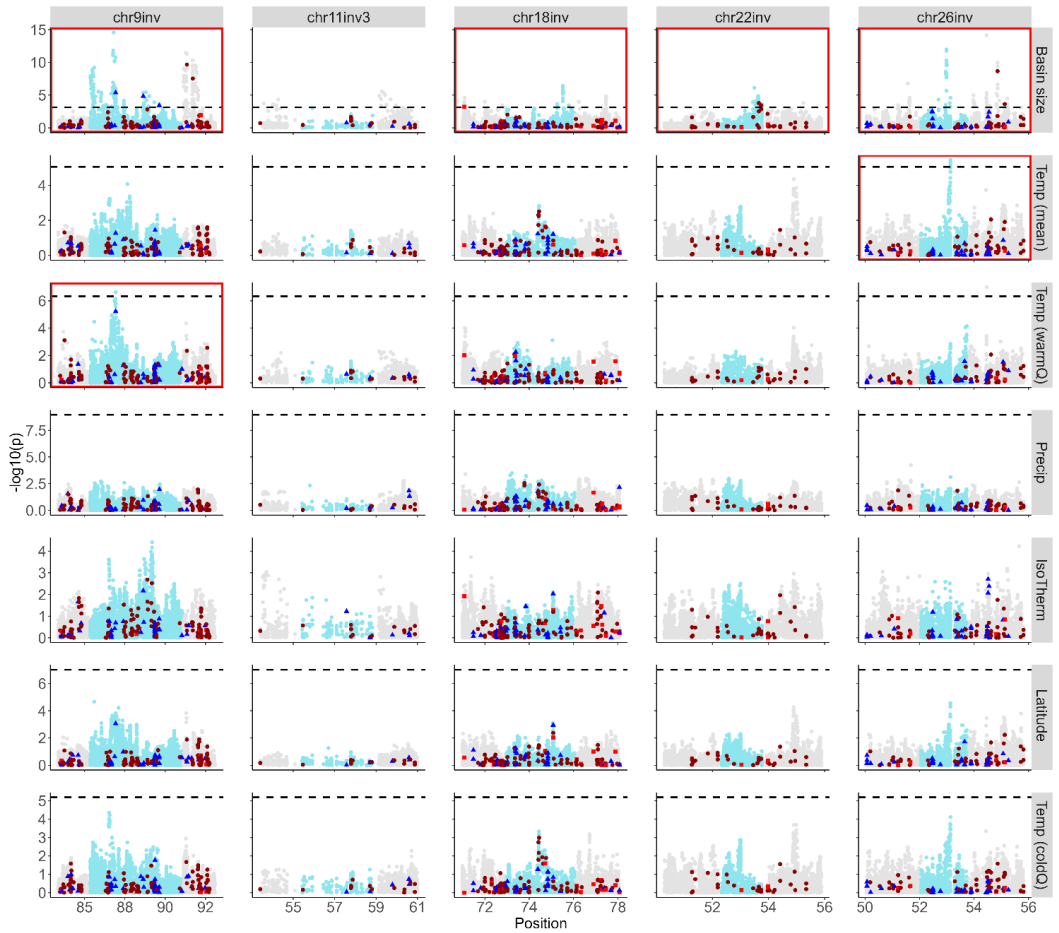
**chr18del:73,857,910-73,858,170**



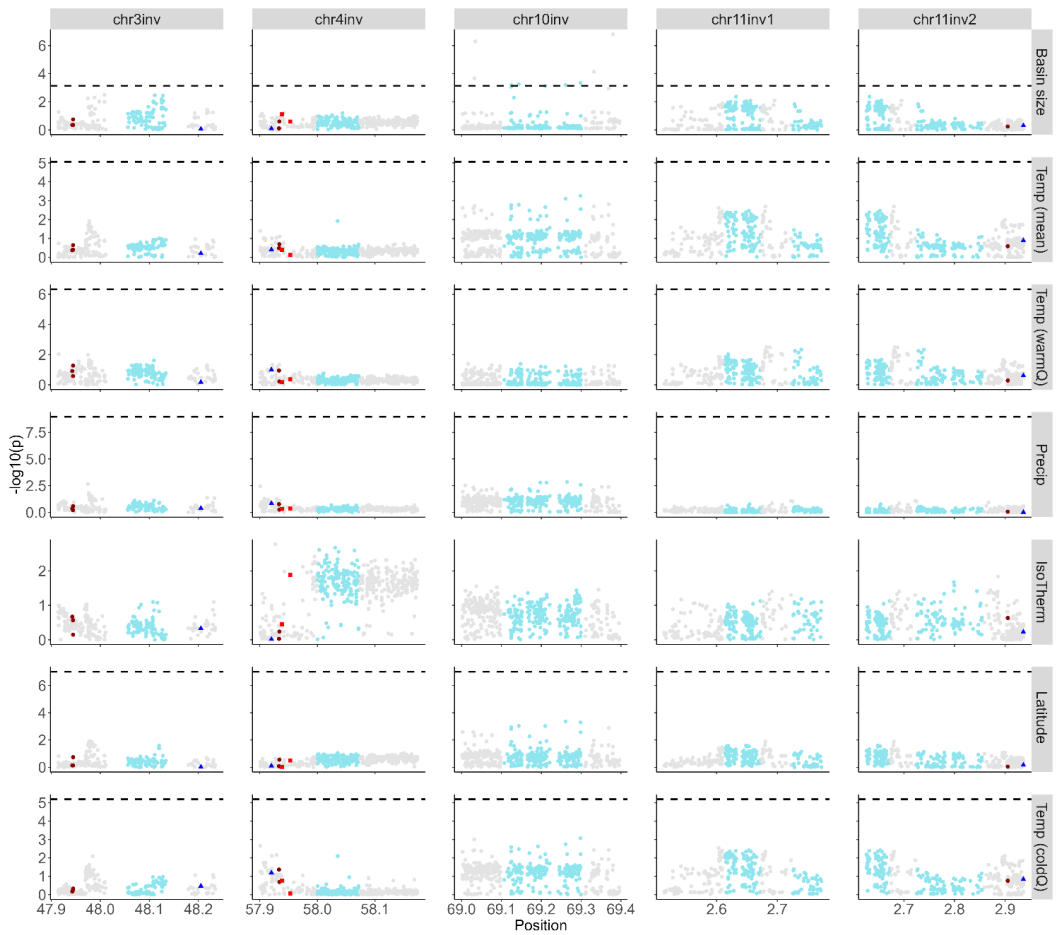
**Figure S9. Deletion segregating with chr18inv.** Location of a 260 bp deletion within chr18inv overlapping the 3'-end of ENSSSAG00000044266 annotated as *P2RY5* (P2Y purinoceptor 5) in the Ensembl Rapid Release annotation (Ssal\_v3.1 version 104.1). The deletion, segregating perfectly with the ancestral configuration of chr18inv in North America populations, is indicated by a red stapled box in the figure.



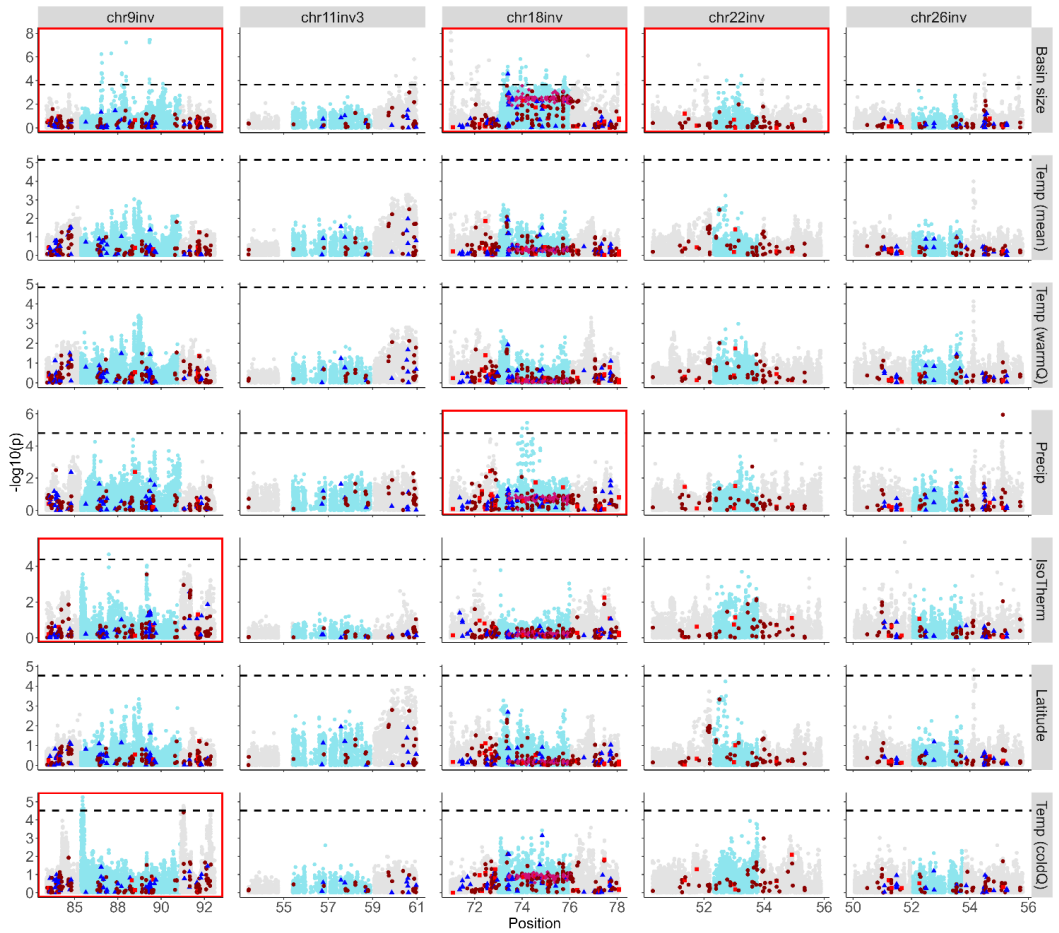
**Figure S10. Dating of chr18inv.** Split plot (smc++) showing estimated date of origin of the inversion. The red line represents the ancestral homozygotes and the blue line the inverted homozygotes. The inversion is estimated to have originated ~5000 generations ago, which is equivalent to approximately 15,000 years with a three-year generation time. This date is when the glacial retreat started, and Atlantic salmon began a postglacial range expansion.



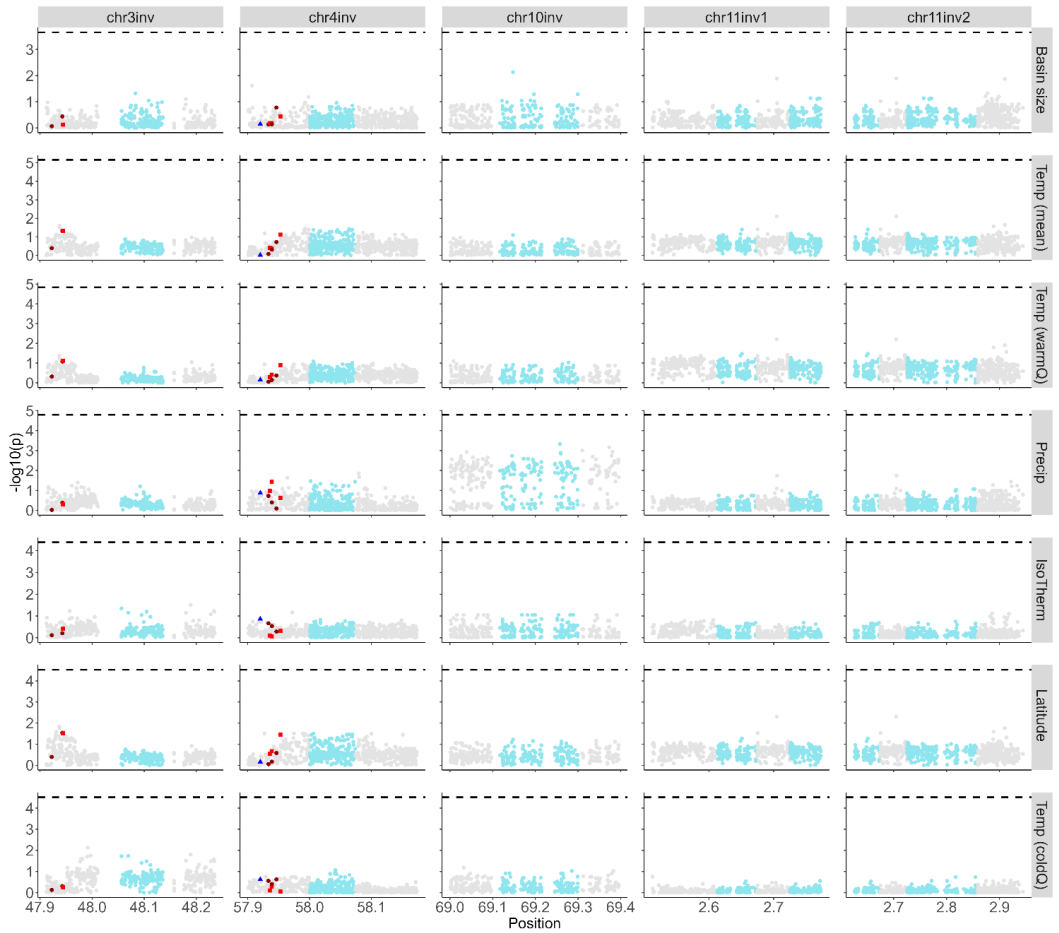
**Figure S11a. Genotype-environment association for SNPs found in European populations within larger inversions (>1Mb).** SNPs inside inverted sequence is blue and 2Mb and up- and downstream flanking sequence is grey for; drainage basin area (sqkm), annual mean temperature, mean temperature of warmest quarter, annual precipitation, isothermality, latitude and mean temperature of coldest quarter. Dashed horizontal line marks the adjusted  $p < 0.05$  significance threshold. Subplots with significant SNPs are marked with red frames. A significance threshold could not be calculated for annual precipitation. Red squares show missense variants annotated as “high” impact by SNPeff, dark red circles show “moderate” impact, and blue triangles mark deleterious mutations ( $\geq |2.5|$  PROVEAN scores).



**Figure S11b. Genotype-environment association for SNPs found in European populations within smaller inversions (<1Mb).** SNPs inside inverted sequence is blue and 2Mb and up- and downstream flanking sequence is grey for; drainage basin area (sqkm), annual mean temperature, mean temperature of warmest quarter, annual precipitation, isothermality, latitude and mean temperature of coldest quarter. Dashed horizontal line marks the adjusted  $p < 0.05$  significance threshold. Subplots with significant SNPs are marked with red frames. A significance threshold could not be calculated for annual precipitation. Red squares show missense variants annotated as “high” impact by SNPeff, dark red circles show “moderate” impact, and blue triangles mark deleterious mutations ( $\geq 2.5$  PROVEAN scores).



**Figure S11c. Genotype-environment association for SNPs found in North American populations within larger inversions (>1Mb).** SNPs inside inverted sequence is blue and 2Mb and up- and downstream flanking sequence is grey for; drainage basin area (sqkm), annual mean temperature, mean temperature of warmest quarter, annual precipitation, isothermality, latitude and mean temperature of coldest quarter. Dashed horizontal line marks the adjusted  $p < 0.05$  significance threshold. Subplots with significant SNPs are marked with red frames. Red squares show missense variants annotated as “high” impact by SnpEff, dark red circles show “moderate” impact, and blue triangles mark deleterious mutations ( $\geq |2.5|$  PROVEAN scores). Pink diamonds represent TAG-SNPs for chr18inv.



**Figure S11d. Genotype-environment association for SNPs found in North American populations within smaller inversions (<1Mb).** SNPs inside inverted sequence is blue and 2Mb and up- and downstream flanking sequence is grey for; drainage basin area (sqkm), annual mean temperature, mean temperature of warmest quarter, annual precipitation, isothermality, latitude and mean temperature of coldest quarter. Dashed horizontal line marks the adjusted  $p < 0.05$  significance threshold. Subplots with significant SNPs are marked with red frames. Red squares show missense variants annotated as “high” impact by SNPeff, dark red circles show “moderate” impact, and blue triangles mark deleterious mutations ( $\geq 2.5$  PROVEAN scores).

## Supplementary methods

### *Long-read based SV-detection*

We detected inversions in the long-read sequenced samples with both read-mapping and assembly comparisons. Read mapping was performed with Winnowmap2 [1], using AQGE as a reference. Sam-files were sorted and converted into BAM-files with Samtools v1.3.1 [2]. The calling was carried out with three separate long-read SV-calling programs, Sniffles v1.2.12 [3], SVIM v1.2.0 [4] and NanoVar 1.3.9 [5], using default settings for SVIM and NanoVar. The minimum number of reads required (-s) was set to 1/3 of the median length when running Sniffles. SVs called as type 'breakpoint', i.e. unresolved variants, and other excess information was filtered out using custom scripts available at [https://github.com/kristinastenlokk/long\\_read\\_SV](https://github.com/kristinastenlokk/long_read_SV). To increase accuracy, VCFs were merged across program with Jasmine v1.1.0 [6], retaining only variants detected with at least two programs.

Inversion calls were additionally filtered by detection in at least two samples and a lower size limit was set to 10kb. Duplicated and overlapping variants were filtered out by stringent manual curation.

[1] Jain, C., Rhie, A., Hansen, N., Koren, S. & Phillippy, A. M. J. b. 2020 A long read mapping method for highly repetitive reference sequences. *bioRxiv*. (DOI:<https://doi.org/10.1101/2020.11.01.363887>).

[2] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. J. B. 2009 The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079. (DOI:<https://doi.org/10.1093/bioinformatics/btp352>).

[3] Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., Von Haeseler, A. & Schatz, M. C. J. N. m. 2018 Accurate detection of complex structural variations using single-molecule sequencing. *Nature methods* **15**, 461-468. (DOI:<https://doi.org/10.1038/s41592-018-0001-7>).

[4] Heller, D. & Vingron, M. J. B. 2019 SVIM: structural variant identification using mapped long reads. *Bioinformatics* **35**, 2907-2915. (DOI:<https://doi.org/10.1093/bioinformatics/btz041>).


[5] Tham, C. Y., Tirado-Magallanes, R., Goh, Y., Fullwood, M. J., Koh, B. T., Wang, W., Ng, C. H., Chng, W. J., Thiery, A. & Tenen, D. G. J. G. b. 2020 NanoVar: accurate characterization of patients' genomic structural variants using low-depth nanopore sequencing. *Genome biology* **21**, 1-15. (DOI:<https://doi.org/10.1186/s13059-020-01968-7>).

[6] Kirsche, M., Prabhu, G., Sherman, R., Ni, B., Aganezov, S. & Schatz, M. C. J. B. 2021 Jasmine: Population-scale structural variant comparison and analysis. *bioRxiv*. (DOI:<https://doi.org/10.1101/2021.05.27.445886>).



# PAPER III

# Genome assembly, structural variants, and genetic differentiation between lake whitefish young species pairs (*Coregonus* sp.) with long and short reads

Claire Mérot<sup>1,2</sup>  | Kristina S. R. Stenløkk<sup>3</sup> | Clare Venney<sup>1</sup> | Martin Laporte<sup>1,4</sup> | Michel Moser<sup>3</sup> | Eric Normandeau<sup>1</sup> | Mariann Árnýasi<sup>3</sup> | Matthew Kent<sup>3</sup> | Clément Rougeux<sup>1</sup> | Jullien M. Flynn<sup>5</sup> | Sigbjørn Lien<sup>3</sup> | Louis Bernatchez<sup>1</sup>

<sup>1</sup>Département de Biologie, Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Québec, Québec, Canada

<sup>2</sup>UMR 6553 Ecobio, OSUR, CNRS, Université de Rennes, Rennes, France

<sup>3</sup>Department of Animal and Aquacultural Sciences (IHA), Faculty of Life Sciences (BIOVIT), Centre for Integrative Genetics (CIGENE), Norwegian University of Life Sciences (NMBU), Ås, Norway

<sup>4</sup>Ministère des Forêts, de la Faune et des Parcs (MFFP) du Québec, Québec, Québec, Canada

<sup>5</sup>Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York, USA

## Correspondence

Claire Mérot, Département de biologie, Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Québec, QC, Canada.

Email: [claire.merot@gmail.com](mailto:claire.merot@gmail.com)

## Funding information

This research was supported by a Discovery research grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) to L.B., the Canadian Research Chair in genomics and conservation of aquatic resources, as well as Ressources Aquatiques Québec (RAQ). C.M. was supported by a Banting Postdoctoral fellowship from the Government of Canada.

**Handling Editor:** Loren Rieseberg

## Abstract

Nascent pairs of ecologically differentiated species offer an opportunity to get a better glimpse at the genetic architecture of speciation. Of particular interest is our recent ability to consider a wider range of genomic variants, not only single-nucleotide polymorphisms (SNPs), thanks to long-read sequencing technology. We can now identify structural variants (SVs) such as insertions, deletions and other rearrangements, allowing further insights into the genetic architecture of speciation and how different types of variants are involved in species differentiation. Here, we investigated genomic patterns of differentiation between sympatric species pairs (Dwarf and Normal) belonging to the lake whitefish (*Coregonus clupeaformis*) species complex. We assembled the first reference genomes for both *C. clupeaformis* sp. Normal and *C. clupeaformis* sp. Dwarf, annotated the transposable elements and analysed the genomes in the light of related coregonid species. Next, we used a combination of long- and short-read sequencing to characterize SVs and genotype them at the population scale using genome-graph approaches, showing that SVs cover five times more of the genome than SNPs. We then integrated both SNPs and SVs to investigate the genetic architecture of species differentiation in two different lakes and highlighted an excess of shared outliers of differentiation. In particular, a large fraction of SVs differentiating the two species correspond to insertions or deletions of transposable elements (TEs), suggesting that TE accumulation may represent a key component of genetic divergence between the Dwarf and Normal species. Together, our results suggest that SVs may play an important role in speciation and that, by combining second- and third-generation sequencing, we now have the ability to integrate SVs into speciation genomics.

## KEYWORDS

population genomics, speciation, structural variants, teleost, transposable elements, whole genome sequencing

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2022 The Authors. *Molecular Ecology* published by John Wiley & Sons Ltd.

## 1 | INTRODUCTION

Understanding the processes underlying the evolution of species and how genomes diverge during speciation is a fundamental goal of evolutionary genomics (Jiggins, 2019; Seehausen et al., 2014). The accumulation of genomic data has allowed scientists to test evolutionary scenarios and infer the timing and circumstances of species divergence (Wolf & Ellegren, 2017). Reciprocally, knowledge about the ecological, geographical and demographic context of speciation helps to interpret the patterns of genetic differentiation between species (Jiggins, 2019; Ravinet et al., 2017). However, the genome-wide landscape of differentiation should be interpreted with caution as it results from complex interactions between gene flow, recombination, demography and selection (Cruickshank & Hahn, 2014; Ravinet et al., 2017; Stevison & McGaugh, 2020). Analysing differentiation between evolutionarily “young” pairs of species has nevertheless proven to be informative, revealing widespread heterogeneity among and between chromosomes (Henderson & Brelsford, 2020; Martin et al., 2019), sometimes identifying genes underlying reproductive isolation (Hejase et al., 2020), and informing about the number and distribution of divergent loci (Dufresnes et al., 2021). Cases of “natural replicates,” including species pairs with similar ecological and phenotypic divergence, are of particular interest, along with instances of repeated hybridization due to secondary contacts. These instances provide important insights into the genomic architecture of species differentiation (Nadeau & Kawakami, 2019) and have revealed that similar patterns between pairs of species may be the result of both (i) shared genetic features such as low-recombination areas in which intraspecific diversity is depleted by linked selection and interspecific  $F_{ST}$  is inflated (Burri et al., 2015); and (ii) shared barrier loci under divergent selection or involved in reproductive isolation (Marques et al., 2016; Meier et al., 2018).

Most of our knowledge on speciation genomics is based on single-nucleotide polymorphisms (SNPs), mainly because such variants are easily accessible with short-read sequencing (Ho et al., 2019; Mérot et al., 2020). However, genomes also vary in structure with loss, gain or rearrangement of sequences between individuals and between species. Such structural variants (SVs) are now recognized to be ubiquitous and to affect a larger fraction of the genomes than SNPs (Catanach et al., 2019; Feulner et al., 2013). SVs may also have large phenotypic effects, may impact recombination and may be involved in speciation (Feulner & De-Kayne, 2017; Kirkpatrick & Barton, 2006; Wellenreuther & Bernatchez, 2018). The best recognized cases are large chromosomal rearrangements such as inversions or fusions, which are hypothesized to favour speciation by preventing recombination between alternative haplotypes (Faria & Navarro, 2010). This is supported by empirical evidence that large rearrangements can accumulate genetic incompatibilities between closely related species of *Drosophila* (Noor et al., 2001) or fish (Berdan et al., 2021). Whole-genome duplication events are particularly prone to favour rapid diversification (Landis et al., 2018) because the rediploidization of duplicated paralogues may differ between lineages and generate hybrid incompatibilities, as observed in yeast (Scannell et al., 2006).

However, small SVs, such as insertions, deletions and small duplications, may also contribute to reproductive isolation. For instance, a duplicated gene in *Drosophila melanogaster* leads to hybrid male sterility (Ting et al., 2004) while in crows a 2.25-kb transposon indel underlies plumage differences, a trait involved in mate choice between two crow species (Weissensteiner et al., 2020). More generally, the insertion, deletion, duplication and/or misregulation of transposable elements (TEs) appear to be responsible for bursts of diversification and various pre- and postzygotic barriers, particularly in plants (Serrato-Capuchina & Matute, 2018) but also in vertebrates (Laporte et al., 2019). Overall, a better understanding of the genomic architecture of species differentiation requires the integration of SVs into speciation genomics (Feulner & De-Kayne, 2017; Mérot et al., 2020; Nadeau & Kawakami, 2019). Moreover, considering both SNPs and SVs is essential to understand the cumulative effects of those different forms of genetic variation on speciation.

Two aspects of long-read sequencing, combined with the development of new bioinformatics tools, have made it possible to investigate SVs between genomes (Ho et al., 2019; Logsdon et al., 2020). First, long-reads have improved the contiguity and quality of genome assemblies, which is particularly relevant for large and complex genomes as well as for regions riddled with repeated elements (Huddleston et al., 2014). Second, long reads can be directly used to detect SVs by aligning the sequences on a reference and analysing split-reads and coverage (Mahmoud et al., 2019). Together, these have proven very powerful for making catalogues of SVs within and between species. For instance, a human genome carries on average 4,442 SVs detected by short reads (Abel et al., 2020) and 27,662 SVs detected with long reads (Chaisson et al., 2019). Potential restrictions when generating long reads are the requirement for high-molecular-weight DNA, and potentially higher costs and lower quality. Consequently, population-level analysis of SVs via long reads is not as accessible as short-read sequencing. One promising possibility is to combine technologies by performing a first step of SV discovery on a limited set of high-quality samples sequenced with long reads, and a second step of SV genotyping on more samples sequenced with short reads (Logsdon et al., 2020; Mérot et al., 2020).

The lake whitefish, *Coregonus clupeaformis*, is a species complex present in numerous cold water lakes throughout North America. In the northeastern part of the continent, it comprises two reproductively isolated species, referred to as *C. clupeaformis* sp. Normal and *C. clupeaformis* sp. Dwarf, which differ ecologically by occupying the benthic and the limnetic habitat, respectively (Bernatchez et al., 2010a; Gagnaire et al., 2013a). Demographic modelling and the analysis of mitochondrial lineages showed that the two species originated from two glacial lineages that started to diverge in allopatry during the last glaciation, roughly 60,000 years ago, before coming into secondary contact about 12,000 years ago (Bernatchez & Dodson, 1990; Jacobsen et al., 2012; Rougeux et al., 2017). This secondary contact occurred independently in several lakes of a suture zone of northeastern America, and provoked a strong character displacement in the Dwarf species toward the use of the planktonic trophic niche, further enhancing speciation through ecological divergence

(Bernatchez et al., 2010b; Landry et al., 2007). The two species show limited gene flow (estimated between one and 30 migrants per generation in the two lakes under study; Rougeux et al., 2017), and the rare hybrids have low fitness due to malformation, early mortality, ecological mismatch and reduced fertility (Bernatchez et al., 2010a; Renaut & Bernatchez, 2011; Rogers & Bernatchez, 2006). Habitat divergence is associated with species differences in a series of morphological, life-history, physiological, transcriptomic and cytological traits (Dalziel et al., 2017; Dion-Côté et al., 2015; Laporte et al., 2015, 2016; Rogers & Bernatchez, 2007; Rogers et al., 2002). The process of ecological and phenotypic divergence following secondary contact probably occurred independently, but with the same genetic background, in several postglacial lakes (Rougeux et al., 2017). Multiple pairs of sympatric species thus provide valuable natural replicates to investigate parallelism and the genetic architecture of speciation. Moreover, as for all salmonid species, *C. clupeaformis* ancestors have undergone a past whole-genome duplication about 80–100 million years ago followed by ongoing rediploidization (Allendorf & Thorgaard, 1984; Lien et al., 2016; Macqueen & Johnston, 2014), resulting in a large, complex genome of ~2.4–3.5 Gb depending on the estimates (Hardie & Hebert, 2003; Lockwood et al., 1991). Therefore, structural genetic polymorphism is expected to be extensive in *C. clupeaformis*, though current studies have not assessed the contribution of SVs to differentiation between Dwarf and Normal species.

In this study, we used a combination of long- and short-read sequencing (Figure 1) to investigate the genetic architecture of speciation and address the contribution of SVs to the genomic differentiation of *C. clupeaformis* sp. Normal and *C. clupeaformis* sp. Dwarf. The main goal was to provide high-quality genomic resources for *C. clupeaformis* in order to investigate parallel and nonparallel genomic patterns of differentiation between Dwarf and Normal species in two independent North American lakes. First, we assembled the reference genome of *C. clupeaformis* sp. Normal based on one sample sequenced with long reads and a genetic map. We documented the specificities of the genome to explore the remaining traces of previous whole-genome duplication and annotated the whitefish TEs. Second, we generated a catalogue of SVs varying between and within Dwarf and Normal species using three data sets: assembly comparison with a *de novo* assembly of a sympatric *C. clupeaformis* sp. Dwarf individual, high-quality long reads of two samples (one Dwarf and one Normal), and short reads of 32 samples (17 Dwarf and 15 Normal) at medium coverage (5×). Third, we analysed genome-wide landscapes of differentiation between Dwarf and Normal species in two lakes by genotyping the whole catalogue of SVs using genome-graph-based mapping, as well as SNPs, in the 32 samples sequenced with short reads. We tested the hypothesis that the two lakes would show parallel patterns of differentiation between Dwarf and Normal and compared signals observed with different kinds of variants (SNPs vs. SVs). Our study provides a unique opportunity to characterize the contribution of both SNPs and SVs to differentiation between young species pairs, with important implications for our understanding of speciation in general.

## 2 | METHODS

### 2.1 | Sampling, DNA extraction, and sequencing of *Coregonus clupeaformis*

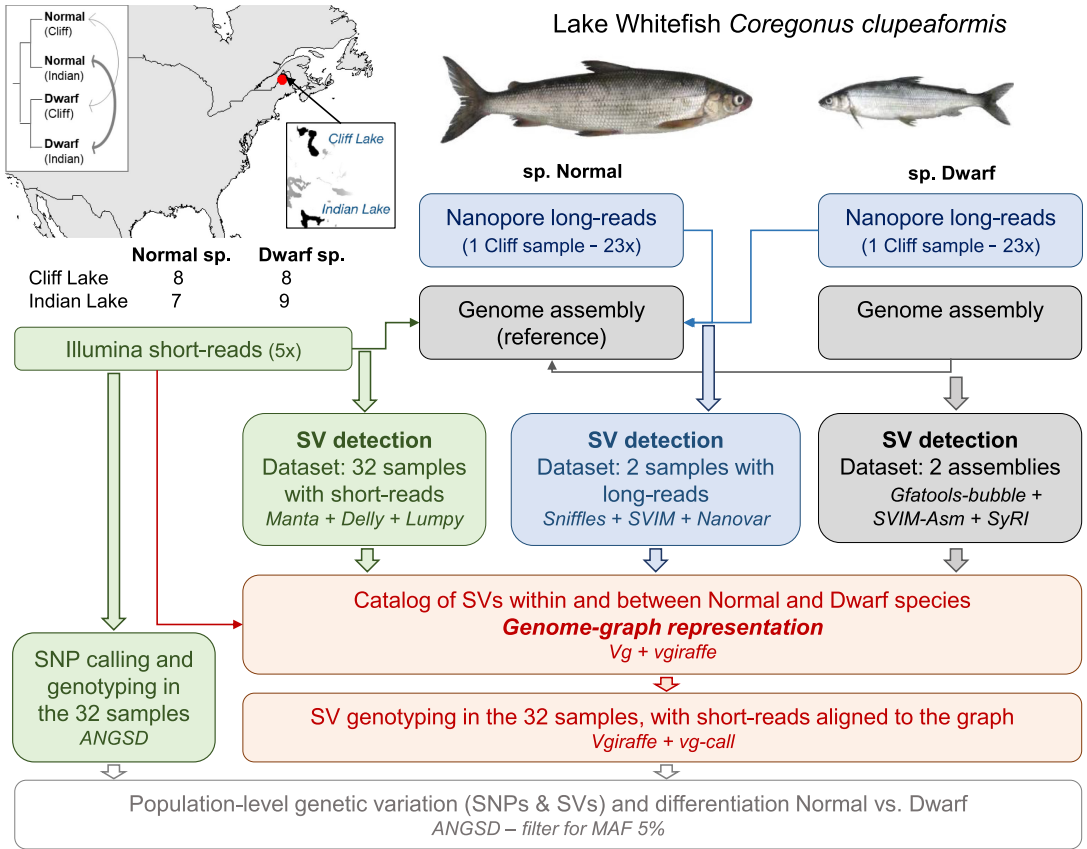
#### 2.1.1 | Long-read sequencing

For long-read sequencing and the assembly of both reference genomes, we sampled one adult of *C. clupeaformis* sp. Normal and one adult *C. clupeaformis* sp. Dwarf from Cliff Lake, Maine (46.3991, -69.2491). Fish were caught live with gillnets, killed, immediately dissected to obtain fresh tissue samples and sexed following a protocol described previously in Evans and Bernatchez (2012). Muscle samples were flash frozen in liquid nitrogen and later stored at -80°C. High-molecular-weight DNA was extracted from 40 mg frozen liver from both species using a Qiagen Genomic Tip 100/G kit (Qiagen). DNA integrity was assessed visually by separating fragments on a 0.5% TAE agarose gel, which revealed a predominant band of high-molecular-weight DNA >45 kb. Smaller fragments were removed by performing size selection, with >20-kb cutoff, using a High Pass Plus cassette (BPLUS10) run on a Blue Pippin (Sage Scientific). Using 1.6 µg of size-selected DNA, four sequencing libraries were independently generated for each sample using the SQK-LSK109 sequencing kit (Oxford Nanopore Technologies), according to the “Genomic DNA by Ligation Nanopore” protocol. For each species, three PromethION flow cells (vR9.4.1; ONT) were loaded with library material. Run performance was monitored, and once the number of sequencing pores dropped below 10% of the starting number, the run was stopped and a nuclease flush was performed using the NFL\_9076\_v109\_revA Nuclease Flush protocol from Oxford Nanopore. Additional library material was loaded onto flow-cells (by species) and sequencing was initiated. In total, three flow cells were used to sequence the Dwarf sample (with three reloads among them) and three flow cells for the Normal sample (with three reloads). Raw nanopore reads were base-called using GUPPY (version 3.0.5. flip-flop HAC model). Data metrics before quality filtering were 72.1 Gb (N50 = 27.1 kb) for the Dwarf sample and 80 Gb (N50 = 27.9 kb) for the Normal sample.

#### 2.1.2 | Short-read sequencing

For population-level analysis, we sampled and sequenced 32 *C. clupeaformis* including eight Normal and eight Dwarf from Cliff Lake, Maine (46.3991, -69.2491), and seven Normal and nine Dwarf from Indian Lake, Maine (46.2574, -69.2987) with Illumina short reads. Fish were caught live with gillnets, killed and immediately dissected to obtain fresh tissue samples. Samples were stored in RNAlater and DNA was extracted using a modified version of a salt extraction protocol (Aljanabi & Martinez, 1997). Shotgun libraries were prepared and sequenced aiming for 5× coverage with 150-bp paired-end reads on a HiSeq4000 instrument at the McGill Genome Québec Innovation center (Montréal).

Paired short reads were trimmed and filtered for quality with FASTP version 0.20.0 using default parameters (Chen et al., 2018), aligned



**FIGURE 1** Overview of the study design. Sampling and sequencing design, which included 32 wild samples of *Coregonus clupeaformis* Normal sp. and Dwarf sp. from Cliff Lake and Indian Lake in Maine (USA), sequenced by Illumina short reads, as well as two samples from Cliff Lake (one Normal and one Dwarf), sequenced by Nanopore long reads to assemble genomes. The insets represent the geographical locations of the two lakes sampled for this study and a schematic phylogeny of the different populations based on relationships inferred in Rougeux et al. (2017), the arrows representing ongoing gene flow (one migrant per generation in Cliff Lake, 1–30 migrants per generation in Indian Lake). The flowchart displays the main features of the pipeline of analysis performed to detect and genotype structural variants (SVs) with different data sets

to the reference genome of the Normal *C. clupeaformis* (see below) with BWA-MEM (Li & Durbin, 2009), and filtered to keep mapping quality over 10 with SAMTOOLS version 1.8 (Li et al., 2009). Duplicate reads were removed with MarkDuplicates (PICARDTOOLS version 1.119). We realigned around indels with GATK IndelRealigner (McKenna et al., 2010) and soft clipped overlapping read ends using clipOverlap in BAMUTIL version 1.0.14 (Breese & Liu, 2013). The pipeline is available at [https://github.com/enormandeu/wgs\\_sample\\_preparation](https://github.com/enormandeu/wgs_sample_preparation).

## 2.2 | Assembly and annotation of two reference genomes for *C. clupeaformis*

### 2.2.1 | De novo assembly and polishing

Long reads were filtered for a minimum length of 4000 bp and minimum average quality PHRED score of 7. This resulted in a total of

62.9 Gb (N50 = 28.5 kb, N90 = 16.3 kb) for the Normal and 60.8 Gb (N50 = 27.4 kb, N90 = 15.0 kb) for the Dwarf, and hence a coverage of ~23× considering a genome size around 2.7 Gb. For the Normal assembly, filtered long reads were independently assembled using FLYE (Kolmogorov et al., 2019) (version 2.5, default parameters) three times using overlap sizes of 8, 10 and 15 kb (Table S1). The three resulting assemblies were merged into a final assembly with QUICKMERGE (Chakraborty et al., 2016) (version 0.3, options: -hco 5.0 -c 1.5 -l 2000000 -ml 10000). For the Dwarf assembly, filtered long reads were assembled using FLYE (version 2.5, default parameters) using overlap sizes of 8, 10 and 12 kb and the assembly with the best N50 was chosen (10 kb). The final assemblies were first polished with their respective long reads using MARGINPOLISH (version 1.2.0 <https://github.com/UCSC-nanopore-cgl/MarginPolish>) for the Normal and PEPPERPOLISH (default settings, model: pepper\_r941\_guppy305\_human.pk1), a successor program with similar performance, for the Dwarf. In a second step, each assembly was polished with short

reads using PILON (Walker et al., 2014) requiring a minimal coverage of 3× to polish (version 1.23, --mindepth 3). BUSCO (Benchmarking Universal Single-Copy Orthologs) scores were computed to assess gene space completeness by looking for the presence or absence of highly conserved genes (BUSCO version 3.0.2, reference database: actinopterygii\_odb9 -sp zebrafish). BUSCO scores for the FLYE-polished assemblies were C: 94.4% [S: 50.9%, D: 43.5%], F: 1.7%, M: 3.9%, n: 4,584 for the Normal and C: 94.6% [S: 59.1%, D: 35.5%], F: 0.9%, M: 4.5%, n: 4,584 for the Dwarf. In other words, out of 4584 searched BUSCO gene groups about 94% were detected as singletons (S) or duplicates (D), a small fraction were missing (M) or fragmented (F).

## 2.2.2 | Scaffolding into chromosomes with a genetic map

To anchor the contigs into chromosomes, we rebuilt a linkage map from previously published data (Gagnaire et al., 2013a; Rogers et al., 2007). The map is based on a backcross family whose mother is a Dwarf × Normal hybrid and father is a pure Dwarf (all details in Rogers et al., 2007). The 100 full-sibs and their two parents were sequenced with reduced-representation sequencing in a previous study (Gagnaire, Normandeau, et al., 2013). Raw reads were aligned on the new contig-level assembly of the Normal genome with BWA-MEM using the default parameters (Li & Durbin, 2009). Genotype likelihoods were obtained with SAMTOOLS mpileup (Li et al., 2009) following the pipeline and parameters provided in LEP-MAP3 documentation (Rastas, 2017). Only positions with at least 3× coverage were kept. A linkage map was built using LEP-MAP3 (Rastas, 2017) following a pipeline available at [https://github.com/claiermerot/lepmap3\\_pipeline](https://github.com/claiermerot/lepmap3_pipeline). With the *Filtering* module, markers with more than 50% of missing data, that were noninformative, and with extreme segregation distortion ( $\chi^2$  test,  $p < 10^{-12}$ ) were excluded. Markers were assigned to linkage groups (LGs) using the *SeparateChromosomes* module with increasing values of the logarithm of the odds (LOD) from 8 to 11 and a minimum size of 20 markers. Markers unassigned to LGs, or released from LG correction, were subsequently joined to LGs using the module *JoinSingle* with decreasing values of LOD until LOD = 3 and a minimum LOD difference of 1. This procedure assigned 5188 markers into 40 LGs. Within each LG, markers were ordered with 10 iterations of the *OrderMarker* module. The marker order from the run with the best likelihood was retained and refined 10 times with the *evaluateOrder* flag with five iterations each. To account for the lower recombination rate in male salmonids compared to females, we adjusted the parameter of recombination rates accordingly (recombination1 = 0.0005; recombination2 = 0.0025). Exploration for more stringent filtering for missing data, different values of LOD or by keeping only female-informative markers resulted in very consistent and collinear maps but with fewer markers, whose density is critical to accurately scaffold the genome.

Since *C. clupeaformis* sp. Normal and sp. Dwarf have the same number of chromosomes (Dion-Côté et al., 2015) and the genetic

map was built from a backcross family, we used the same map to anchor both the Normal and the Dwarf genome assemblies. Scaffolds were assembled into chromosomes using *Chromonomer* (Catchen et al., 2020), which anchors and orients scaffolds based on the order of markers in the linkage map. Default parameters were used. In both assemblies, chromosomes were renamed to match homologous chromosomes in the reference genome of the European sister species, *C. lavaretus* “Balchen” (De-Kayne et al., 2020), as detailed in Table S2. For all subsequent analyses, the Normal whitefish genome was chosen as the reference because of its higher contiguity (N50 = 6.1 Mb for the Normal, N50 = 2.2 Mb for the Dwarf) and because a higher fraction of the genome could be anchored into chromosomes in the Normal (83%) than the Dwarf (73%). It is also worth noting that, by using the same linkage map to anchor chromosomes in both the Dwarf and Normal genome, the current assemblies do not allow us to investigate large-scale chromosomal rearrangements.

## 2.2.3 | Annotation for genes and TEs

Gene content annotation of both genomes was made with the NCBI Prokaryotic Genome Annotation Pipeline using the following transcriptome sources available on NCBI: Dion-Côté: PRJNA237376; Rougeux: 72 liver RNA samples from 2018, NCBI: PRJNA448004; Carruthers: SRR6321817, SRR6321818, SRR6321819, SRR6321820, SRR6321821, SRR6321822, SRR6321823, SRR6321824; Pasquier: SRP058861 lake whitefish, SRP045143 European whitefish.

We used REPEATMODELER2 (Flynn et al., 2020) to build a library of TEs from the *C. clupeaformis* sp. Normal assembly. We had to slightly modify the REPEATMODELER LTR pipeline because LTRHARVEST failed for an unknown reason. We instead substituted it with an equivalent program, LTRFINDER-PARALLEL (Ou & Jiang, 2019), to identify long terminal repeats (LTRs) in the genome. We combined the LTR-specific library with the general repeat library as done in canonical REPEATMODELER2. After obtaining the TE library, we relabelled the fasta headers of sequences that were identified in the LTR pipeline but were assigned an “Unknown” classification due to lack of homology to database sequences, to broadly classify them as LTR elements.

We then used REPEATMASKER to annotate the locations of each repeat family in both the Normal and the Dwarf genomes. We used parseRM.pl (<https://github.com/4ureliek/Parsing-RepeatMasker-Outputs/blob/master/parseRM.pl>) to summarize the genomic abundance of each TE subclass (LTR, LINE, SINE, DNA-TIR, Helitron), correcting for overlapping masking which sometimes occurs with REPEATMASKER. We also used parseRM.pl to produce a landscape plot of the genome composition, where the TE-subclass composition is shown in 1% divergence windows (compared to each TE copy's respective consensus sequence), where low-divergence sequences suggest more recent insertions and higher divergence sequences suggest older insertions.

## 2.2.4 | Synteny, map, chromosomes and genome analysis

To analyse synteny with related species, we first compared the linkage map to the previously published maps of *C. clupearformis* (Gagnaire, Normandeau, et al., 2013), *C. lavaretus* "Albock" (De-Kayne & Feulner, 2018) and *C. artedii* (Blumstein et al., 2020) using MAPCOMP (Sutherland et al., 2016), a program designed to compare syntenic relationships among markers between linkage maps of any related species using an intermediate genome, here our reference genome. Correspondence between chromosomes and linkage groups across maps of different *Coregonus* sp. is provided in Table S2 and Figures S1–S3.

Next, we aligned the repeat-masked *C. clupearformis* sp. Normal and sp. Dwarf genomes to the European whitefish reference, *C. lavaretus* sp. Balchen (De-Kayne et al., 2020), and to each other, with NUCMER (-l 100 -c 500; Marçais et al., 2018) and used SYMAP version 4.2 (Soderlund et al., 2011) to extract syntenic blocks along the genome. Syntenic blocks were visualized in R using the package Circlize (Gu et al., 2014).

To investigate chromosome types (acrocentric, metacentric), we used phased information from the linkage map by applying a method developed by Limborg et al. (2016), which uses phased progeny genotypes to detect individual recombination events. The cumulative number of recombination events between the first marker and increasingly distant markers was computed from both extremities of each chromosome and this recombination frequency (RFm) is expected to reach a plateau over a chromosome arm (see Limborg et al. (2016) for details and Figure S4).

As salmonids have experienced an ancestral whole-genome duplication, most of the chromosomes are expected to be homologous to another one, and some pairs still recombine to a certain extent, resulting in pseudotetrasomal regions or chromosomes (Glasauer & Neuhauss, 2014; Lien et al., 2016; Sutherland et al., 2016). To investigate this homology, we explored self-synteny by aligning the repeat-masked *C. clupearformis* sp. Normal genome on itself with NUCMER (-maxmatch -l 100 -c 500; Marçais et al., 2018) and extracted syntenic blocks with SYMAP version 4.2 (Soderlund et al., 2011). The degree of sequence similarity within each of the syntenic blocks was calculated after a subsequent alignment with LASTZ (Harris, 2007), following the procedure described in Lien et al. (2016). To assign *C. clupearformis* chromosomes to ancestral chromosomes following the nomenclature proposed by Sutherland et al. (2016) based on northern pike (*Esox Lucius*) linkage groups, we aligned the repeat-masked Normal genome to the northern pike reference genome with MINIMAP2 (Li, 2018) and visualized alignment using D-GENIES (Cabanettes & Klopp, 2018).

We further explored whether the assembly included duplicated or collapsed regions by quantifying variation of coverage along the genome. Total depth of aligned short reads across the 32 samples was calculated using ANGSD (Korneliussen et al., 2014) at each position with the option -doDepth -dumpCounts, and averaged by sliding windows of 100 kb. The coordinates of putatively collapsed regions, defined as regions having a depth greater than the average depth plus twice the standard deviation and showing no homology with another chromosome, are provided in Table S3.

## 2.3 | Detection and characterisation of SVs

We performed SV detection based on three data sets: (i) the genome assemblies of the Normal and the Dwarf; (ii) the long reads of the two samples (Normal and Dwarf) used to build the genome assemblies; and (iii) the short reads of 32 samples (Normal and Dwarf). SV detection with the three data sets shared consistent features. First, all SVs were defined relative to the reference genome of *C. clupearformis* sp Normal. Second, to enhance SV detection, SVs were detected by three independent software packages, but to better limit the amount of false positives, we kept only SVs detected by at least two out of three SV callers in each data set as proposed previously (De Coster et al., 2019; Weissensteiner et al., 2020). Third, we focused on variants over 50 bp (Ho et al., 2019) and restricted our analysis to insertions (INS), deletions (DEL), duplications (DUP) and inversions (INV) to simplify the use of multiple tools, including merging software and genome-graph representations. Fourth, to avoid artefacts due to genome misassemblies, we filtered out SVs which overlapped a scaffold junction (characterized by a gap of 10 Ns).

### 2.3.1 | SV detection based on the comparison of de novo assemblies

SVs between the Normal and the Dwarf haploid assemblies were identified using three independent approaches detailed below. All methods included an alignment step of the query assembly (*C. clupearformis* sp. Dwarf) on the reference assembly (*C. clupearformis* sp. Normal). To avoid artefacts due to scaffolding with a map, we chose to use the contig-level assembly for the Dwarf genome.

- (i) We built a genome-graph with the two assemblies using MINIGRAPH (Li et al., 2020) with the -xggs options and retrieved SVs in bed format with GFATOOLS-BUBBLE. The graph with variants was further reformatted into a vcf with full sequence information using VG SUITE (Hickey et al., 2020).
- (ii) We aligned the assemblies with MINIMAP2 (Li, 2018) and parameters -a -x asm5 --cs -r2k, and extracted SVs with SVIM-ASM (Heller & Vingron, 2020) and the following parameters: --haploid --min\_sv\_size 50 --max\_sv\_size 200000 --tandem\_duplications\_as\_insertions --interspersed\_duplications\_as\_insertions.
- (iii) We ordered the scaffolds of the Dwarf assembly according to the Normal reference using RAGTAG (Alonge et al., 2019) and aligned the assemblies with MINIMAP2 (Li, 2018) and parameters "-ax asm5" and ran SVRI (Goel et al., 2019) with standard parameters.

After filtering, the three VCFs were joined using JASMINE (Kirsche et al., 2021) using the following parameters: "--ignore\_strand --mutual\_distance --max\_dist\_linear=0.5 --min\_dist=100," and we kept SVs detected by at least two approaches. All scripts are available at [https://github.com/clairemérot/assembly\\_SV](https://github.com/clairemérot/assembly_SV).

### 2.3.2 | SV detection based on long reads

We mapped long reads from both the Dwarf and the Normal samples to the Normal reference using `WINNOMAP2` version 2.0 with the “--MD” flag to better resolve repetitive regions of the genome (Jain et al., 2020). SAM files were sorted and converted into BAM files with `SAMTOOLS` version 1.3.1 (Li et al., 2009). SV detection was performed with three long-read-specific SV calling programs: `SNIFLES` version 1.0.12 (Sedlazeck et al., 2018) (-l 50 -s 7 -n -1), `SVIM` version 1.2.0 (Heller & Vingron, 2019) (--insertion\_sequences) and `NANOVAR` version 1.3.9 (Tham et al., 2020) with default settings. VCF files were filtered using custom R scripts to remove excess information and read names were added to preserve insertion sequences in the final VCF. We kept SVs detected by at least two callers after merging with `JASMINE` version 1.1.0 (Kirsche et al., 2021) including refinement of insertion sequences with `IRIS` “max\_dist\_linear=0.1 min\_dist=50 --default\_genotype --mutual\_distance min\_support=2 --output\_genotypes --normalize\_type --run\_iris iris\_args=--keep\_long\_variants.” All scripts are available at [https://github.com/kristinastenlokk/long\\_read\\_SV](https://github.com/kristinastenlokk/long_read_SV).

### 2.3.3 | SV detection based on short reads

SVs among the 32 samples sequenced with short reads were identified using three independent approaches: (i) `MANTA` (Chen et al., 2016), (ii) the `SMOOVE` pipeline (<https://github.com/brentp/smoove>) which is based on `LUMPY` (Layer et al., 2014) and (iii) `DELLY` (Rausch et al., 2012). All of the approaches rely on the filtered bam files resulting from the alignment of the short reads to the Normal reference (as described above). All SV callers were run with default parameters except for `SMOOVE` which was run by subsets of chromosomes, and `DELLY` by subsets of individuals. VCF outputs were formatted and filtered with custom scripts called “delly\_filter,” “manta\_filter,” and “smoove\_filter” to include full sequence information. The three VCFs were joined using `JASMINE` (Kirsche et al., 2021) and the following parameters: “--ignore\_strand --mutual\_distance --max\_dist\_linear=0.5 --min\_dist=50 --max\_dist=5000 --allow\_intrasample,” and we kept SVs detected by at least two approaches. All scripts are available at [https://github.com/clairmerot/SR\\_SV](https://github.com/clairmerot/SR_SV).

### 2.3.4 | Analysis and annotation of SVs

SVs detected by the three kinds of data sets (assembly comparison, long reads, short reads) were joined using `JASMINE` (Kirsche et al., 2021) and the following parameters: “--ignore\_strand --mutual\_distance --max\_dist\_linear=0.5 --min\_dist=100 --min\_overlap 0.5.” This merging tool represents the set of all SVs as a network, and uses a modified minimum spanning forest algorithm to determine the best way of merging the variants based on position information (chromosome, start, end, length) and their type (DEL, INS, DUP, INV),

requiring a minimum overlap between SVs and a maximum distance between breakpoints. We explored different parameter values without noticing major differences in the final merging, and hence the final choice of intermediate parameters (50% of the length). We reported the overlap of SVs detected in more than one data set according to its type and its size. The sequences included in SVs (e.g., the reference sequence in the case of a deletion, or the alternative sequence in the case of an insertion) were annotated for TEs using `REPEATMASKER` and the TE library of the Normal *C. clupeaformis* (see above). We explored the length of SV sequences covered by TE or simple repeats quantitatively (Tables S4 and S5) and also categorized them as associated with TE or other kinds of repeats if more than 50% of the SV sequence was covered by a given TE family or other kind of repeats.

## 2.4 | Analysis of single-nucleotide and structural polymorphisms

### 2.4.1 | SNP calling and genotyping

To detect SNPs and genotype them, we analysed the short reads of the 32 samples, in bam format, with the program `ANGSD` version 0.931 (Korneliussen et al., 2014), which accounts for genotype uncertainty and is appropriate for medium-coverage whole genome sequencing (Lou et al., 2020). Input reads were filtered to remove low-quality reads and to keep mapping quality above 30 and base quality above 20. Genotype likelihoods were estimated with the `GATK` method (-GL 2). The major allele was the most frequent allele (-doMajorMinor 1). We filtered to keep positions covered by at least one read in at least 50% of individuals, with a total coverage below 800 (25 times the number of individuals) to avoid including repeated regions in the analysis. From this list of variant and in-variant positions, we selected SNPs outside SVs and with a minor allele frequency (MAF) above 5%. We subsequently used this SNP list with their respective major and minor alleles for most analyses, including principal components analysis (PCA),  $F_{ST}$  and allelic frequency difference (AFD).

### 2.4.2 | SV genotyping

To genotype the identified SVs in the 32 samples, we used a genome-graph approach with the `vg` suite of tools (Garrison et al., 2018; Hickey et al., 2020). Briefly, the full catalogue of SVs discovered (through assembly comparison and long- and short-read SV calling) was combined with the reference genome to build a variant-aware graph using the module `vg autoindex -giraffe`. Short reads from the 32 samples were then aligned to the graph with the module `vg giraffe` (Sirén et al., 2020). For each SV represented in the graph through a reference and an alternative path, a genotype likelihood was calculated with the module `vg call`. We then combined the VCFs of SV genotype likelihoods across the 32 samples. For population-level



analysis, mirroring the filters applied for SNPs, we retained SVs covered by at least one read in at least 50% of samples, and with an alternative allele frequency between 5% and 95%. The pipeline used is available at [https://github.com/claïmerot/genotyping\\_SV](https://github.com/claïmerot/genotyping_SV). Subsequent analytical steps were performed in ANGSD, using the VCF of SV genotype likelihoods as input, to perform population-level analysis within a probabilistic framework to account for the uncertainty linked to medium coverage.

### 2.4.3 | Genetic differentiation according to lake and species

An individual covariance matrix was extracted from the genotype likelihoods of SNPs and SVs in beagle format using PCANGSD (Meisner & Albrechtsen, 2018). The matrix was decomposed into PCs with R using a scaling 2 transformation which adds an eigenvalue correction (Legendre & Legendre, 2012). Pairwise  $F_{ST}$  differentiation between all populations was estimated based on the allele frequency spectrum per population (-doSaf) and using the realSFS function in ANGSD. Minor allelic frequencies per population (MAF) were estimated based on genotype likelihoods using the function doMaf in ANGSD. We then computed AFD between sympatric species in each lake for each variant as  $AFD = MAF(Dwarf) - MAF(Normal)$ . AFD is a polarized difference of frequency that varies between -1 and 1, meaning that when we compared AFD between lakes they can be either with the same sign (the same allele has a higher frequency in the same species in both lakes) or opposite sign (the allele more frequent in the Dwarf in one lake is more frequent in the Normal in the other lake). For  $F_{ST}$  and AFD estimates, positions were restricted to the polymorphic SNPs/SVs (>5% MAF) previously assigned as major or minor allele (options -sites and -doMajorMinor 3), and which were covered in at least 50% of the samples in each population. Given the high density of SNPs,  $F_{ST}$  and mean absolute AFD were also calculated by windows of 100 kb for visualization and correlation tests. The most differentiated variants between species were defined in each lake as those within the upper 95% quantile for  $F_{ST}$  and either below the 2.5% or above the 97.5% quantile for AFD. By chance only, we would expect that 0.25% of variants are in the upper  $F_{ST}$  quantile in both lakes ( $5\% \times 5\%$ ), 0.125% of variants are in AFD outliers in both lakes with the same sign ( $2.5\% \times 2.5\% \times 2$ ), and 0.125% of variants are in AFD outliers in both lakes with opposite sign. We used Fisher's exact test to determine whether the number of outlier variants overlapping between lakes exceeded this expectation.

Using BEDTOOLS, we extracted the list of genes overlapping with the most differentiated SNPs/SVs. We then tested for the presence of overrepresented GO terms using GOATOOLS (version 0.6.1,  $p_{\text{val}} = .05$ ) and filtered the outputs of GOATOOLS to keep only GO terms for biological processes with an FDR value of  $\leq 0.1$ .

Using our annotation of TEs and repeated sequences on SVs, we tested whether some families of TEs were over-represented in the subset of outlier SVs relative to the whole pool of SVs studied at the population level using a Fisher exact test.

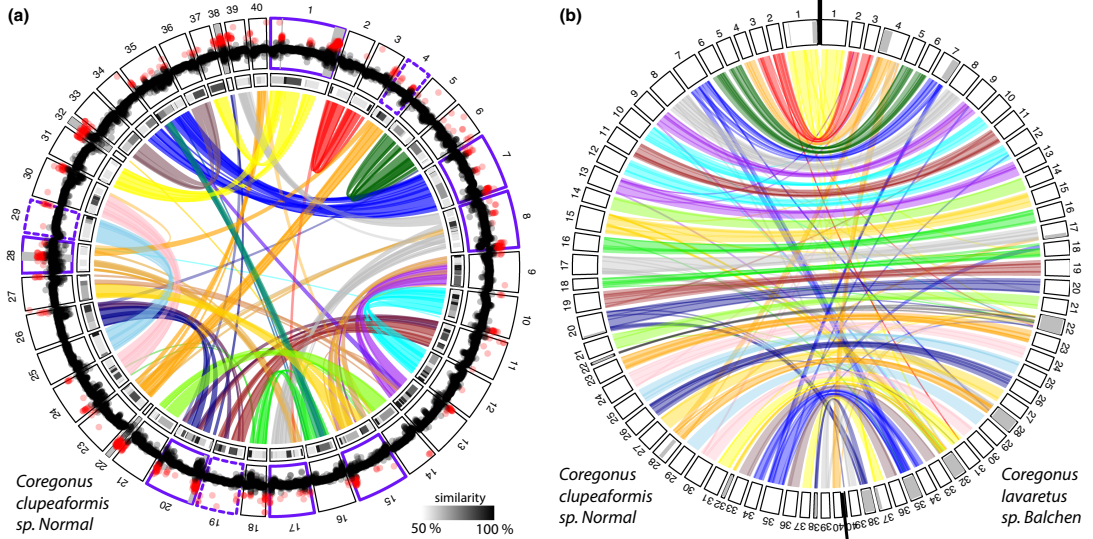
Finally, several quantitative trait loci (QTLs) for behavioural, morphological and life-history traits differentiating Normal and Dwarf previously identified in Gagnaire, Normandeau, et al. (2013) and Rogers et al. (2007) were positioned on the Normal reference genome. We compared the positions of those QTLs relative to the most differentiated regions and extracted the list of genes hit by an outlier variant and falling within a 1-Mb window around the QTL.

## 3 | RESULTS

### 3.1 | High-quality reference assembly for *Coregonus clupeaformis* sp. Normal

Using long-read sequencing, we built the first reference genome assembly for *C. clupeaformis* sp. Normal (ASM1839867v1). The *de novo* assembly showed good contiguity with an N50 of 6.1 Mb and a L50 of 101 contigs. A linkage map allowed us to anchor and orient 83% of the genome into 40 linkage groups, the expected number of chromosomes for *C. clupeaformis* (Dion-Côté et al., 2015; Phillips et al., 1996), although some of the linkage groups, chromosome 22 in particular, may only represent a fraction of a chromosome. Studying recombination along those linkage groups, we identified seven metacentric chromosomes, three putatively metacentric (or submetacentric) chromosomes and 30 acrocentric chromosomes (Figure S4; Figure 2a). The final assembly included 40 putative chromosomes and 6,427 unanchored scaffolds with an N50 of 57 Mb for a total genome size of 2.68 Gb (Table 1). This reference genome had a high level of completeness, with 94% of universal single-copy orthologous genes in a BUSCO analysis based on the actinopterygii database. A relatively high percentage of duplicated busco groups (44%) was observed, which is probably a consequence of the salmonid-specific whole genome duplication (Allendorf & Thorgaard, 1984; Smith et al., 2021).

The genome of *C. clupeaformis* sp. Normal was composed of 60.5% TEs (Figure S5, Table S4). The greatest TE subclass representation in terms of total base pairs was DNA-TIR elements, taking up 24% of the genome. LINES and LTRs were approximately equally abundant at about 13% of the genome each. Elements that were unclassified took up 9% of the genome. SINES took up <1% of the genome, and rolling-circle/helitron elements were essentially absent. Our repeat identification pipeline identified 3490 distinct families. LTR elements were the most diverse with 1521 families identified, almost half the total number of families. Comparatively, 373 families were identified as DNA-TIR elements and 250 as LINES. The genome of *C. clupeaformis* sp. Normal is composed of TEs at a variety of stages of decomposition (Figure S6). A proxy for age of a given insertion is its sequence divergence from the consensus sequence, since the longer the insertions have been present, the more time there has been for accumulation of random mutations. The landscape plot shows that an equal amount (in terms of bp) of LINES, LTRs and DNA-TIRs are present in recent insertions (less than 1% diverged from the consensus sequence). DNA-TIR elements near



**FIGURE 2** Self-synteny and coverage in *Coregonus clupeaformis* sp. Normal, and synteny with *C. lavaretus* sp. Balchen. (a) Circular plot showing syntenic relationship between homoeologous chromosomes (inner track) and their level of sequence similarity (medium track) in the genome of *C. clupeaformis* sp. Normal. The outer track displays mean coverage by windows of 100 kb in the short-read alignments. Points coloured in red show coverage higher than 1.5 times the average coverage (3.7 $\times$ ). Chromosomes surrounded by a purple outline are metacentric chromosomes, with dashed lines for putatively metacentric chromosomes. (b) Circular plot showing syntenic relationship between *C. clupeaformis* sp. Normal and *C. lavaretus* sp. Balchen. On both plots, chromosomes are coloured according to the ancestral origin (based on the PK nomenclature proposed in Sutherland et al., 2016). Regions coloured in grey represent collapsed duplicated regions in genome assemblies

**TABLE 1** Statistics of the genome assemblies of *Coregonus clupeaformis* sp. Normal and sp. Dwarf

Species	<i>Coregonus clupeaformis</i> sp. Normal	<i>Coregonus clupeaformis</i> sp. Dwarf
Genome size	2.68 Gb	2.76 Gb
N50 (contig level)	6.1 Mb	2.2 Mb
L50 (contig level)	101 contigs	274 contigs
Fraction anchored in chromosomes	83%	73%
N50 (final assembly)	57 Mb	52 Mb
busco score (Actinopterygii)	C: 94.4% [S: 50.9%, D: 43.5%], F: 1.7%, M: 3.9%, n: 4584	C: 94.6% [S: 59.1%, D: 35.5%], F: 0.9%, M: 4.5%, n: 4584
Fraction of TEs	60.5%	62.4%

the 5% divergence level are the most abundant, indicating an older burst of activity.

The genome of *C. clupeaformis* sp. Normal showed high synteny with the closely related European Alpine whitefish, *C. lavaretus* "Balchen" (Figure 2a), allowing the identification of 39 homologous chromosomes which were named accordingly. Chromosome 40 of *C. lavaretus* sp. Balchen was small and had no homologous chromosome in the genome of *C. clupeaformis* sp. Normal. Chromosome 40 of *C. clupeaformis* sp. Normal aligned with a fraction of chromosome 4 in the *C. lavaretus* sp. Balchen assembly and may or may not be one arm

of the putatively metacentric chromosome 4. Some chromosomes (Chr7, Chr8, Chr15, Chr17, Chr20, Chr28, Chr35) included syntenic blocks matching two chromosomes in the related species. Some of those blocks probably correspond to duplicated regions collapsed in one of the assemblies, as they also exhibit higher than average coverage (Figure 2). Those blocks may also belong to pseudotetrasomal chromosomes, which are homeologous chromosomes resulting from ancient whole-genome duplication and that still recombine to a certain extent (Allendorf et al., 2015; Blumstein et al., 2020; Lien et al., 2016; Waples et al., 2016).

The identification of ancestral chromosomes by alignment to other linkage maps (Figures S1–S3) and to the northern pike genome (Figure S7), as well as self-synteny (Figure 2a), allowed us to further identify the pairs of homeologous chromosomes. A few regions (Chr22, Chr 32, the end of Chr1) show no matching region in the genome of *C. clupearformis* sp. Normal but high coverage, suggesting that the assembly may have locally collapsed two highly similar regions (Figure 2a; Table S3). Self-synteny assessment also supports fusion events between ancestral chromosomes that were previously reported in the three *Coregonus* species, *C. lavaretus*, *C. artedii* and *C. clupearformis* (Blumstein et al., 2020; Sutherland et al., 2016) such as PK05–PK06 (Chr01), PK10–PK24 (chr8), PK11–PK21 (Chr7), PK01–PK14 (Chr15), PK16–PK23 (Chr4) and PK8–PK9 (Chr20), as well as putative complex rearrangements between PK10–PK20–PK23 (Chr17, Chr28, Chr4). Those eight chromosomes also correspond to those identified as metacentric in our study and in the *Cisco artedii* (Blumstein et al., 2020).

### 3.2 | Discovery of SVs using a combination of sequencing tools

To identify SVs between Normal and Dwarf species, we built a *de novo* assembly for *C. clupearformis* sp. Dwarf (ASM2061545v1) based on long-read sequences. This assembly shows high contiguity with an N50 of 2.2 Mb and L50 of 274 contigs, of which 73% could be placed into chromosomes using the linkage map. The final Dwarf assembly included 40 chromosomes and 7,294 unanchored scaffolds with an N50 of 52 Mb for a total genome size of 2.76 Gb. The Dwarf genome also showed high synteny with *C. lavaretus* sp. Balchen (Figure S8). Like the Normal genome, the genome of *C. clupearformis* sp. Dwarf was composed of about 61% TEs at various ages, with similar repartition between different class and families (Figures S5 and S6, Table S6). The Dwarf genome also contains a high fraction (95%) of universal single-copy orthologous genes (*actinopterygii*), among which 36% were duplicated. This fraction is nevertheless smaller than in the Normal genome (44%), which may possibly reflect more collapsed duplicated regions in the Dwarf.

Comparison of the Dwarf assembly to the Normal reference revealed 244,717 SVs, of which 89,909 were detected by at least two tools and were kept as “high-confidence SVs.” Approximately half of the SVs were deletions and half were insertions (Figure 3a). Duplications were counted as insertions, and only a limited number of inversions were detected (2815, out of which only 77 were found by two tools).

Since a comparison of haploid assemblies is only able to detect SVs in the Dwarf relative to the Normal, and may be sensitive to assembly errors, we next investigated SV polymorphisms based on long reads. This revealed a higher number of high-confidence SVs with a total of 194,861 SVs detected by at least two tools. Those included SVs putatively heterozygous in the Normal and the Dwarf genomes and resulted in a high number of novel deletions and insertions.

Only two samples (one Dwarf and one Normal) were sequenced with long reads; hence we hoped to cover a wider range of population structural polymorphism by using short reads on 32 individuals (15 Normal and 17 Dwarf) to detect SVs. This method nevertheless appeared less powerful than SV detection based on long reads as 84,673 SVs were detected, with only 28,579 detected by at least two tools. This is possibly due to the smaller size of short reads and limited depth of sequencing in our data set (about 5x), which is suboptimal for SV calling. The large majority of SVs detected in this data set were deletions ( $n = 77,899$ ; 92%), followed by duplications ( $n = 5,927$ ; 7%), a few inversions ( $n = 24$ ; 0.02%) and insertions ( $n = 15$ ; 0.01%) (Figure 3a).

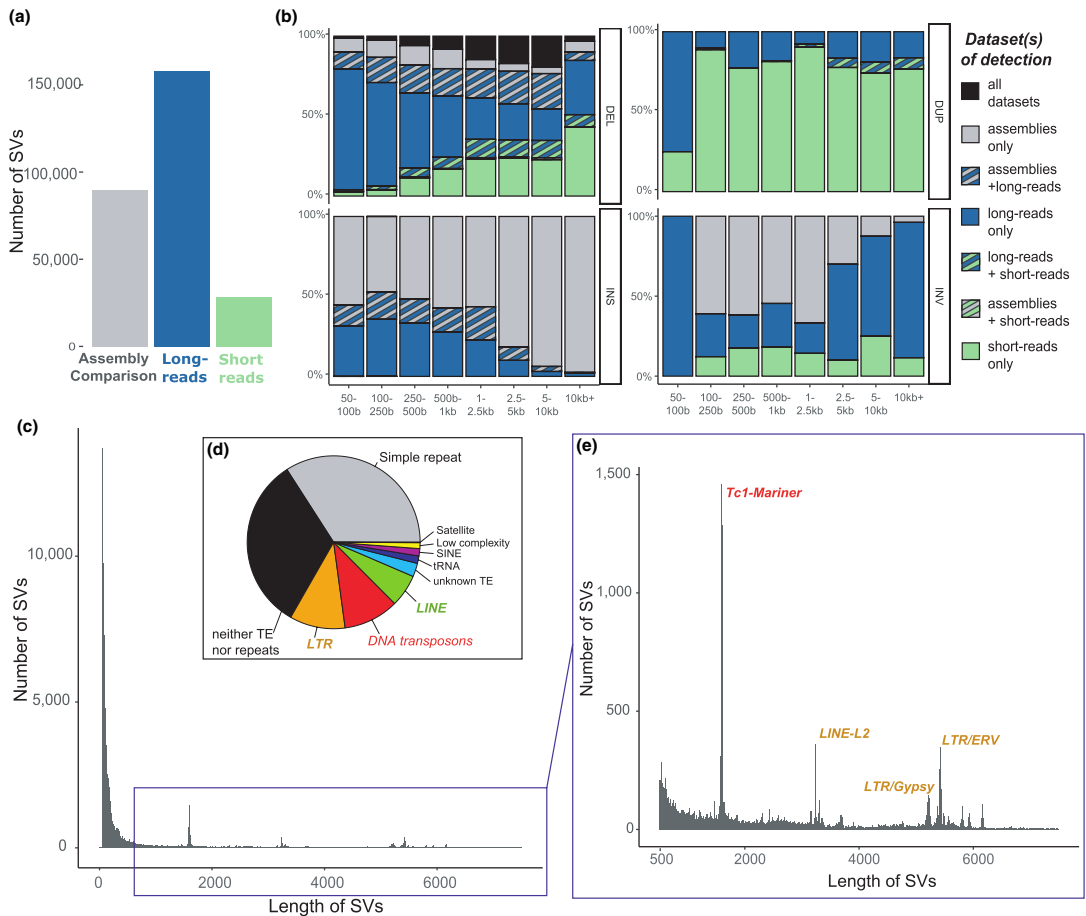
There was limited overlap between the different approaches with 7,525 SVs detected in the three data sets and 38,202 detected in two data sets out of a total of 222,927 SVs. This limited overlap, which varies depending on type and size, probably reflects the different sensitivity and detection power of the calling methods associated with each data set. Almost no overlap was observed for inversions and duplications, probably reflecting the difficulties in characterizing such SVs. For insertions, the overlap between long reads and assembly comparison approaches tended to decrease with size, possibly due to more approximate breakpoints, while for deletions it increased with size (Figure 3b).

The distribution of SV sizes was highly skewed towards smaller SVs below 500 bp (Figure 3c). We observed heterogeneous peaks in the SV size distribution corresponding to insertions or deletions of TEs (Figure 3e). The sequence of SVs around the 1600-bp peak matches with TC1-Mariner. SVs around 3700 bp correspond to Line-L2 indels while the peaks between 5000 and 6000 bp are different kinds of LTR (Gypsy, ERV1). Overall, TEs were important factors driving SVs in *C. clupearformis* as their sequences were composed of 73% of TEs (compared to 60% for the entire genome, Table S4). This enrichment was mostly due to retroelements (49% in SV sequences compared to 25% in the genome), mostly LTR and Gypsy (Table S5). This resulted in about a third of all SVs in the catalogue being associated with an insertion or deletion of a TE (Figure 3d). Satellite repeats and simple repeats (e.g., microsatellites) cover a smaller fraction of the SV sequences (5%, Table S4) but they were found in about a third of SVs. A third of SVs did not match any TE nor any repeated regions.

### 3.3 | Polymorphism and differentiation in *C. clupearformis* sp. Normal and sp. Dwarf

To assess genetic variation at the population level, we estimated genotype likelihoods for SNPs and SVs in the 32 samples sequenced with short reads. Filtering for genetic variants with allelic frequency >5% retained 12,886,292 SNPs and 103,857 SVs. Those “frequent” SVs cover a total of 66 Mb, representing polymorphism affecting approximately five times more nucleotides in the genome than SNPs.

Decomposing genetic variation with a PCA revealed a strong clustering of individuals by species and by lake. This was consistent whether considering SNPs or SVs, although SVs tended to show greater separation between the two species along the first



**FIGURE 3** Overview of SVs detected within and between *Coregonus clupeaformis* sp. Normal and sp. Dwarf. (a) Number of SVs detected in the three data sets by at least two tools. (b) Proportion of SVs detected in one or several data sets according to type and size. (c) Size distribution of SVs. (d) Proportion of SVs associated with different families of transposable elements and repeated elements. (e) Size distribution of SVs (zoomed on the range 500–7500 bp)

PC (Figure 4a,b). This suggests a higher level of shared interspecific variation between lakes for SVs than for SNPs.

$F_{ST}$  was moderate to high between lakes and between species, with values ranging from 0.052 up to 0.167 based on SVs and from 0.084 to 0.182 based on SNPs (Figure 4c). The Normal and Dwarf were more differentiated in Cliff Lake than in Indian Lake using both kinds of variants (Cliff Lake:  $F_{ST} = 0.175/0.167$ ; Indian Lake:  $F_{ST} = 0.098/0.062$ ) and such species differentiation was widespread along the genome (Figure 5). Within each lake, the landscape of interspecific  $F_{ST}$  displayed similarities between SNPs and SVs, and 100-kb window-based  $F_{ST}$  showed significant correlations when calculated on SNPs and on SVs (Cliff:  $R^2 = 0.71$ , Indian:  $R^2 = 0.63$ ). This suggests that there may be linked variants (e.g., small deletions and SNPs) and that the two kinds of mutations may affect each other, for instance if some SVs reduce recombination.

As the two lakes represent parallel situations of coexistence between the Normal and the Dwarf species of *C. clupeaformis* (Rougeux et al., 2017), we investigated whether genetic differentiation follows similar patterns. The most differentiated genetic variants, defined as the SNPs and SVs in the top 95%  $F_{ST}$  quantile within each lake, showed three times the expected number of shared variants across lakes, suggesting that areas of differentiation between species are conserved in parallel across lakes. When measuring species differentiation as a polarized difference in allelic frequencies (AFD statistic), this overlap was even stronger. There was a five-fold excess for AFD outliers in the same end of the distribution (positive in both lakes and negative in both lakes). In other words, the variants with high allelic frequency differences between species are more likely than expected by chance to display the same Normal allele and Dwarf allele in both lakes (Table 2).

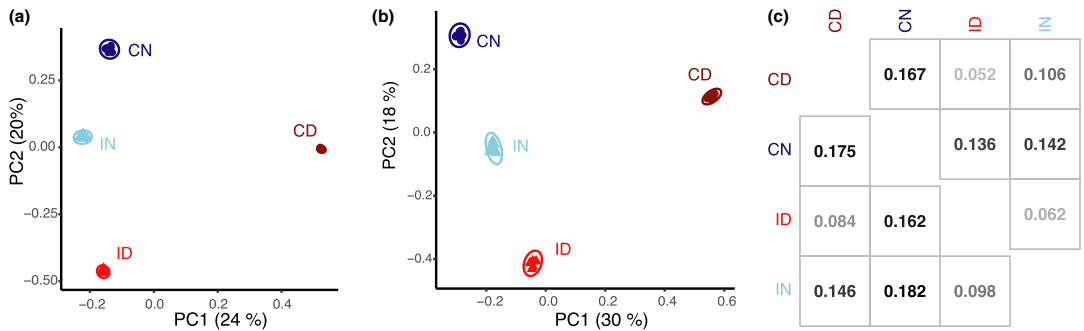


FIGURE 4 Genomic variation in *Coregonus clupeaformis* sp. Normal and sp. Dwarf. Principal component analysis (PCA) based on (a) SNPs and (b) SVs. Each point is an individual coloured by lake and by species. (c)  $F_{ST}$  between lakes and species based on SNPs (below diagonal) and SVs (above diagonal). CN, Normal from Cliff Lake; CD, Dwarf from Cliff Lake; IN, Normal from Indian Lake; ID, Dwarf from Indian Lake

Relative to all SVs, the most differentiated SVs, both within each lake and shared between lakes, were significantly enriched in TE-associated SVs. In other words, while SVs containing DNA transposons represent 15% of all SVs, they represent 27% of outlier SVs. In contrast, SVs associated with simple repeats were underrepresented in outliers of differentiation, while SVs without TEs or repeats showed no bias. This excess of TE-linked SVs in outliers was driven by all categories of TE: DNA transposons, LINES, SINES and LTRs. The most significantly enriched families in both lakes were the DNA transposons Tc1-mariner and hAT-Ac, and the retrotransposons LTR-Gypsy and LTR-ERV1, LINE-L1, LINE-L2 and LINE-RexBabar (Table 3; Table S7).

The most differentiated variants overlapped with thousands of genes. Out of a total of 34,913 genes with SNPs, 15,732 genes had at least one outlier SNP in Cliff Lake, 17,344 in Indian Lake and 4,678 in both lakes. Out of a total of 13,886 genes with SVs, 1396 genes had at least one outlier SV in Cliff Lake, 1,622 in Indian Lake and 242 in both lakes. Gene ontology analysis revealed significant enrichment in behaviour, morphogenesis, cell signalling, immunity and metabolism, among many other functions (Table S8). To narrow down putative candidate genes possibly involved in phenotypic differentiation, we focused on outliers overlapping with QTLs previously mapped with the linkage map (Gagnaire, Normandeau, et al., 2013; Rogers et al., 2007). A total of 27 QTLs for various traits differentiating Dwarf and Normal (growth rate, maturity, gill raker, etc.) could be positioned on the new reference genome, although some of them had a relatively wide interval (Table S9; Figure 5). They overlapped with 45,823 SNPs and 414 SVs that were identified as outliers of differentiation in both lakes. The list of genes belonging to a QTL and overlapping with at least one outlier is provided in Table S10.

## 4 | DISCUSSION

By combining long- and short-read sequencing on two species of the lake whitefish complex, *Coregonus clupeaformis* sp. Normal and *C. clupeaformis* sp. Dwarf, our study generated new genomic resources

and provided insights into the genomic architecture of recent speciation. First, we produced a reference genome assembly for both *C. clupeaformis* sp. Normal and *C. clupeaformis* sp. Dwarf, as well as an extensive catalogue of SVs. Second, studying SVs at the population level showed that SVs represent a large amount of variation within and between Normal and Dwarf sympatric species, less numerous but more extensive than SNPs in terms of the total number of nucleotides. Third, by comparing young species pairs in two lakes, we highlighted shared genetic differentiation and supported a predominant role of TEs in the divergence between the Normal and the Dwarf. Hereafter, we discuss how our results and methods contribute to a better understanding of the genomic architecture of speciation and the involvement of structural polymorphism.

Generating high-quality reference genomes for nonmodel species is becoming a requirement to understand the evolution of genomic variation during the speciation process (Nadeau & Kawakami, 2019; Ravinet et al., 2017). Here, using the genome of *C. clupeaformis* sp. Normal from North America as a reference facilitated the accurate detection of population-level variants, both SNPs and SVs, because the reference is from the same species, or a closely related species, and from the same geographical area. Moreover, contiguity and chromosome-level information allowed a finer understanding of the role played by recombination, large rearrangements and chromosome-level variability (fusion/fission, karyotypic polymorphism, etc.). In our study, the use of long reads proved incomparable to resolve the complexity of the genomes of *C. clupeaformis* sp. Normal and sp. Dwarf. Using Nanopore sequencing data, we have been able to reach a high contiguity, allowing us to search for SVs. Assembly comparison, as well as direct detection based on long reads, show that one Normal and one Dwarf individual differ by more than 100,000 high-confidence SVs. Given the stringency of our quality filters, and the lack of power to detect complex rearrangements or inversions, this number should be seen as a lower bound of the amount of SVs. In particular, most of the detected SVs remain in a range of small size (<1 kb) or relatively medium size. This catalogue of SVs can therefore be supplemented by including more individuals, longer sequences and additional genomes. Regardless,

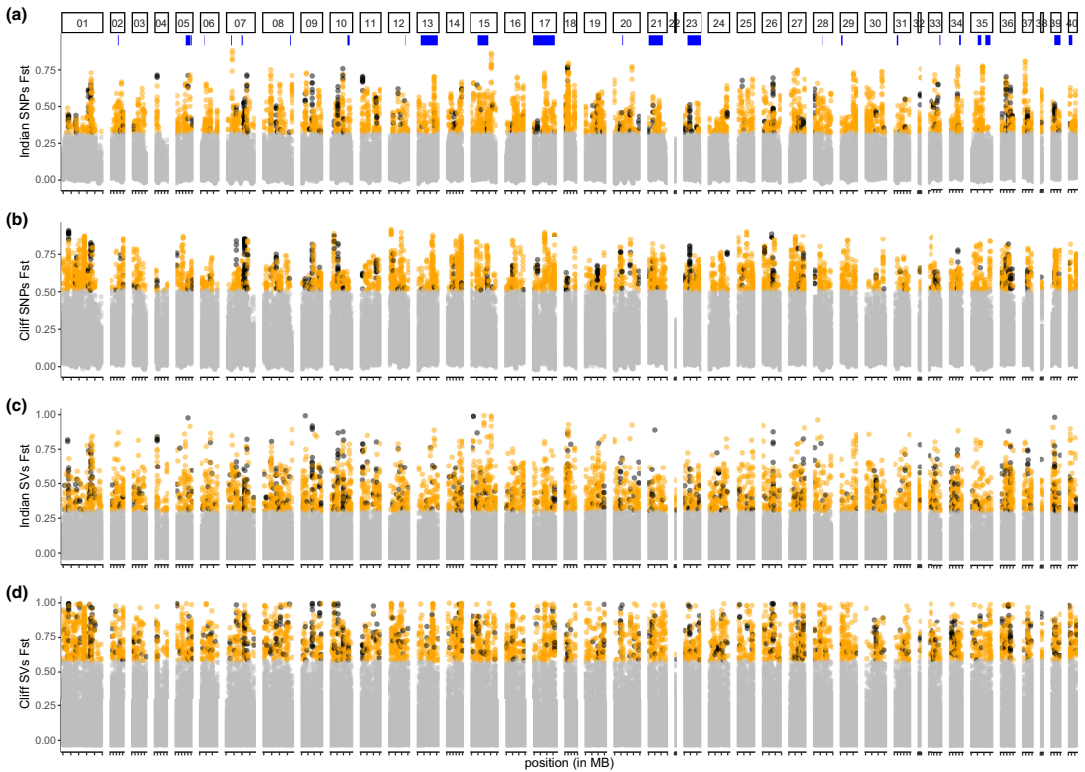


FIGURE 5 Genomic differentiation along the genome between *Coregonus clupeaformis* sp. Normal and sp. Dwarf.  $F_{ST}$  between Normal and Dwarf based on SNPs, by windows of 100 kb, in (a) Indian Lake and (b) Cliff Lake.  $F_{ST}$  between Normal and Dwarf based on SVs in (c) Indian Lake and (d) Cliff Lake. Windows and variants that exceed the 95% quantile in one lake are coloured orange. Shared polymorphisms between lakes (i.e., variants found as outliers in both lakes) are shown in black. Blue segments under chromosome numbers indicate the positions of QTLs associated with behavioural and morphological differences between Normal and Dwarf species, as identified in Gagnaire, Pavey, et al. (2013)

TABLE 2 Overlap across the two lakes in the most differentiated variants between species

Data set and method	Number of variants	Expected number of overlapping outliers	Observed number of overlapping outliers	Odds-ratio	$p$ -value (Fisher test)
SNP $F_{ST}$	11,389,952	28,475	94,572	3.3	<.001
SNP AFD same sign	11,389,952	14,237	80,474	5.7	<.001
SNP AFD opposite sign	11,389,952	14,237	17,947	1.3	<.001
SV $F_{ST}$	93,773	234	727	3.1	<.001
SV AFD same sign	93,773	117	618	5.3	<.001
SV AFD opposite sign	93,773	117	123	1.1	.75

Note:  $F_{ST}$  outliers were defined as the top 5% of the  $F_{ST}$  distribution. AFD is the allelic frequency difference between the *Coregonus clupeaformis* sp. Normal and *C. clupeaformis* sp. Dwarf (polarized as Dwarf-Normal). "Same sign" indicates that the outliers are in the same end of the AFD distribution in both lakes (either upper 97.5 quartile or lower 2.5 quartile), while "opposite sign" indicates that outliers are not in the same end of the AFD distribution in both lakes. In other words, outliers with opposite signs are variants in which the allele that is more frequent in the Dwarf in one lake is the allele that is more frequent in the Normal in the other lake.

the large number of high-confidence SVs identified in this study reinforces the importance of considering the possible role of SVs in evolutionary processes such as adaptation and speciation.

Regarding SVs of larger size (>100 kb), we acknowledge that the detection power of our data set was limited. Because the final scaffolding of the two genomes is based on a single (and not so dense)

TABLE 3 Enrichment in SVs associated with transposable elements in outliers of differentiation between *Coregonus clupeaformis* sp. Normal and *C. clupeaformis* sp. Dwarf

Type of SV	Population-scale analysis (MAF > 5%)		$F_{ST}$ outliers in both lakes				
	Number of SVs	Proportion of SVs	Number of SVs	Proportion of SVs	Odd-ratio	<i>p</i> -value (Fisher test)	<i>q</i> value (B & H correction)
Neither TE nor repeats	30,082	32%	213	29%	0.9	.885	1.00
Simple repeats	24,142	26%	22	3%	0.1	.000	1.00
Satellite	35	0%	1	0%	3.7	.243	0.38
Low complexity	774	1%	2	0%	0.3	.982	1.00
RNA repeats	2,303	2%	24	3%	1.3	.100	0.18
<b>TEs</b>							
dnaTE	<b>13,970</b>	<b>15%</b>	<b>193</b>	<b>27%</b>	<b>1.8</b>	<b>&lt;.001</b>	<b>&lt;0.001</b>
LINE	<b>6,725</b>	<b>7%</b>	<b>70</b>	<b>10%</b>	<b>1.3</b>	<b>.014</b>	<b>0.03</b>
SINE	<b>2,254</b>	<b>2%</b>	<b>39</b>	<b>5%</b>	<b>2.2</b>	<b>&lt;.001</b>	<b>&lt;0.001</b>
LTR	<b>10,691</b>	<b>11%</b>	<b>120</b>	<b>17%</b>	<b>1.4</b>	<b>&lt;.001</b>	<b>&lt;0.001</b>
Unknown TE	<b>2,776</b>	<b>3%</b>	<b>43</b>	<b>6%</b>	<b>2.0</b>	<b>&lt;.001</b>	<b>&lt;0.001</b>
RC/Helitron	21	0%	NA	NA	0.0	1.000	1.00

Lines in bold correspond to significant enrichment with  $p < .05$ .

linkage map, made from a Normal  $\times$  Dwarf hybrid family (Gagnaire et al., 2013a; Rogers et al., 2007), we could not search for large chromosomal rearrangements simply by contrasting the two genomes. This is unfortunate because large rearrangements such as inversions, fusions and translocations may be relevant for speciation as they often differ between closely-related sympatric species and contribute to reproductive isolation (Berdan et al., 2021; Faria & Navarro, 2010; Noor et al., 2001). In the case of *C. clupeaformis*, on the one hand, we do not expect a major effect of chromosomal rearrangements. First, the differentiation observed in SNPs and SVs is widespread along the genome and does not display the typical spatial clustering of differentiated regions observed between species pairs such as *Littorina saxatilis* (Morales et al., 2019) or *Helianthus* sp. (Todesco et al., 2020). Second, cytogenetic analysis showed that the *C. clupeaformis* sp. Normal and sp. Dwarf from these same lakes have an identical number of chromosomes (Dion-Côté et al., 2017). On the other hand, cytogenetic exploration showed subtle chromosomal polymorphism within and between them (Dion-Côté et al., 2017). For instance, chromosome 1 is longer in the Normal than in the Dwarf in Cliff Lake due to heterochromatin differences (Dion-Côté et al., 2017), a pattern that we also observed in the genome (121 vs. 99 Mb, Figure S8). We also note some peculiarities such as Chr22, for which sequences in *C. clupeaformis* sp. Normal are homologous to sequences belonging to Chr22 in the genome of *C. lavaretus* sp. Balchen but which we never managed to order as a full linkage group, probably because of the lack of recombination in the family used for the linkage map. Since the mother used for the linkage map is a hybrid Dwarf  $\times$  Normal, any rearrangement differing between species (and affecting recombination at the heterozygote stage) may be absent from the final map, and hence from the present genomes. These chromosomal differences may lead to issues

with recombination during meiosis (Dion-Côté et al., 2015; Faria & Navarro, 2010), contributing to reproductive isolation and speciation (Hoffmann & Rieseberg, 2008; Kirkpatrick & Barton, 2006). In the future, it would be worthwhile to explore large-scale chromosomal rearrangements in *C. clupeaformis* sp. Normal and *C. clupeaformis* sp. Dwarf in depth to understand the role of chromosomal polymorphism in speciation. However, this will require improved genome scaffolding based on Hi-C chromatin contacts (which was attempted here without success) or separate linkage maps.

Beyond the contrast between *C. clupeaformis* sp. Normal and *C. clupeaformis* sp. Dwarf, the new genome assemblies also provide relevant information about the evolution of genomes at a higher taxonomic level. Salmonids have experienced a recent whole-genome duplication, followed by different events of rediploidization, as well as important chromosomal rearrangements such as fusions (Blumstein et al., 2020; Glasauer & Neuhauss, 2014; Lien et al., 2016; Macqueen & Johnston, 2014). Here, as often observed in salmonids, synteny was high between *C. clupeaformis* sp. Normal, *C. clupeaformis* sp. Dwarf and closely related species such as the European whitefish *C. lavaretus* sp. Balchen. The same groups of chromosomes appear to be metacentric and bear residual tetrasomy in *C. clupeaformis* as in its related species *C. ardetii* (Blumstein et al., 2020). Chromosomal comparison with *C. ardetii* and *C. lavaretus* also suggested shared fusion and fission of ancestral chromosomes and a consistent karyotype between the different coregonids (Blumstein et al., 2020; De-Kayne & Feulner, 2018). This would suggest that the majority of rediploidization processes occurred before the split of the different *Coregonus* species, which would all share a relatively similar karyotype. That being said, it should be kept in mind that the residual tetrasomy observed on a subset of chromosomes makes it difficult to fully ascertain synteny vs. rearrangements within and

between species on those chromosomes. Moreover, *C. clupeaformis* genomes remain extremely complex with several regions that end up collapsed by genome assembly (at least 126 Mb, 5% of the chromosome-anchored genome), as was previously reported in other salmonid genomes (De-Kayne et al., 2020; Lien et al., 2016). Therefore, while the *Coregonus* reference genome assemblies provide an important first step, refining the assemblies and complementing by cytogenetic or chromatin-contact data will be valuable to further explore the timing and modalities of rediploidization in coregonids, and its possible contribution to speciation.

Salmonid genomes are also littered with TEs and *C. clupeaformis* was no exception: interspersed repeats accounted for about 60% of the genome. This amount is comparable to *Salmo salar* (60%; Lien et al., 2016) and *Coregonus lavaretus* "Balchen" (52%; De-Kayne et al., 2020). Moreover, not all TE copies are shared by all individuals and our results highlighted that they were responsible for a third of the SVs detected within and between species. This is also consistent with observations made on other species, such as Atlantic salmon *Salmo salar* (Bertolotti et al., 2020) or crows *Corvus* sp. (Weissensteiner et al., 2020), in which young and active TEs generate numerous insertions and deletions between samples. It has been hypothesized that bursts of transposon activity may contribute to speciation (de Boer et al., 2007), or at least that TEs may rapidly generate genetic variation differentiating species (Serrato-Capuchina & Matute, 2018). Our data strongly support this hypothesis since the most differentiated SVs between Dwarf and Normal in both lakes were enriched in several classes of TEs. A large part of the fixed genetic variation between species corresponds to an insertion or a deletion of a given TE. It is worth noting that this pattern is widespread across the genome rather than centred on a few loci. Such extensive differentiation suggests a progressive and differential TE accumulation without gene flow, probably in allopatry during the Pleistocene glaciation (~15,000 generations/60,000 years ago) that may have contributed to the maintenance of reproductive isolation during the postglacial sympatric phase following secondary contact (~3000 generations/12,000 years ago) (Jacobsen et al., 2012; Rougeux et al., 2017). Accumulations of different TEs between lineages may be quite rapid as active TEs have a high mutation rate, as observed in *Daphnia* with an order of  $10^{-5}$  gain or loss per copy per generation (Ho et al., 2021). TEs can also contribute to reproductive isolation by altering gene structure, expression pattern and chromosome organization (Dubin et al., 2018; Goodier, 2016). In fact, TE deregulation is known to generate postzygotic breakdown in Dwarf × Normal hybrids (Dion-Côté et al., 2014), which has been associated with epigenetic (DNA methylation) reprogramming in hybrids (Laporte et al., 2019). Moreover, this supported the hypothesis that TE transcriptional derepression, perhaps due to different TE silencing mechanisms that evolved in allopatry, may be the cause for both massive misregulation of gene expression and abnormal embryonic development and death in hybrids (Dion-Côté et al., 2014; Renaut et al., 2009). Both in previous studies and in our study, the same TE families emerged as associated with species differentiation, namely Tc1-mariner and hAT-Ac as well as LTR-Gypsy, Line-L2 and Line-RexBabar. Together, cumulative

evidence points towards a major role of several TE families in the reproductive isolation of Dwarf and Normal, involving TEs distributed throughout the genome rather than in a few barrier loci.

A peculiarity of the speciation between *C. clupeaformis* sp. Normal and *C. clupeaformis* sp. Dwarf is the character displacement in the Acadian lineage towards a dwarf limnetic species upon secondary contact with the American lineage, a process which occurred independently in separate lakes of the suture zone, resulting in two ecologically distinct sympatric species, the Dwarf and the Normal (Bernatchez et al., 2010a; Landry et al., 2007; Rougeux et al., 2017). Previous work revealed that strong parallelism at the phenotypic level between lakes was accompanied by weak parallelism at the genome level (Gagnaire, Pavey, et al., 2013; Lu & Bernatchez, 1999; Rougeux et al., 2019). With a higher density of variants being screened, our results corroborate those from these previous studies. The pattern of differentiation between species was indeed specific to each lake. However, it is worth noting the excess of shared outliers of differentiation, for both SNPs and SVs, and that differences of allelic frequencies were more often in the same direction (e.g., higher allelic frequency in dwarf species in both lakes) than expected by chance. A large fraction of such parallelism probably reflects historical divergence between allopatric lineages, possibly reinforced by the result of comparable ecological response to selection. It is also possible that shared regions of differentiation reflect regions of the genome more resistant to gene flow, such as low recombination regions, as observed in *Ficedula* flycatchers (Burri et al., 2015). General patterns of TE enrichment in outlier SVs, as well as gene ontology enrichment, also converged between lakes. This suggests that the processes driving genetic divergence between species were probably similar between lakes, namely through shared historical divergence and similar ecological selection imposed by the use of distinct trophic niches (Bernatchez et al., 2010a). However, they were buffered by lake-specific contingency at finer molecular level, for instance, associated with the effect of genetic drift on available standing genetic variation within each lake (Gagnaire, Pavey, et al., 2013).

Studying two types of genetic variants in tandem, SVs and SNPs, at the population level showed similar patterns and level of differentiation between species and between lakes. On the one hand, this confirms that evaluating population/species structure requires neither a diversity of variants nor a large amount of markers. In fact, the  $F_{ST}$  values observed at the scale of the entire genome for both types of variants and in both lakes were strikingly similar to values measured based on a much smaller subset of markers. For instance, based on the RADseq genotyping of about 2500 SNP loci, Gagnaire, Pavey, et al. (2013) reported  $F_{ST}$  values of 0.12 and 0.10 between Dwarf and Normal from Cliff Lake and from Indian Lake respectively, compared to values of 0.18 and 0.10 here for SNPs and 0.17 and 0.06 for SVs. On the other hand, studying different kinds of variants with similar filters shows a large amount of nonrare SVs (i.e., SVs found in more than two of 64 chromosomes; 32 diploid individuals). Because of their size, the accumulation of SVs at intermediate frequency in natural populations thus represents a non-negligible aspect of genetic variation, as they covered at least five times more of the genome



than SNPs. This point is increasingly underlined by studies in population genomics and evolutionary genomics (Catanach et al., 2019; Mérot et al., 2020; Weissensteiner et al., 2020) and means that a full understanding of genetic variation cannot overlook SVs. However, it remains difficult to study SVs at the population level. Short reads are more accessible when sequencing a large number of individuals but they proved to be less powerful for characterizing SVs (Mahmoud et al., 2019). For instance, here we found around five times fewer SVs with 32 samples sequenced with short reads than with two samples sequenced with long reads. Our study also used short reads at shallow/medium coverage (~5x) which may be suboptimal to detect and genotype SNPs and SVs with confidence. However, there are ways to handle the uncertainty associated with a low number of supporting reads, such as working within a genotype likelihood framework (Buerkle & Gompert, 2013; Lou et al., 2020). Recent studies have proposed relying on mixed data sets (e.g., combining long- and short-read sequence data, combining high and shallow coverage) to achieve together a good catalogue of SVs and then perform population genomic studies based on their variation (Logsdon et al., 2020). We have achieved this in this study by first characterizing SVs using high-depth long reads on a limited number of samples, and second by genotyping known SVs with medium-coverage short reads on a greater number of samples. To achieve this, genome-graph-based approaches were particularly relevant, allowing us to build a variation-aware reference graph (Garrison et al., 2018), and then perform unbiased mapping of reads to this graph (Sirén et al., 2020). Such two-step approaches have also been used in a handful of studies looking at SVs in chocolate trees *Theobroma cacao* (Hämälä et al., 2021), soybeans *Glycine max* (Lemay et al., 2021) and potato beetle *Leptinotarsa decemlineata* (Cohen et al., 2021). Based on this, we believe that the combination of second- and third-generation sequencing is promising to study structural polymorphism within a population genomics framework and will allow the inclusion of SVs in studies of speciation and adaptation genomics.

#### ACKNOWLEDGEMENTS

We thank M. Suga for contributing to fieldwork, K. Wellband for help with annotation, M-A. Lemay for help with SV analysis, M. Leitwein for help with the analysis of metacentric chromosomes, and A-L. Ferchaud and R. De Kayne for their help with synteny analysis and plotting. We are very grateful to the teams that developed vg, giraffe and pgbp for their guidance in using the genome-graph tools as well as J. Laroche at the IBIS Bioinformatic platform ([www.ibis.ulaval.ca](http://www.ibis.ulaval.ca)) for his support. We are also grateful to Editor L. Rieseberg and three anonymous reviewers for their extensive and constructive comments on a previous version of the manuscript.

#### CONFLICT OF INTEREST

The authors declare no conflict of interest.

#### AUTHOR CONTRIBUTIONS

This study is part of the long-term research programme of L.B. on the adaptive radiation and ecological speciation of lake whitefish.

L.B., C.R. and C.M. conceived the study. M.Á. and M.K. performed the Nanopore sequencing. M.M. and S.L. assembled the genome and E.N. and C.M. scaffolded the genome. J.M.F. analysed the TEs. C.M. and K.S. did the analyses with contributions from C.V., E.N., M.L. and S.L. C.M. wrote the manuscript with contributions from all authors.

#### DATA AVAILABILITY STATEMENT

Final genome assemblies are available on NCBI under the id ASM1839867v1 and ASM2061545v1. Long- and short-read sequences are available on NCBI under project nos. PRJNA715481, PRJNA767633, PRJNA820751. Code is available on Github for the different pipelines listed in the methods ([https://github.com/clairmerot/wgs\\_sample\\_preparation](https://github.com/clairmerot/wgs_sample_preparation), [https://github.com/clairmerot/lepmap3\\_pipeline](https://github.com/clairmerot/lepmap3_pipeline), [https://github.com/clairmerot/assembly\\_SV](https://github.com/clairmerot/assembly_SV), [https://github.com/clairmerot/long\\_read\\_SV](https://github.com/clairmerot/long_read_SV), [https://github.com/clairmerot/SR\\_SV](https://github.com/clairmerot/SR_SV), [https://github.com/clairmerot/genotyping\\_SV](https://github.com/clairmerot/genotyping_SV)).

#### ORCID

Claire Mérot  <https://orcid.org/0000-0003-2607-7818>

#### REFERENCES

- Abel, H. J., Larson, D. E., Regier, A. A., Chiang, C., Das, I., Kanchi, K. L., & Reeves, C. (2020). Mapping and characterization of structural variation in 17,795 human genomes. *Nature*, 583(7814), 83–89.
- Aljanabi, S. M., & Martinez, I. (1997). Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. *Nucleic Acids Research*, 25(22), 4692–4693. <https://doi.org/10.1093/nar/25.22.4692>
- Allendorf, F. W., Bassham, S., Cresko, W. A., Limborg, M. T., Seeb, L. W., & Seeb, J. E. (2015). Effects of crossovers between homeologs on inheritance and population genomics in polyploid-derived salmonid fishes. *Journal of Heredity*, 106(3), 217–227. <https://doi.org/10.1093/jhered/esv015>
- Allendorf, F. W., & Thorgaard, G. H. (1984). Tetraploidy and the evolution of salmonid fishes. In: B. J. Turner (Eds), *Evolutionary genetics of fishes*, Monographs in Evolutionary Biology (pp. 1–53). Springer.
- Alonge, M., Soyk, S., Ramakrishnan, S., Wang, X., Goodwin, S., Sedlazeck, F. J., Lippman, Z. B., & Schatz, M. C. (2019). RaGOO: Fast and accurate reference-guided scaffolding of draft genomes. *Genome Biology*, 20(1), 1–17. <https://doi.org/10.1186/s13059-019-1829-6>
- Berdan, E. L., Fuller, R. C., & Kozak, G. M. (2021). Genomic landscape of reproductive isolation in Lucania killifish: The role of sex loci and salinity. *Journal of Evolutionary Biology*, 34(1), 157–174.
- Bernatchez, L., & Dodson, J. J. (1990). Allopatric origin of sympatric populations of lake whitefish (*Coregonus clupeaformis*) as revealed by mitochondrial-DNA restriction analysis. *Evolution*, 44(5), 1263–1271.
- Bernatchez, L., Renaut, S., Whiteley, A. R., Derome, N., Jeukens, J., Landry, L., & Rogers, S. M. (2010a). On the origin of species: Insights from the ecological genomics of lake whitefish. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1547), 1783–1800.
- Bernatchez, L., Renaut, S., Whiteley, A. R., Derome, N., Jeukens, J., Landry, L., & Rogers, S. M. (2010b). On the origin of species: Insights from the ecological genomics of lake whitefish. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1547), 1783–1800.
- Bertolotti, A. C., Layer, R. M., Gundappa, M. K., Gallagher, M. D., Pehlivanoglu, E., Nome, T., Robledo, D., Kent, M. P., Røsaeg, L. L., Holen, M. M., Mulugeeta, T. D., Ashton, T. J., Hindar, K., Særgrov, H., Florø-Larsen, B., Erkinaro, J., Primmer, C. R., Bernatchez, L., Martin,

- S. A. M., ... Macqueen, D. J. (2020). The structural variation landscape in 492 Atlantic salmon genomes. *Nature Communications*, 11(1), 5176. <https://doi.org/10.1038/s41467-020-18972-x>
- Blumstein, D. M., Campbell, M. A., Hale, M. C., Sutherland, B. J., McKinney, G. J., Stott, W., & Larson, W. A. (2020). Comparative genomic analyses and a novel linkage map for cisco (*Coregonus artedii*) provide insights into chromosomal evolution and rediploidization across salmonids. *G3: Genes, Genomes, Genetics*, 10(8), 2863–2878.
- Breese, M. R., & Liu, Y. (2013). NGSUtils: A software suite for analyzing and manipulating next-generation sequencing data sets. *Bioinformatics*, 29(4), 494–496. <https://doi.org/10.1093/bioinformatics/bts731>
- Buerkle, 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>
- Burri, R., Nater, A., Kawakami, T., Mugal, C. F., Olason, P. I., Smeds, L., & Garamszegi, L. Z. (2015). Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of *Ficedula flycatchers*. *Genome Research*, 25(11), 1656–1665.
- Cabanettes, F., & Klopp, C. (2018). D-GENIES: Dot plot large genomes in an interactive, efficient and simple way. *PeerJ*, 6, e4958. <https://doi.org/10.7717/peerj.4958>
- Catanach, A., Crowhurst, R., Deng, C., David, C., Bernatchez, L., & Wellenreuther, M. (2019). The genomic pool of standing structural variation outnumbers single nucleotide polymorphism by threefold in the marine teleost *Chrysophrys auratus*. *Molecular Ecology*, 28(6), 1210–1223.
- Catchen, J., Amores, A., & Bassham, S. (2020). Chromonomer: A tool set for repairing and enhancing assembled genomes through integration of genetic maps and conserved synteny. *G3 Genes/genomes/genetics*, 10(11), 4115–4128. <https://doi.org/10.1534/g3.120.401485>
- Chaisson, M. J. P., Sanders, A. D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E. J., Rodriguez, O. L., Guo, L. I., Collins, R. L., Fan, X., Wen, J., Handsaker, R. E., Fairley, S., Kronenberg, Z. N., Kong, X., Hormozdiari, F., Lee, D., Wenger, A. M., ... Lee, C. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications*, 10(1), 1–16. <https://doi.org/10.1038/s41467-018-08148-z>
- Chakraborty, M., Baldwin-Brown, J. G., Long, A. D., & Emerson, J. (2016). Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Research*, 44(19), e147.
- Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17), i884–i890. <https://doi.org/10.1093/bioinformatics/bty560>
- Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A. J., Kruglyak, S., & Saunders, C. T. (2016). Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, 32(8), 1220–1222. <https://doi.org/10.1093/bioinformatics/btv710>
- Cohen, Z., Hawthorne, D., & Schoville, S. (2021). The role of structural variants in pest adaptation and genome evolution of the Colorado potato beetle, *Leptinotarsa decemlineata* (Say). *Authorea Preprints*.
- Cruikshank, T. E., & Hahn, M. W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, 23(13), 3133–3157. <https://doi.org/10.1111/mec.12796>
- Dalziel, A. C., Laporte, M., Rougeux, C., Guderley, H., & Bernatchez, L. (2017). Convergence in organ size but not energy metabolism enzyme activities among wild Lake Whitefish (*Coregonus clupeaformis*) species pairs. *Molecular Ecology*, 26(1), 225–244.
- de Boer, J. G., Yazawa, R., Davidson, W. S., & Koop, B. F. (2007). Bursts and horizontal evolution of DNA transposons in the speciation of pseudotetraploid salmonids. *BMC Genomics*, 8(1), 1–10.
- De Coster, W., De Rijk, P., De Roeck, A., De Pooter, T., D'Hert, S., Strazisar, M., Slegers, K., & Van Broeckhoven, C. (2019). Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. *Genome Research*, 29(7), 1178–1187. <https://doi.org/10.1101/gr.244939.118>
- De-Kayne, R., & Feulner, P. G. (2018). A European whitefish linkage map and its implications for understanding genome-wide synteny between salmonids following whole genome duplication. *G3 Genes/genomes/genetics*, 8(12), 3745–3755. <https://doi.org/10.1534/g3.118.200552>
- De-Kayne, R., Zoller, S., & Feulner, P. G. (2020). A de novo chromosome-level genome assembly of *Coregonus* sp. "Balchen": One representative of the Swiss Alpine whitefish radiation. *Molecular Ecology Resources*, 20(4), 1093–1109.
- Dion-Côté, A.-M., Renaud, S., Normandeau, E., & Bernatchez, L. (2014). RNA-seq reveals transcriptomic shock involving transposable elements reactivation in hybrids of young lake whitefish species. *Molecular Biology and Evolution*, 31(5), 1188–1199. <https://doi.org/10.1093/molbev/msu069>
- Dion-Côté, A., Symonová, R., Lamaze, F. C., Pelikánová, Š., Ráb, P., & Bernatchez, L. (2017). Standing chromosomal variation in Lake Whitefish species pairs: The role of historical contingency and relevance for speciation. *Molecular Ecology*, 26(1), 178–192.
- Dion-Côté, A.-M., Symonová, R., Ráb, P., & Bernatchez, L. (2015). Reproductive isolation in a nascent species pair is associated with aneuploidy in hybrid offspring. *Proceedings of the Royal Society B: Biological Sciences*, 282(1802), 20142862. <https://doi.org/10.1098/rspb.2014.2862>
- Dubin, M. J., Scheid, O. M., & Becker, C. (2018). Transposons: A blessing curse. *Current Opinion in Plant Biology*, 42, 23–29. <https://doi.org/10.1016/j.pbi.2018.01.003>
- Dufresnes, C., Brelsfors, A., Jeffries, D. L., Mazepa, G., Suchan, T., Canestrelli, D., Nicieza, A., Fumagalli, L., Dubey, S., Martínez-Solano, I., Litvinchuk, S. N., Vences, M., Perrin, N., & Crochet, P.-A. (2021). Mass of genes rather than master genes underlie the genomic architecture of amphibian speciation. *Proceedings of the National Academy of Sciences*, 118(36), e2103963118. <https://doi.org/10.1073/pnas.2103963118>
- Evans, M., & Bernatchez, L. (2012). Oxidative phosphorylation gene transcription in whitefish species pairs reveals patterns of parallel and nonparallel physiological divergence. *Journal of Evolutionary Biology*, 25(9), 1823–1834. <https://doi.org/10.1111/j.1420-9101.2012.02570.x>
- Faria, R., & Navarro, A. (2010). Chromosomal speciation revisited: Rearranging theory with pieces of evidence. *Trends in Ecology & Evolution*, 25(11), 660–669. <https://doi.org/10.1016/j.tree.2010.07.008>
- Feulner, P. G. D., Chain, F. J. J., Panchal, M., Eizaguirre, C., Kalbe, M., Lenz, T. L., Mundry, M., Samonte, I. E., Stoll, M., Milinski, M., Reusch, T. B. H., & Bornberg-bauer, E. (2013). Genome-wide patterns of standing genetic variation in a marine population of three-spined sticklebacks. *Molecular Ecology*, 22(3), 635–649. <https://doi.org/10.1111/j.1365-294X.2012.05680.x>
- Feulner, P., & De-Kayne, R. (2017). Genome evolution, structural rearrangements and speciation. *Journal of Evolutionary Biology*, 30(8), 1488–1490. <https://doi.org/10.1111/jeb.13101>
- Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*, 117(17), 9451–9457. <https://doi.org/10.1073/pnas.1921046117>
- Gagnaire, P., Normandeau, E., Pavey, S. A., & Bernatchez, L. (2013). Mapping phenotypic, expression and transmission ratio distortion QTL using RAD markers in the Lake Whitefish (*Coregonus clupeaformis*). *Molecular Ecology*, 22(11), 3036–3048.

- Gagnaire, P., Pavey, S. A., Normandeau, E., & Bernatchez, L. (2013). The genetic architecture of reproductive isolation during speciation-with-gene-flow in lake whitefish species pairs assessed by RAD sequencing. *Evolution*, 67(9), 2483–2497. <https://doi.org/10.1111/evo.12075>
- Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., Jones, W., Garg, S., Markello, C., Lin, M. F., Paten, B., & Durbin, R. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, 36(9), 875–879. <https://doi.org/10.1038/nbt.4227>
- Glasauer, S. M., & Neuhauss, S. C. (2014). Whole-genome duplication in teleost fishes and its evolutionary consequences. *Molecular Genetics and Genomics*, 289(6), 1045–1060. <https://doi.org/10.1007/s00438-014-0889-2>
- Goel, M., Sun, H., Jiao, W.-B., & Schneeberger, K. (2019). SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biology*, 20(1), 1–13. <https://doi.org/10.1186/s13059-019-1911-0>
- Goodier, J. L. (2016). Restricting retrotransposons: A review. *Mobile DNA*, 7(1), 1–30. <https://doi.org/10.1186/s13100-016-0070-z>
- Gu, Z., Gu, L., Eils, R., Schlessner, M., & Brors, B. (2014). Circlize implements and enhances circular visualization in R. *Bioinformatics*, 30(19), 2811–2812. <https://doi.org/10.1093/bioinformatics/btu393>
- Hämälä, T., Wafala, E. K., Guiltinan, M. J., Ralph, P. E., dePamphilis, C. W., & Tiffin, P. (2021). Genomic structural variants constrain and facilitate adaptation in natural populations of *Theobroma cacao*, the chocolate tree. *Proceedings of the National Academy of Sciences*, 118(35).
- Hardie, D. C., & Hebert, P. D. (2003). The nucleotypic effects of cellular DNA content in cartilaginous and ray-finned fishes. *Genome*, 46(4), 683–706. <https://doi.org/10.1139/g03-040>
- Harris, R. S. (2007). *Improved pairwise alignment of genomic DNA*. The Pennsylvania State University.
- Hejase, H. A., Salman-Minkov, A., Campagna, L., Hubisz, M. J., Lovette, I. J., Gronau, I., & Siepel, A. (2020). Genomic islands of differentiation in a rapid avian radiation have been driven by recent selective sweeps. *Proceedings of the National Academy of Sciences*, 117(48), 30554. <https://doi.org/10.1073/pnas.2015987117>
- Heller, D., & Vingron, M. (2019). SVIM: structural variant identification using mapped long reads. *Bioinformatics*, 35(17), 2907–2915. <https://doi.org/10.1093/bioinformatics/btaz041>
- Heller, D., & Vingron, M. (2020). SVIM-asm: Structural variant detection from haploid and diploid genome assemblies. *Bioinformatics*, 36(22–23), 5519–5521.
- Henderson, E. C., & Brelsford, A. (2020). Genomic differentiation across the speciation continuum in three hummingbird species pairs. *BMC Evolutionary Biology*, 20(1), 1–11. <https://doi.org/10.1186/s12862-020-01674-9>
- Hickey, G., Heller, D., Monlong, J., Sibbesen, J. A., Sirén, J., Eizenga, J., Dawson, E. T., Garrison, E., Novak, A. M., & Paten, B. (2020). Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biology*, 21(1), 1–17. <https://doi.org/10.1186/s13059-020-1941-7>
- Ho, E. K. H., Bellis, E. S., Calkins, J., Adrion, J. R., Latta, L. C. IV, & Schaack, S. (2021). Engines of change: Transposable element mutation rates are high and variable within *Daphnia magna*. *PLOS Genetics*, 17(11), e1009827. <https://doi.org/10.1371/journal.pgen.1009827>
- Ho, S. S., Urban, A. E., & Mills, R. E. (2019). Structural variation in the sequencing era. *Nature Reviews Genetics*, 21(3), 171–189. <https://doi.org/10.1038/s41576-019-0180-9>
- 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, 21–42. <https://doi.org/10.1146/annurev.ecolsys.39.110707.173532>
- Huddleston, J., Ranade, S., Malig, M., Antonacci, F., Chaisson, M., Hon, L., Sudmant, P. H., Graves, T. A., Alkan, C., Dennis, M. Y., Wilson, R. K., Turner, S. W., Korlach, J., & Eichler, E. E. (2014). Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Research*, 24(4), 688–696. <https://doi.org/10.1101/gr.168450.113>
- Jacobsen, M. W., Hansen, M. M., Orlando, L., Bekkevold, D., Bernatchez, L., Willerslev, E., & Gilbert, M. T. P. (2012). Mitogenome sequencing reveals shallow evolutionary histories and recent divergence time between morphologically and ecologically distinct European whitefish (*Coregonus* spp.). *Molecular Ecology*, 21(11), 2727–2742.
- Jain, C., Rhie, A., Zhang, H., Chu, C., Walenz, B. P., Koren, S., & Phillippy, A. M. (2020). Weighted minimizer sampling improves long read mapping. *Bioinformatics*, 36(Suppl\_1), i111–i118. <https://doi.org/10.1093/bioinformatics/btaa435>
- Jiggins, C. D. (2019). Can genomics shed light on the origin of species? *PLoS Biology*, 17(8), e3000394. <https://doi.org/10.1371/journal.pbio.3000394>
- Kirkpatrick, M., & Barton, N. (2006). Chromosome inversions, local adaptation and speciation. *Genetics*, 173(1), 419–434. <https://doi.org/10.1534/genetics.105.047985>
- Kirsche, M., Prabhu, G., Sherman, R., Ni, B., Aganezov, S., & Schatz, M. C. (2021). Jasmine: Population-scale structural variant comparison and analysis. *BioRxiv*.
- Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, 37(5), 540–546. <https://doi.org/10.1038/s41587-019-0072-8>
- Korneliusson, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics*, 15(1), 356. <https://doi.org/10.1186/s12859-014-0356-4>
- Landis, J. B., Soltis, D. E., Li, Z., Marx, H. E., Barker, M. S., Tank, D. C., & Soltis, P. S. (2018). Impact of whole-genome duplication events on diversification rates in angiosperms. *American Journal of Botany*, 105(3), 348–363. <https://doi.org/10.1002/ajb2.1060>
- Landry, L., Vincent, W., & Bernatchez, L. (2007). Parallel evolution of lake whitefish dwarf ecotypes in association with limnological features of their adaptive landscape. *Journal of Evolutionary Biology*, 20(3), 971–984. <https://doi.org/10.1111/j.1420-9101.2007.01304.x>
- Laporte, M., Dalziel, A. C., Martin, N., & Bernatchez, L. (2016). Adaptation and acclimation of traits associated with swimming capacity in Lake Whitefish (*Coregonus clupeaformis*) ecotypes. *BMC Evolutionary Biology*, 16(1), 1–13. <https://doi.org/10.1186/s12862-016-0732-y>
- Laporte, M., Le Luyer, J., Rougeux, C., Dion-Côté, A.-M., Krick, M., & Bernatchez, L. (2019). DNA methylation reprogramming, TE derepression, and postzygotic isolation of nascent animal species. *Science*, Advances, 5(10), eaaw1644. <https://doi.org/10.1126/sciadv.aaw1644>
- Laporte, M., Rogers, S. M., Dion-Côté, A.-M., Normandeau, E., Gagnaire, P.-A., Dalziel, A. C., & Bernatchez, L. (2015). RAD-QTL mapping reveals both genome-level parallelism and different genetic architecture underlying the evolution of body shape in lake whitefish (*Coregonus clupeaformis*) species pairs. *G3: Genes, Genomes, Genetics*, 5(7), 1481–1491.
- Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). LUMPY: A probabilistic framework for structural variant discovery. *Genome Biology*, 15(6), R84. <https://doi.org/10.1186/gb-2014-15-6-r84>
- Legendre, P., & Legendre, L. F. (2012). *Numerical ecology*, Vol. 24. Elsevier.
- Lemay, M.-A., Sibbesen, J. A., Torkamaneh, D., Hamel, J., Levesque, R. C., & Belzile, F. (2021). Combined use of Oxford Nanopore and Illumina sequencing yields insights into soybean structural variation biology. *BioRxiv*.
- Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18), 3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>
- 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>

- Li, H., Feng, X., & Chu, C. (2020). The design and construction of reference pangenome graphs with minigraph. *Genome Biology*, 21(1), 1–19. <https://doi.org/10.1186/s13059-020-02168-z>
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16), 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Kent, M. P., Nome, T., & Zimin, A. (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature*, 533(7602), 200–205.
- Limborg, M. T., McKinney, G. J., Seeb, L. W., & Seeb, J. E. (2016). Recombination patterns reveal information about centromere location on linkage maps. *Molecular Ecology Resources*, 16(3), 655–661. <https://doi.org/10.1111/1755-0998.12484>
- Lockwood, S. F., Seavey, B. T., Dillinger, R. E. Jr, & Bickham, J. W. (1991). Variation in DNA content among age classes of broad whitefish (*Coregonus nasus*) from the Sagavanirktok River delta. *Canadian Journal of Zoology*, 69(5), 1335–1338.
- Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, 21(10), 597–614. <https://doi.org/10.1038/s41576-020-0236-x>
- Lou, R. N., Jacobs, A., Wilder, A., & Therkildsen, N. O. (2020). A beginner's guide to low-coverage whole genome sequencing for population genomics. *Molecular Ecology* 30, 5966–5993.
- Lu, G., & Bernatchez, L. (1999). Correlated trophic specialization and genetic divergence in sympatric lake whitefish ecotypes (*Coregonus clupeaformis*): Support for the ecological speciation hypothesis. *Evolution*, 53(5), 1491–1505.
- Macqueen, D. J., & Johnston, I. A. (2014). A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. *Proceedings of the Royal Society B: Biological Sciences*, 281(1778), 20132881. <https://doi.org/10.1098/rspb.2013.2881>
- Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., & Sedlazeck, F. J. (2019). Structural variant calling: The long and the short of it. *Genome Biology*, 20(1), 246. <https://doi.org/10.1186/s13059-019-1828-7>
- Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLoS Computational Biology*, 14(1), e1005944. <https://doi.org/10.1371/journal.pcbi.1005944>
- Marques, D. A., Lucek, K., Meier, J. I., Mwaiko, S., Wagner, C. E., Excoffier, L., & Seehausen, O. (2016). Genomics of rapid incipient speciation in sympatric threespine stickleback. *PLoS Genetics*, 12(2), e1005887. <https://doi.org/10.1371/journal.pgen.1005887>
- Martin, S. H., Davey, J. W., Salazar, C., & Jiggins, C. D. (2019). Recombination rate variation shapes barriers to introgression across butterfly genomes. *PLoS Biology*, 17(2), e2006288. <https://doi.org/10.1371/journal.pbio.2006288>
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & 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>
- Meier, J. I., Marques, D. A., Wagner, C. E., Excoffier, L., & Seehausen, O. (2018). Genomics of parallel ecological speciation in Lake Victoria cichlids. *Molecular Biology and Evolution*, 35(6), 1489–1506. <https://doi.org/10.1093/molbev/msy051>
- Meisner, J., & Albrechtsen, A. (2018). Inferring population structure and admixture proportions in low-depth NGS data. *Genetics*, 210(2), 719–731. <https://doi.org/10.1534/genetics.118.301336>
- Mérot, C., Oomen, R. A., Tigano, A., & Wellenreuther, M. (2020). A road-map for understanding the evolutionary significance of structural genomic variation. *Trends in Ecology & Evolution*, 35(7), 561–572. <https://doi.org/10.1016/j.tree.2020.03.002>
- Morales, H. E., Faria, R., Johannesson, K., Larsson, T., Panova, M., Westram, A. M., & Butlin, R. K. (2019). Genomic architecture of parallel ecological divergence: Beyond a single environmental contrast. *Science Advances*, 5(12), eaav9963. <https://doi.org/10.1126/sciadv.aav9963>
- Nadeau, N. J., & Kawakami, T. (2019). Population genomics of speciation and admixture. In O. P. Rajora (Ed.), *Population genomics: Concepts, approaches and applications* (pp. 613–653). Springer International Publishing.
- Noor, M. A., Grams, K. L., Bertucci, L. A., Almendarez, Y., Reiland, J., & Smith, K. R. (2001). The genetics of reproductive isolation and the potential for gene exchange between *Drosophila pseudoobscura* and *D. persimilis* via backcross hybrid males. *Evolution*, 55(3), 512–521.
- Ou, S., & Jiang, N. (2019). LTR\_FINDER\_parallel: Parallelization of LTR\_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mobile DNA*, 10(1), 1–3. <https://doi.org/10.1186/s13100-019-0193-0>
- Phillips, R. B., Reed, K. M., & Ráb, P. (1996). Revised karyotypes and chromosome banding of coregonid fishes from the Laurentian Great Lakes. *Canadian Journal of Zoology*, 74(2), 323–329. <https://doi.org/10.1139/z96-040>
- Rastas, P. (2017). Lep-MAP3: Robust linkage mapping even for low-coverage whole genome sequencing data. *Bioinformatics*, 33(23), 3726–3732. <https://doi.org/10.1093/bioinformatics/btx494>
- Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., & Korbel, J. O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18), i333–i339. <https://doi.org/10.1093/bioinformatics/bts378>
- Ravinet, M., Faria, R., Butlin, R. K., Galindo, J., Bierne, N., Rafajlović, M., Noor, M. A. F., Mehlig, B., & 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>
- Renaut, S., & Bernatchez, L. (2011). Transcriptome-wide signature of hybrid breakdown associated with intrinsic reproductive isolation in lake whitefish species pairs (*Coregonus* spp. Salmonidae). *Heredity*, 106(6), 1003–1011. <https://doi.org/10.1038/hdy.2010.149>
- Renaut, S., Nolte, A., & Bernatchez, L. (2009). Transcriptomic investigation of post-zygotic isolation in lake whitefish (*Coregonus clupeaformis*). *Molecular Biology and Evolution*, 26, 925–936.
- Rogers, S., & Bernatchez, L. (2006). The genetic basis of intrinsic and extrinsic post-zygotic reproductive isolation jointly promoting speciation in the lake whitefish species complex (*Coregonus clupeaformis*). *Journal of Evolutionary Biology*, 19(6), 1979–1994. <https://doi.org/10.1111/j.1420-9101.2006.01150.x>
- Rogers, S., & Bernatchez, L. (2007). The genetic architecture of ecological speciation and the association with signatures of selection in natural lake whitefish (*Coregonus* sp. Salmonidae) species pairs. *Molecular Biology and Evolution*, 24(6), 1423–1438. <https://doi.org/10.1093/molbev/msm066>
- Rogers, S. M., Gagnon, V., & Bernatchez, L. (2002). Genetically based phenotype-environment association for swimming behavior in lake whitefish ecotypes (*Coregonus clupeaformis* Mitchill). *Evolution*, 56(11), 2322–2329. <https://doi.org/10.1111/j.0014-3820.2002.tb00155.x>
- Rogers, S. M., Isabel, N., & Bernatchez, L. (2007). Linkage maps of the dwarf and normal lake whitefish (*Coregonus clupeaformis*) species complex and their hybrids reveal the genetic architecture of population divergence. *Genetics*, 175(1), 375–398.
- Rougeux, C., Bernatchez, L., & Gagnaire, P.-A. (2017). Modeling the multiple facets of speciation-with-gene-flow toward inferring the divergence history of lake whitefish species pairs (*Coregonus clupeaformis*). *Genome Biology and Evolution*, 9(8), 2057–2074. <https://doi.org/10.1093/gbe/evx150>
- Rougeux, C., Gagnaire, P.-A., Præbel, K., Seehausen, O., & Bernatchez, L. (2019). Polygenic selection drives the evolution of convergent

- transcriptomic landscapes across continents within a Nearctic sister species complex. *Molecular Ecology*, 28(19), 4388–4403. <https://doi.org/10.1111/mec.15226>
- Scannell, D. R., Byrne, K. P., Gordon, J. L., Wong, S., & Wolfe, K. H. (2006). Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature*, 440(7082), 341–345.
- Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., Von Haeseler, A., & Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, 15(6), 461–468. <https://doi.org/10.1038/s41592-018-0001-7>
- Seehausen, O., Butlin, R. K., Keller, I., Wagner, C. E., Boughman, J. W., Hohenlohe, P. A., Peichel, C. L., Saetre, G.-P., Bank, C., Brännström, Å., Brelsford, A., Clarkson, C. S., Eroukmanoff, F., Feder, J. L., Fischer, M. C., Foote, A. D., Franchini, P., Jiggins, C. D., Jones, F. C., ... Widmer, A. (2014). Genomics and the origin of species. *Nature Reviews Genetics*, 15(3), 176–192. <https://doi.org/10.1038/nrg3644>
- Serrato-Capuchina, A., & Matute, D. R. (2018). The role of transposable elements in speciation. *Genes*, 9(5), 254. <https://doi.org/10.3390/genes9050254>
- Sirén, J., Monlong, J., Chang, X., Novak, A. M., Eizenga, J. M., Markello, C., & Carroll, A. (2020). Genotyping common, large structural variations in 5,202 genomes using pangenomes, the Giraffe mapper, and the vg toolkit. *BioRxiv*.
- Smith, S. R., Normandeau, E., Djambazian, H., Nawarathna, P. M., Berube, P., Muir, A. M., & Luikart, G. (2021). A chromosome-anchored genome assembly for Lake Trout (*Salvelinus namaycush*). *Molecular Ecology Resources*, 22(2), 679–694.
- Soderlund, C., Bomhoff, M., & Nelson, W. M. (2011). SyMAP v3. 4: A turnkey synteny system with application to plant genomes. *Nucleic Acids Research*, 39(10), e68.
- Stevison, L. S., & McGaugh, S. E. (2020). It's time to stop sweeping recombination rate under the genome scan rug. *Molecular Ecology*, 29(22), 4249–4253. <https://doi.org/10.1111/mec.15690>
- Sutherland, B. J. G., Gosselin, T., Normandeau, E., Lamothe, M., Isabel, N., Audet, C., & Bernatchez, L. (2016). Salmonid chromosome evolution as revealed by a novel method for comparing RADseq linkage maps. *Genome Biology and Evolution*, 8(12), 3600–3617. <https://doi.org/10.1093/gbe/evw262>
- Tham, C. Y., Tirado-Magalanes, R., Goh, Y., Fullwood, M. J., Koh, B. T. H., Wang, W., Ng, C. H., Chng, W. J., Thiery, A., Tenen, D. G., & Benoukraf, T. (2020). NanoVar: Accurate characterization of patients' genomic structural variants using low-depth nanopore sequencing. *Genome Biology*, 21(1), 1–15. <https://doi.org/10.1186/s13059-020-01968-7>
- Ting, C.-T., Tsaur, S.-C., Sun, S., Browne, W. E., Chen, Y.-C., Patel, N. H., & Wu, C.-I. (2004). Gene duplication and speciation in *Drosophila*: Evidence from the *Odysseus* locus. *Proceedings of the National Academy of Sciences*, 101(33), 12232–12235.
- Tedesco, M., Owens, G. L., Bercovich, N., Légaré, J.-S., Soudi, S., Burge, D. O., & Imerovski, I. (2020). Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature*, 584(7822), 602–607.
- Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, 9(11), e112963. <https://doi.org/10.1371/journal.pone.0112963>
- Waples, R. K., Seeb, L. W., & Seeb, J. E. (2016). Linkage mapping with paralogs exposes regions of residual tetrasomic inheritance in chum salmon (*Oncorhynchus keta*). *Molecular Ecology Resources*, 16(1), 17–28.
- Weissensteiner, M. H., Bunikis, I., Catalán, A., Francoijs, K.-J., Knief, U., Heim, W., Peona, V., Pophaly, S. D., Sedlazeck, F. J., Suh, A., Warmuth, V. M., & Wolf, J. B. W. (2020). Discovery and population genomics of structural variation in a songbird genus. *Nature Communications*, 11(1), 1–11. <https://doi.org/10.1038/s41467-020-17195-4>
- Wellenreuther, M., & Bernatchez, L. (2018). Eco-evolutionary genomics of chromosomal inversions. *Trends in Ecology & Evolution*, 33(6), 427–440. <https://doi.org/10.1016/j.tree.2018.04.002>
- Wolf, J. B., & Ellegren, H. (2017). Making sense of genomic islands of differentiation in light of speciation. *Nature Reviews Genetics*, 18(2), 87–100. <https://doi.org/10.1038/nrg.2016.133>

## SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

**How to cite this article:** Mérot, C., Stenlökk, K. S. R., Venney, C., Laporte, M., Moser, M., Normandeau, E., Árnýasi, M., Kent, M., Rougeux, C., Flynn, J. M., Lien, S., & Bernatchez, L. (2022). Genome assembly, structural variants, and genetic differentiation between lake whitefish young species pairs (*Coregonus* sp.) with long and short reads. *Molecular Ecology*, 00, 1–20. <https://doi.org/10.1111/mec.16468>

Supplementary Materials for

Genome assembly, structural variants, and genetic differentiation  
between Lake Whitefish young species pairs (*Coregonus* sp.) with long  
and short reads

Claire Mérot<sup>1\*</sup>, Kristina S R Stenløkk<sup>2</sup>, Clare Venney<sup>1</sup>, Martin Laporte<sup>1,3</sup>, Michel Moser<sup>2</sup>, Eric Normandeau<sup>1</sup>, Mariann Árnýasi<sup>2</sup>, Matthew Kent<sup>2</sup>, Clément Rougeux<sup>1</sup>, Jullien M. Flynn<sup>4</sup>, Sigbjørn Lien<sup>2</sup>, Louis Bernatchez<sup>1</sup>

Table S1: Statistics for the genome assemblies before anchoring into chromosomes. ....	2
Table S2: Correspondence between chromosomes in Coregonids.....	3
Table S3: Coordinates of putatively collapsed duplicated regions in the reference assembly of the Normal Lake Whitefish.....	4
Table S4: Proportion of interspersed repeats in the Normal genome and in the sequences of the structural variants .....	5
Table S5: Breakdown of the proportion of transposable elements for the main families in the Normal genome and in the sequences of the structural variants.....	5
Table S6: Portion of the genome masked by different families of TEs .....	6
Table S7: Enrichment in SVs associated with transposable elements in outliers of differentiation between Dwarf and Normal Whitefish. ....	8
Table S9: Coordinates of previously identified QTL in the new reference genome.....	11
Figure S1: Homologous chromosomes with <i>Coregonus artedii</i> .....	12
Figure S2: Homologous chromosomes with <i>Coregonus lavaretus</i> .....	13
Figure S3: Homologous chromosomes with the previous map of <i>Coregonus clupeaformis</i> .....	14
Figure S4: Recombination frequency estimates (RFm) for intervals between markers along each of the 40 linkage groups (LG). ....	15
Figure S5: Proportion of transposable elements in interspersed repeats .....	16
Figure S6: Distribution of transposable elements according to their divergence from the consensus. ....	17
Figure S7: Alignment of the Normal Lake Whitefish genome to the Northern Pike genome using D-genies visualisation.....	18
Figure S8: Synteny between <i>C. clupeaformis</i> sp. Dwarf, <i>C. clupeaformis</i> sp. Normal, and <i>C. lavaretus</i> sp. Balchen.....	19
References in supplementary materials.....	19

Table S1: Statistics for the genome assemblies before anchoring into chromosomes.

<b>Whitefish genome assemblies</b>	<b>total length [bp]</b>	<b>number of contigs</b>	<b>longest contig [bp]</b>	<b>L50 [bp]</b>	<b>N50 [n]</b>	<b>L90 [bp]</b>	<b>N90 [n]</b>
Normal Final assembly	2682618941	7076	43030526	6096834	101	308820	870
Normal Whitefish_Flye08K	2712325911	8777	24349152	3219133	201	168561	1863
Normal Whitefish_Flye10K	2772957034	8762	25401145	2368812	257	166291	2143
Normal Whitefish_Flye15K	2785624403	9475	11362423	970569	707	131193	3746
Dwarf Whitefish Flye10k	2764848066	8433	21574413	2160360	274	159968	2272
Dwarf Whitefish Flye08k	2780381566	10128	21923556	2153278	291		
Dwarf Whitefish Flye12k	2807411528	9494	14914173	1671611	377		

Table S2: Correspondence between chromosomes in Coregonids

Correspondence between the reference genomes of *Coregonus clupeaformis* and *C. lavaretus* and linkage groups in genetic maps of *Coregonus clupeaformis*, *C. artedii*, and *C. lavaretus*.

Chromosome name in <i>Coregonus clupeaformis</i> genome (this study)	Chromosome name in <i>Coregonus lavaretus</i> genome (DeKayne et al, 2020)	Linkage group in <i>Coregonus lavaretus</i> genetic map (DeKayne & Feulner, 2018)	Linkage group in <i>Coregonus artedii</i> genetic map (Blumstein et al, 2020)	Linkage group in <i>Coregonus clupeaformis</i> previous genetic map (Gagnaire et al, 2013)	Linkage group in <i>Coregonus clupeaformis</i> new genetic map (this study)
Chr01	WFS01	1	6	5	4
Chr02	WFS02	29	30	16	30
Chr03	WFS03	30	21	29	31
Chr04	WFS04	3 (+36)	7	2 (+3)?	34
Chr05	WFS05	16 (+33)	23	28	3
Chr06	WFS06	33 (+16)	14	35	18
Chr07	WFS07	13 (+34)	5	12	15
Chr08	WFS08	9	4	24	2
Chr09	WFS09	6	20	13	14
Chr10	WFS10	5	24	38	5
Chr11	WFS11	15	11	11	11
Chr12	WFS12	12	22	21	16
Chr13	WFS13	7	19	8	19
Chr14	WFS14	22	26	26	1
Chr15	WFS15	2	1	4	12
Chr16	WFS16	25	17	34	21
Chr17	WFS17	20	8	10	10
Chr18	WFS18	31	32	37	25
Chr19	WFS19	4	13	30	6
Chr20	WFS20	8	3	6	7
Chr21	WFS21	19	25	25	24
Chr22	WFS22	38+39	NA	7	36
Chr23	WFS23	21	28	32	17
Chr24	WFS24	28	16	14	20
Chr25	WFS25	10	31	22	39
Chr26	WFS26	17	27	36	13
Chr27	WFS27	11	38	33	8
Chr28	WFS28	36 (+20)	2	1	29
Chr29	WFS29	18	10	31	9
Chr30	WFS30	14	33	27	28
Chr31	WFS31	35	18	15	26
Chr32	WFS32	37 (+2)	12	4	35
Chr33	WFS33	23	37	17	27
Chr34	WFS34	26	36	39	22
Chr35	WFS35	34 (+13)	9	NA/18?	23
Chr36	WFS36	24	29	40	38
Chr37	WFS37	27 (+1)	15	5	32
Chr38	WFS38	8	34	6	37
Chr39	WFS39	32	35	23	33
Chr40	NA	3 (+36)	7	2+3?	40



Table S3: Coordinates of putatively collapsed duplicated regions in the reference assembly of the Normal Lake Whitefish

Chr01	18000000	20500000
Chr01	103700000	121031747
Chr03	43640000	45080000
Chr04	32010000	34920000
Chr07	11410000	12870000
Chr07	81600000	82980000
Chr10	62190000	66060000
Chr13	1	3020000
Chr15	73710000	75020000
Chr18	26260000	27550000
Chr20	1800000	8400000
Chr22	1	7216770
Chr24	1	8220000
Chr28	20000000	49420000
Chr30	56520000	64630000
Chr32	1	11602415
Chr35	20840000	23000000
Chr38	1	8900000
Chr39	1	7110000

Table S4: Proportion of interspersed repeats in the Normal genome and in the sequences of the structural variants

	Normal reference genome			Sequences of SVs		
	number of sequences	length of sequences	% of the length	number of sequences	length of sequences	% of the length
Unmasked		1008956356	37.61%		49088355	21.63%
Retroelements	1635898	694575878	25.89%	162140	111518063	49.13%
DNA-transposons	1587932	629778518	23.48%	96759	42683508	18.81%
Rolling-circles	8595	1329769	0.05%	745	139058	0.06%
Unclassified TE	1505716	246978169	9.21%	64010	9506168	4.19%
Small RNA	121294	19679506	0.73%	8423	1613074	0.71%
Satellites	7644	1993471	0.07%	454	86506	0.04%
Simple repeats	727239	68691185	2.56%	110507	11415712	5.03%
Low complexity	84004	10696989	0.40%	5018	920708	0.41%

Table S5: Breakdown of the proportion of transposable elements for the main families in the Normal genome and in the sequences of the structural variants

	Normal reference genome			Sequences of SVs		
	number of sequences	length of sequences	% of the length	number of sequences	length of sequences	% of the length
Retroelements/SINEs	165982	19369894	0.72%	11103	1858884	0.82%
Retroelements/Penelope	16711	3023382	0.11%	573	89414	0.04%
Retroelements/LINES	752039	336040143	12.53%	48725	28036418	12.35%
Retroelements/CRE/SLACS	0	0	0%	0	0	0%
Retroelements/L2/CR1/Rex	527901	235787647	8.79%	35376	21505242	9.47%
Retroelements/R1/LOA/Jockey	22192	7475316	0.28%	1042	324057	0.14%
Retroelements/R2/R4/NeSL	816	538660	0.02%	56	10508	0%
Retroelements/RTE/Bov-B	67311	39504265	1.47%	5650	2598236	1.14%
Retroelements/L1/CIN4	27755	11837512	0.44%	1462	1461716	0.64%
Retroelements/LTR	717877	339165841	12.64%	102312	81622761	35.96%
Retroelements/BEL/Pao	1667	3089456	0.12%	282	693587	0.31%
Retroelements/Ty1/Copia	6651	1186106	0.04%	486	198262	0.09%
Retroelements/Gypsy/DIRS1	215716	177453473	6.61%	30682	51256082	22.58%
Retroelements/Retroviral	94119	37728202	1.41%	11140	5625760	2.48%
DNA-transposons/hobo-Activator	215774	54227588	2.02%	13741	5110005	2.25%
DNA-transposons/Tc1-IS630-Pogo	1156068	538733044	20.08%	67533	33021547	14.55%
DNA-transposons/En-Spm	0	0	0%	0	0	0%
DNA-transposons/MuDR-IS905	0	0	0%	0	0	0%
DNA-transposons/PiggyBac	10077	3142994	0.12%	546	253304	0.11%
DNA-transposons/Tourist/Harbinger	11323	3509455	0.13%	825	547015	0.24%
DNA-transposons/Other	719	201371	0.01%	25	4403	0%

Table S6: Portion of the genome masked by different families of TEs

	% of the Dwarf		% of the Normal	
	size in Dwarf	Genome	size in Normal	genome
5S-Deu-L2	4580718	0.17%	4139986	0.15%
CMC-EnSpm	8615241	0.31%	8447378	0.31%
Copia	1103875	0.04%	1057340	0.04%
Crypton-A	812641	0.03%	799915	0.03%
Crypton-V	4066459	0.15%	4562557	0.17%
DIRS	6594109	0.24%	6364078	0.23%
DNA	8085117	0.29%	7730840	0.28%
ERV	1787503	0.06%	1749186	0.06%
ERV1	37428917	1.35%	37122674	1.36%
Ginger-1	147473	0.01%	147412	0.01%
Gypsy	176462088	6.38%	171307131	6.26%
hAT	3200603	0.12%	3117699	0.11%
hAT-Ac	17428675	0.63%	17411111	0.64%
hAT-Blackjack	1192194	0.04%	1175813	0.04%
hAT-Charlie	16524837	0.60%	16193170	0.59%
hAT-hAT5	504449	0.02%	486359	0.02%
hAT-hAT6	678618	0.02%	678122	0.02%
hAT-Tip100	17333573	0.63%	16047021	0.59%
Helitron	1283837	0.05%	1252969	0.05%
I	8078027	0.29%	7744404	0.28%
IS3EU	793720	0.03%	720151	0.03%
Kolobok-E	650086	0.02%	608333	0.02%
Kolobok-T2	1096387	0.04%	1089204	0.04%
L1	11284387	0.41%	12025585	0.44%
L1-Tx1	39839302	1.44%	38773224	1.42%
L2	142551018	5.16%	138666129	5.07%
LINE	1046032	0.04%	1050364	0.04%
LTR	29955	0.00%	29014	0.00%
Maverick	1878546	0.07%	1795159	0.07%
Merlin	175903	0.01%	179830	0.01%
Ngaro	230312	0.01%	217572	0.01%
P	206746	0.01%	201657	0.01%
Pao	3125866	0.11%	3069278	0.11%
Penelope	2681214	0.10%	2856849	0.10%
PIF	137083	0.00%	128569	0.00%
PIF-Harbinger	3832933	0.14%	3529906	0.13%
PIF-ISL2EU	2010155	0.07%	2022879	0.07%
PiggyBac	3058264	0.11%	3141712	0.11%
R2-NeSL	431679	0.02%	537377	0.02%
Rex-Babar	97757592	3.54%	96050620	3.51%
RTE-BovB	154480	0.01%	150975	0.01%
RTE-X	41541679	1.50%	56048671	2.05%
SINE	2013974	0.07%	1956691	0.07%
SINE?	8581657	0.31%	9841653	0.36%
Sola-1	283780	0.01%	288855	0.01%
Sola-2	763024	0.03%	737761	0.03%
TcMar	456471	0.02%	410609	0.02%
TcMar-Fot1	2045407	0.07%	2882523	0.11%
TcMar-ISRm11	8253620	0.30%	7751729	0.28%
TcMar-Tc1	543460096	19.66%	533614040	19.50%
TcMar-Tc2	1031109	0.04%	1071410	0.04%
TcMar-Tigger	991499	0.04%	932645	0.03%
tRNA-Core-RTE	2423485	0.09%	2549640	0.09%
tRNA-Deu-RTE	2727426	0.10%	2667158	0.10%
Unknown	379770302	13.74%	387922260	14.18%
Zisupton	580639	0.02%	544959	0.02%



Table S7: Enrichment in SVs associated with transposable elements in outliers of differentiation between Dwarf and Normal Whitefish.

Element associated with the SV	Dataset for population-level analysis										Outliers in Indian Lake					Outliers in Cliff Lake					Overlap of outlier SVs across lakes				
	N	%	N	%	N	%	p	OR	q	N	%	p	OR	q	N	%	p	OR	q	N	%	p	OR	q	
DNA	93	0.1%	5	0.1%	5	0.1%	0.502	1.1	0.887	7	0.1%	0.200	1.5	0.619	-	-	1.000	0.0	1.000	-	-	1.000	0.0	1.000	
DNA/GMC-EnSpm	78	0.1%	6	0.1%	6	0.1%	0.211	1.5	0.453	3	0.1%	0.747	0.8	1.000	-	-	1.000	0.0	1.000	-	-	1.000	0.0	1.000	
DNA/Crypton-A	7	0.0%	-	-	1	0.0%	1.000	0.0	1.000	1	0.0%	0.323	2.9	0.762	-	-	1.000	0.0	1.000	-	-	1.000	0.0	1.000	
DNA/Crypton-V	106	0.1%	9	0.2%	9	0.2%	0.099	1.7	0.269	4	0.1%	0.774	0.8	1.000	2	0.3%	0.203	2.4	0.782	2	0.3%	0.203	2.4	0.782	
DNA/hAT	24	0.0%	2	0.0%	2	0.0%	0.353	1.7	0.662	1	0.0%	0.705	0.8	1.000	-	-	1.000	0.0	1.000	-	-	1.000	0.0	1.000	
DNA/hAT-Ac	484	0.5%	57	1.2%	57	1.2%	0.000	2.4	0.000	37	0.8%	0.011	1.5	7	12	1.7%	0.001	3.2	0.009	12	1.7%	0.001	3.2	0.009	
DNA/hAT-Blackjack	15	0.0%	-	-	-	-	1.000	0.0	1.000	2	0.0%	0.193	2.7	0.619	-	-	1.000	0.0	1.000	-	-	1.000	0.0	1.000	
DNA/hAT-Charlie	528	0.6%	49	1.0%	49	1.0%	0.000	1.9	0.001	50	1.1%	0.000	1.9	0	9	1.2%	0.026	2.2	0.192	9	1.2%	0.026	2.2	0.192	
DNA/hAT-hAT5	3	0.0%	-	-	-	-	1.000	0.0	1.000	-	-	1.000	0.0	1.000	-	-	1.000	0.0	1.000	-	-	1.000	0.0	1.000	
DNA/hAT-hAT6	5	0.0%	-	-	-	-	1.000	0.0	1.000	-	-	1.000	0.0	1.000	-	-	1.000	0.0	1.000	-	-	1.000	0.0	1.000	
DNA/hAT-Tip100	301	0.3%	29	0.6%	29	0.6%	0.001	1.9	0.007	19	0.4%	0.193	1.3	0.619	2	0.3%	0.678	0.9	1.000	2	0.3%	0.678	0.9	1.000	
DNA/IS3EU	3	0.0%	1	0.0%	1	0.0%	0.177	6.7	0.426	-	-	1.000	0.0	1.000	-	-	1.000	0.0	1.000	-	-	1.000	0.0	1.000	
DNA/Kolobok-E	8	0.0%	-	-	1	0.0%	1.000	0.0	1.000	1	0.0%	0.355	2.5	0.762	-	-	1.000	0.0	1.000	-	-	1.000	0.0	1.000	
DNA/Kolobok-T2	21	0.0%	1	0.0%	1	0.0%	0.658	1.0	1.000	1	0.0%	0.658	1.0	1.000	-	-	1.000	0.0	1.000	-	-	1.000	0.0	1.000	
DNA/Maverick	60	0.1%	5	0.1%	5	0.1%	0.197	1.7	0.453	2	0.0%	0.801	0.7	1.000	1	0.1%	0.376	2.1	1.000	1	0.1%	0.376	2.1	1.000	
DNA/PIF	1	0.0%	-	-	-	-	1.000	0.0	1.000	-	-	1.000	0.0	1.000	-	-	1.000	0.0	1.000	-	-	1.000	0.0	1.000	
DNA/PIF-Harbinger	57	0.1%	2	0.0%	2	0.0%	0.778	0.7	1.000	2	0.0%	0.778	0.7	1.000	-	-	1.000	0.0	1.000	-	-	1.000	0.0	1.000	
DNA/PIF-ISL2EU	69	0.1%	5	0.1%	5	0.1%	0.277	1.4	0.554	4	0.1%	0.461	1.2	0.937	-	-	1.000	0.0	1.000	-	-	1.000	0.0	1.000	
DNA/piggyBac	17	0.0%	3	0.1%	3	0.1%	0.067	3.5	0.237	3	0.1%	0.067	3.5	0.310	1	0.1%	0.130	7.6	0.709	1	0.1%	0.130	7.6	0.709	

DNA/Sola-1	3	0.0%	-	-	1.000	0.0	1.000	0.0	1.000	-	-	1.000	0.0	1.000
DNA/Sola-2	8	0.0%	2	0.0%	0.079	5.0	0.243	1	0.355	0.762	-	1.000	0.0	1.000
DNA/TcMar	1	0.0%	-	-	1.000	0.0	1.000	-	1.000	1.000	-	1.000	0.0	1.000
DNA/TcMar-Fot1	24	0.0%	3	0.1%	0.136	2.5	0.339	2	0.353	0.762	-	1.000	0.0	1.000
DNA/TcMar- ISrm11	285	0.3%	17	0.4%	0.275	1.2	0.554	18	0.200	0.619	-	1.000	0.0	1.000
DNA/TcMar-Tc1	11594	12.4%	<b>870</b>	<b>18.6%</b>	<b>0.000</b>	<b>1.5</b>	<b>0.000</b>	<b>876</b>	<b>18.7%</b>	<b>0.000</b>	<b>0</b>	<b>165</b>	<b>22.7%</b>	<b>0.000</b>
DNA/TcMar-Tc2	118	0.1%	11	0.2%	0.045	1.9	0.168	9	0.153	0.612	1	0.1%	0.601	1.1
DNA/TcMar- Tigger	55	0.1%	3	0.1%	0.526	1.1	0.902	1	0.935	1.000	-	1.000	0.0	1.000
DNA/Zisupton	2	0.0%	-	-	1.000	0.0	1.000	-	1.000	1.000	-	1.000	0.0	1.000
LINE	3	0.0%	-	-	1.000	0.0	1.000	-	1.000	1.000	-	1.000	0.0	1.000
LINE/I	64	0.1%	<b>9</b>	<b>0.2%</b>	<b>0.008</b>	<b>2.8</b>	<b>0.036</b>	<b>5</b>	<b>0.232</b>	<b>0.632</b>	-	<b>1.000</b>	<b>0.0</b>	<b>1.000</b>
LINE/L1	96	0.1%	<b>8</b>	<b>0.2%</b>	<b>0.124</b>	<b>1.7</b>	<b>0.322</b>	<b>14</b>	<b>0.3%</b>	<b>0.001</b>	<b>2</b>	<b>0.3%</b>	<b>0.175</b>	<b>2.7</b>
LINE/L1-Tx1	519	0.6%	34	0.7%	0.081	1.3	0.243	34	0.081	0.347	6	0.8%	0.222	1.5
LINE/L2	3076	3.3%	<b>231</b>	<b>4.9%</b>	<b>0.000</b>	<b>1.5</b>	<b>0.000</b>	<b>260</b>	<b>5.5%</b>	<b>0.000</b>	<b>0</b>	<b>37</b>	<b>5.1%</b>	<b>0.009</b>
LINE/Penelope	24	0.0%	1	0.0%	0.705	0.8	1.000	1	0.705	1.000	1	0.1%	0.176	5.4
LINE/R2-NeSL	7	0.0%	-	-	1.000	0.0	1.000	-	1.000	1.000	-	1.000	0.0	1.000
LINE/Rex-Babar	1821	1.9%	<b>172</b>	<b>3.7%</b>	<b>0.000</b>	<b>1.9</b>	<b>0.000</b>	<b>157</b>	<b>3.3%</b>	<b>0.000</b>	<b>0</b>	<b>21</b>	<b>2.9%</b>	<b>0.055</b>
LINE/RTE-BovB	2	0.0%	2	0.0%	0.013	20.0	0.055	-	1.000	1.000	-	1.000	0.0	1.000
LINE/RTE-X	1113	1.2%	28	0.6%	1.000	0.5	1.000	39	0.991	1.000	3	0.4%	0.991	0.3
Low_complexity	774	0.8%	11	0.2%	1.000	0.3	1.000	17	1.000	1.000	2	0.3%	0.982	0.3
LTR	4	0.0%	-	-	1.000	0.0	1.000	1	0.217	0.619	-	1.000	0.0	1.000
LTR/Copia	16	0.0%	2	0.0%	0.211	2.5	0.453	1	0.564	1.000	-	1.000	0.0	1.000
LTR/DIRS	91	0.1%	10	0.2%	0.022	2.2	0.090	5	0.484	0.937	-	1.000	0.0	1.000
LTR/ERV	21	0.0%	1	0.0%	0.658	1.0	1.000	1	0.658	1.000	-	1.000	0.0	1.000
LTR/ERV1	594	0.6%	<b>54</b>	<b>1.2%</b>	<b>0.000</b>	<b>1.8</b>	<b>0.000</b>	<b>49</b>	<b>1.0%</b>	<b>0.001</b>	<b>6</b>	<b>1.0%</b>	<b>0.186</b>	<b>1.5</b>

LTR/Gypsy	5543	5.9%	<b>384</b>	<b>8.2%</b>	<b>0.000</b>	<b>1.4</b>	<b>0.000</b>	<b>481</b>	<b>10.3%</b>	<b>0.000</b>	<b>1.7</b>	<b>0</b>	<b>61</b>	<b>8.4%</b>	<b>0.007</b>	<b>1.4</b>	<b>0.073</b>
LTR/Ngaro	2	0.0%	-	-	1.000	0.0	1.000	-	-	1.000	0.0	1.000	-	-	1.000	0.0	1.000
LTR/Pao	46	0.1%	5	0.1%	0.095	2.2	0.269	2	0.0%	0.673	0.9	1.000	1	0.1%	0.305	2.8	0.962
LTR/Unknown	4374	4.7%	<b>334</b>	<b>7.1%</b>	<b>0.000</b>	<b>1.5</b>	<b>0.000</b>	<b>343</b>	<b>7.3%</b>	<b>0.000</b>	<b>1.6</b>	<b>0</b>	<b>51</b>	<b>7.0%</b>	<b>0.005</b>	<b>1.5</b>	<b>0.057</b>
no_TE	30082	32.1%	1494	31.9%	0.594	1.0	0.963	1462	31.2%	0.827	1.0	1.000	213	29.3%	0.885	0.9	1.000
RC/Helitron	21	0.0%	-	-	1.000	0.0	1.000	-	-	1.000	0.0	1.000	-	-	1.000	0.0	1.000
rRNA	12	0.0%	1	0.0%	0.470	1.7	0.854	1	0.0%	0.470	1.7	0.937	-	-	1.000	0.0	1.000
Satellite	35	0.0%	1	0.0%	0.827	0.6	1.000	3	0.1%	0.271	1.7	0.706	1	0.1%	0.243	3.7	0.810
Simple_repeat	24142	25.8%	232	4.9%	1.000	0.2	1.000	181	3.9%	1.000	0.1	1.000	22	3.0%	1.000	0.1	1.000
SINE	22	0.0%	-	-	1.000	0.0	1.000	-	-	1.000	0.0	1.000	-	-	1.000	0.0	1.000
SINE/5S-Deu-L2	17	0.0%	1	0.0%	0.585	1.2	0.963	-	-	1.000	0.0	1.000	-	-	1.000	0.0	1.000
SINE/trNACore-RTE	53	0.1%	4	0.1%	0.288	1.5	0.557	4	0.1%	0.288	1.5	0.719	1	0.1%	0.341	2.4	1.000
SINE/trNADeu-RTE	30	0.0%	4	0.1%	0.077	2.7	0.243	3	0.1%	0.207	2.0	0.619	1	0.1%	0.213	4.3	0.782
SINE?	2132	2.3%	<b>198</b>	<b>4.2%</b>	<b>0.000</b>	<b>1.9</b>	<b>0.000</b>	<b>200</b>	<b>4.3%</b>	<b>0.000</b>	<b>1.9</b>	<b>0</b>	<b>37</b>	<b>5.1%</b>	<b>0.000</b>	<b>2.2</b>	<b>0.000</b>
trNA	2291	2.4%	<b>158</b>	<b>3.4%</b>	<b>0.000</b>	<b>1.4</b>	<b>0.001</b>	<b>191</b>	<b>4.1%</b>	<b>0.000</b>	<b>1.7</b>	<b>0</b>	<b>24</b>	<b>3.3%</b>	<b>0.096</b>	<b>1.4</b>	<b>0.575</b>
Unknown	2776	3.0%	<b>230</b>	<b>4.9%</b>	<b>0.000</b>	<b>1.7</b>	<b>0.000</b>	<b>190</b>	<b>4.1%</b>	<b>0.000</b>	<b>1.4</b>	<b>0</b>	<b>43</b>	<b>5.9%</b>	<b>0.000</b>	<b>2.0</b>	<b>0.001</b>

Table S9: Coordinates of previously identified QTL in the new reference genome

Regions including markers identified as significantly associated with one of the phenotypic traits of interest by (Gagnaire, Normandeau, Pavey, & Bernatchez, 2013; Rogers, Isabel, & Bernatchez, 2007)

CHR	start	stop	QTL
Chr02	24354669	24354670	Maturity
Chr05	30243484	44560907	Activity & Depth_selection
Chr05	47955978	47955979	Gonadosomatic index
Chr06	10944482	11510663	Growth rate
Chr07	13382789	13382790	Directional change
Chr07	45142150	49941487	Activity
Chr07	52999016	52999017	Growth rate
Chr08	85507277	85507278	Burst swimming & directional change
Chr10	53381623	60034589	Depth selection
Chr12	51988913	52856337	Depth selection
Chr13	10099482	63978158	Gill raker
Chr13	46005064	47196556	Depth selection
Chr15	23283752	56663130	Gill raker
Chr17	164510	68091188	Directional change
Chr20	27334774	27334775	Directional change
Chr21	1967998	45849371	sex
Chr23	11426032	54324241	Directional change
Chr23	28790621	38959612	Growth rate
Chr28	25985678	26432681	Directional change
Chr29	2195646	6155549	Gill raker
Chr31	8131059	12403428	Activity
Chr33	38762999	41235318	Gonadosomatic index
Chr34	30151992	35406463	Depth selection
Chr35	21065802	31659917	Directional change & Activity
Chr35	44842935	60738884	Burst swimming
Chr39	11171630	30309175	Depth selection
Chr40	1267440	11926388	Maturity



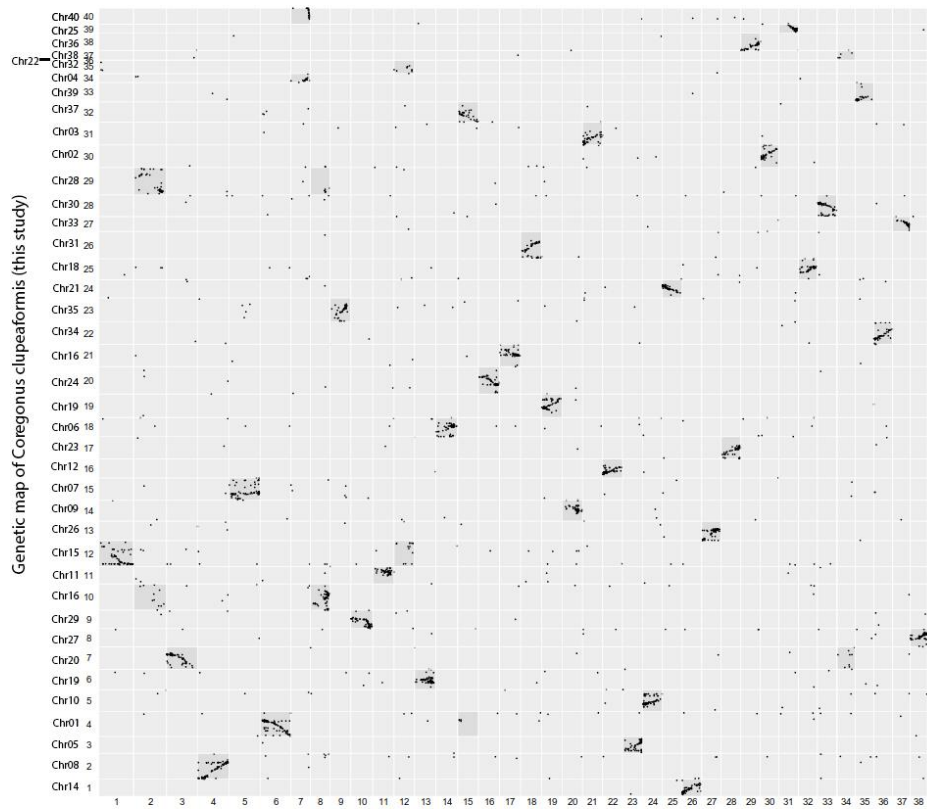


Figure S1: Homologous chromosomes with *Coregonus artedii* *Coregonus clupeaformis* compared with *Coregonus artedii* (Blumstein et al., 2020) with markers paired through *Coregonus clupeaformis* genome identified homology between chromosome arms with MapComp (Sutherland et al., 2016).

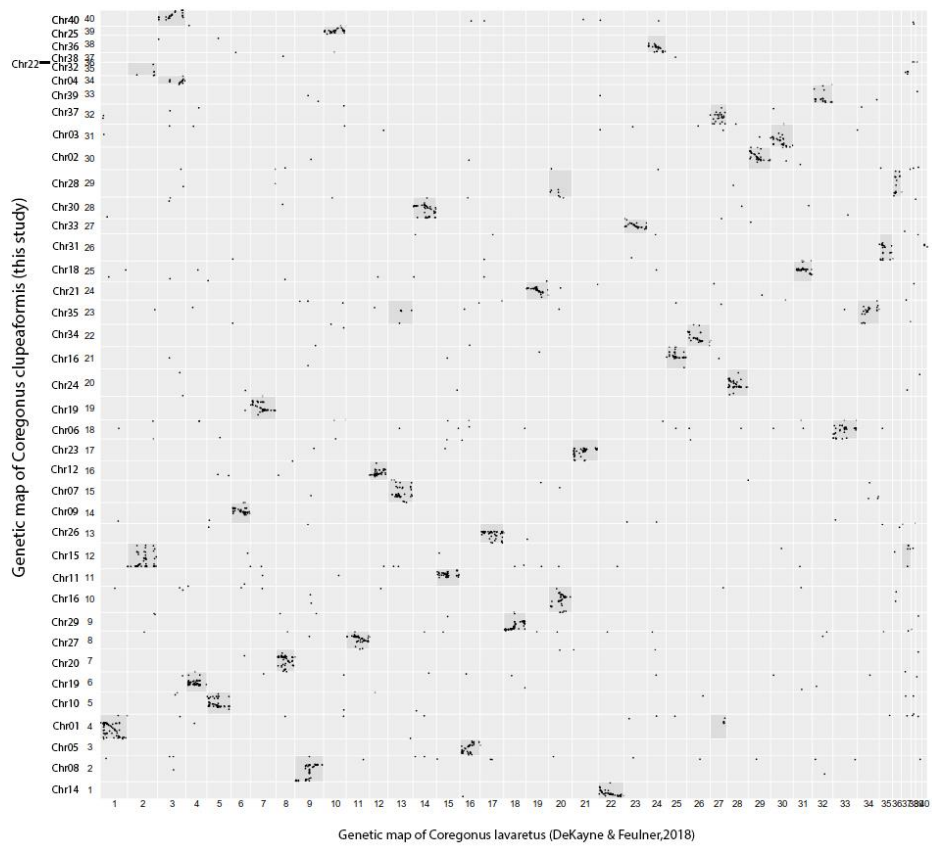
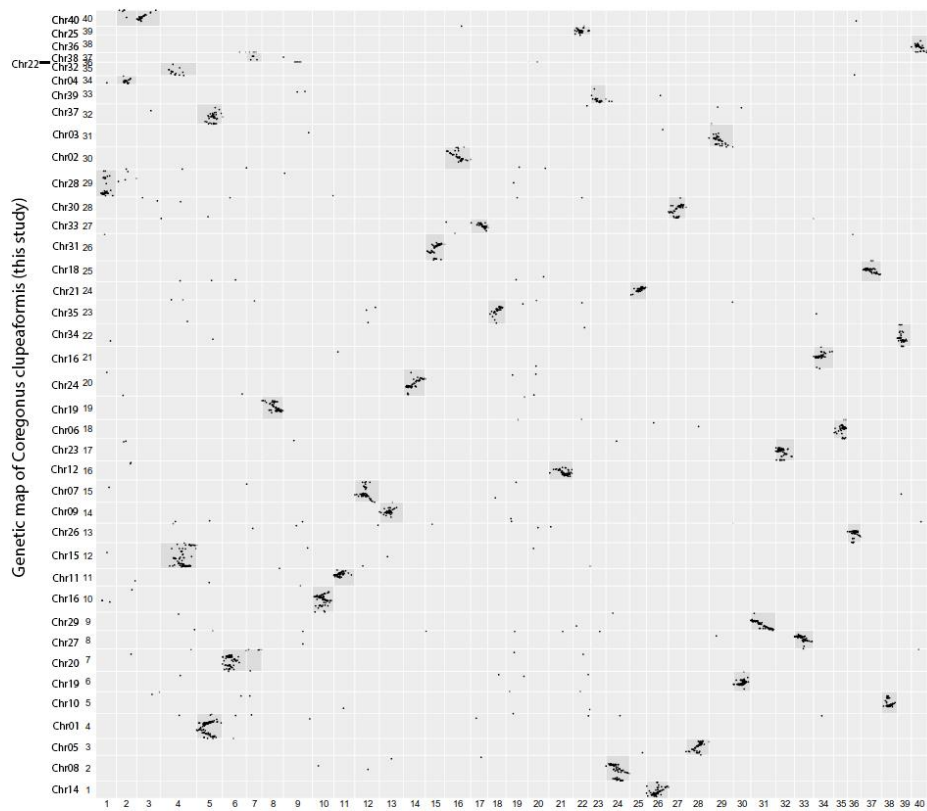


Figure S2: Homologous chromosomes with *Coregonus lavaretus* *Coregonus clupeaformis* compared with *Coregonus lavaretus* (De-Kayne & Feulner, 2018) with markers paired through *Coregonus clupeaformis* genome identified homology between chromosome arms with MapComp (Sutherland et al., 2016).



Genetic map of *Coregonus clupeaformis* (Gagnaire et al, 2013)

Figure S3: Homologous chromosomes with the previous map of *Coregonus clupeaformis* *Coregonus clupeaformis* compared with *Coregonus clupeaformis* (Gagnaire et al., 2013) with markers paired through *Coregonus clupeaformis* genome identified homology between chromosome arms with MapComp (Sutherland et al., 2016).

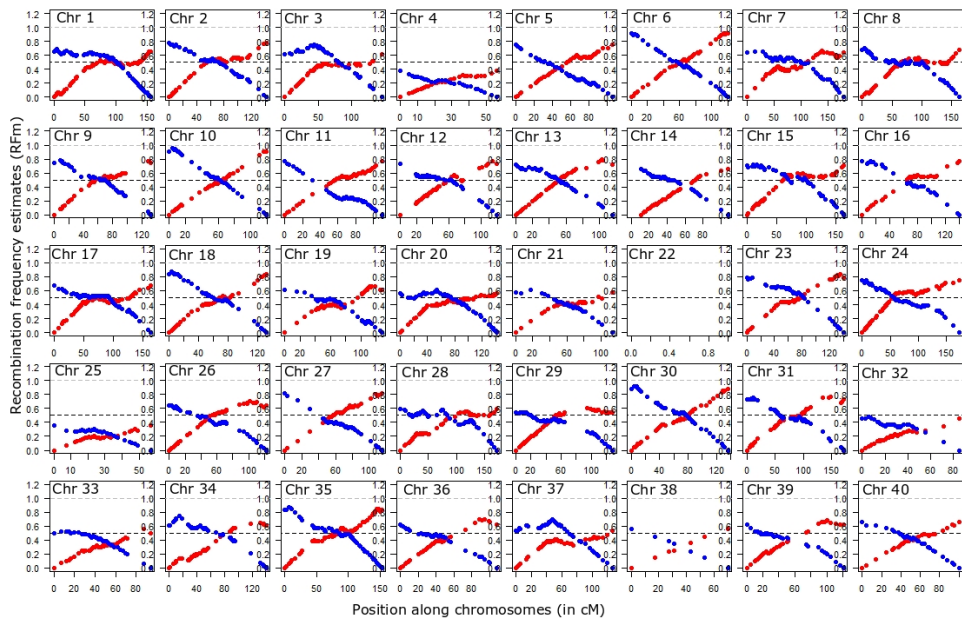


Figure S4: Recombination frequency estimates (RFm) for intervals between markers along each of the 40 linkage groups (LG).

LG are ordered by chromosome names from left to right, then from top to bottom). For each LG, RFm was calculated from both chromosomal extremities (right: red circles; left: blue circles), using each of the two terminal markers as a reference starting point. The RFm plot of Chr01 (top left) illustrates a classical metacentric pattern with a centromere position  $\sim 100\text{cM}$  while Chr05 (5th on the 1st line) illustrates a classical acrocentric pattern, the centromere position remains undetermined with regard to which LG extremity. See (Limborg, McKinney, Seeb, & Seeb, 2016) for methods.

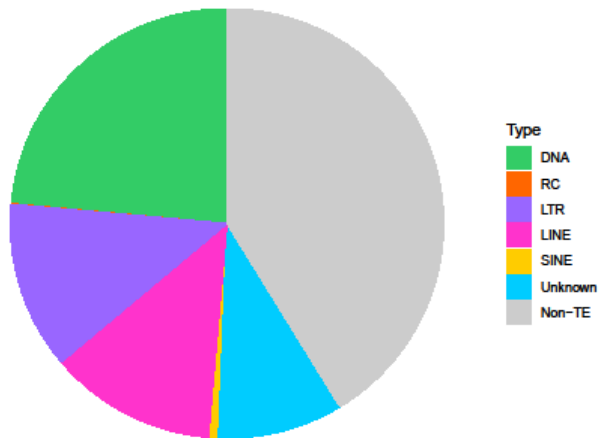
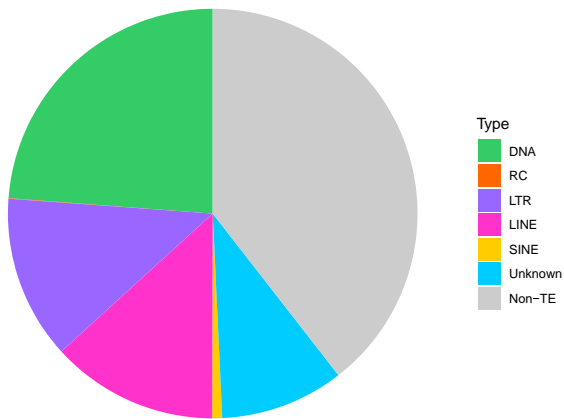


Figure S5: Proportion of transposable elements in interspersed repeats

Top: *Coregonus clupeaformis* sp. Normal (DNA=DNA-TIR 24%, RC-Helitron <1%, LTR 13%, LINEs 13%, SINEs <1%, Unknown=Unclassified TEs 9%, non-TE= interspersed repeats which are not transposable elements, 40%).

Bottom: *Coregonus clupeaformis* sp. Dwarf (DNA=DNA-TIR 23.5%, RC-Helitron <1%, LTR 13%, LINEs 13%, SINEs <1%, Unknown=Unclassified TEs 9%, non-TE= interspersed repeats which are not transposable elements, 40%).

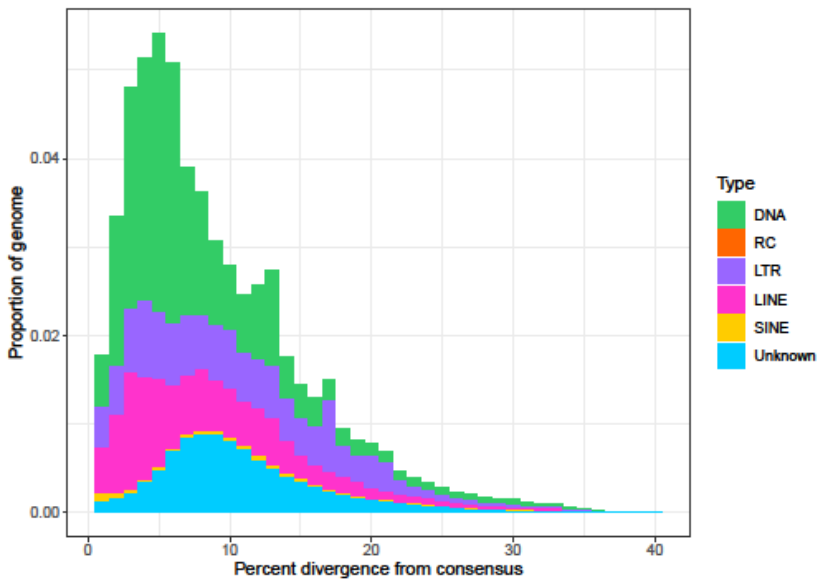
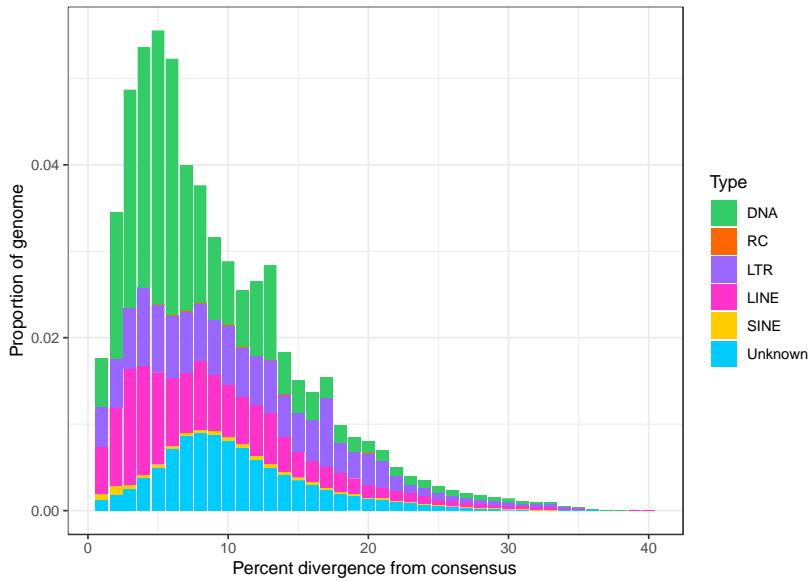


Figure S6: Distribution of transposable elements according to their divergence from the consensus.

Top: *Coregonus clupeaformis* sp. Normal. Bottom: *Coregonus clupeaformis* sp. Dwarf.

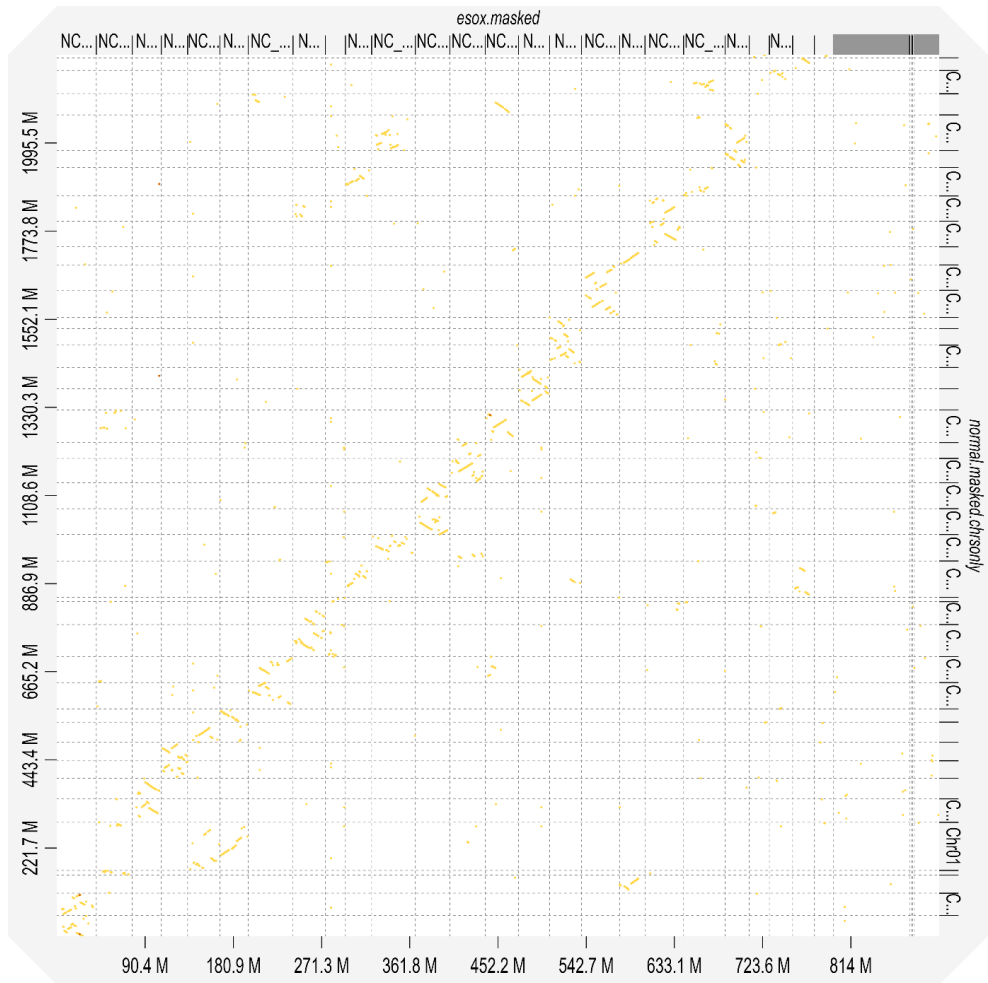


Figure S7: Alignment of the Normal Lake Whitefish genome to the Northern Pike genome using D-genies visualisation. (Cabanettes & Klopp, 2018).

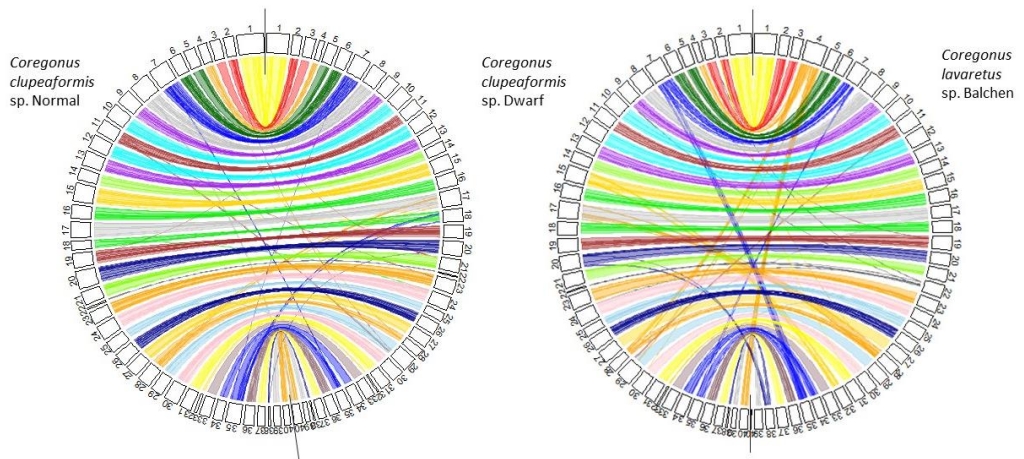


Figure S8: Synteny between *C. clupeaformis* sp. Dwarf, *C. clupeaformis* sp. Normal, and *C. lavaretus* sp. Balchen

Circular plots showing syntenic relationship between *C. clupeaformis* sp. Dwarf and *C. clupeaformis* sp. Normal (left) and *C. lavaretus* sp. Balchen (right). Note that The Dwarf genome was anchored into chromosomes using the same linkage map as the Normal genome (based on a hybrid family).

#### References in supplementary materials

Blumstein, D. M., Campbell, M. A., Hale, M. C., Sutherland, B. J., McKinney, G. J., Stott, W., & Larson,

W. A. (2020). Comparative genomic analyses and a novel linkage map for cisco (*Coregonus artedii*) provide insights into chromosomal evolution and rediploidization across salmonids.

*G3: Genes, Genomes, Genetics*, 10(8), 2863–2878.

Cabanettes, F., & Klopp, C. (2018). D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ*, 6, e4958.

De-Kayne, R., & Feulner, P. G. (2018). A European whitefish linkage map and its implications for understanding genome-wide synteny between salmonids following whole genome duplication. *G3: Genes, Genomes, Genetics*, 8(12), 3745–3755.

Gagnaire, P., Normandeau, E., Pavey, S. A., & Bernatchez, L. (2013). Mapping phenotypic, expression and transmission ratio distortion QTL using RAD markers in the Lake Whitefish (*Coregonus clupeaformis*). *Molecular Ecology*, 22(11), 3036–3048.



- Limborg, M. T., McKinney, G. J., Seeb, L. W., & Seeb, J. E. (2016). Recombination patterns reveal information about centromere location on linkage maps. *Molecular Ecology Resources*, *16*(3), 655–661.
- Rogers, S. M., Isabel, N., & Bernatchez, L. (2007). Linkage maps of the dwarf and normal lake whitefish (*Coregonus clupeaformis*) species complex and their hybrids reveal the genetic architecture of population divergence. *Genetics*, *175*(1), 375–398.
- Sutherland, B. J. G., Gosselin, T., Normandeau, E., Lamothe, M., Isabel, N., Audet, C., & Bernatchez, L. (2016). Salmonid Chromosome Evolution as Revealed by a Novel Method for Comparing RADseq Linkage Maps. *Genome Biology and Evolution*, *8*(12), 3600–3617. doi: 10.1093/gbe/evw262





ISBN: 978-82-575-2056-4

ISSN: 1894-6402



Norwegian University  
of Life Sciences

Postboks 5003  
NO-1432 Ås, Norway  
+47 67 23 00 00  
[www.nmbu.no](http://www.nmbu.no)