The emergence of supergenes from inversions in Atlantic salmon.

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

*contributed equally as senior and corresponding author

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

Abstract

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

Keywords

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

Introduction

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

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

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

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

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

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

Material and Methods

Nanopore long-read sequencing and building of genome assemblies

The Atlantic salmon reference genome (GCA_905237065.2) was built from 70x genome coverage with long-read Oxford Nanopore reads generated from a Norwegian aquaculture salmon (AQGE; Table S1). Long-read libraries were prepared using the SQK-LSK109 kit following the Genomic DNA by ligation protocol and sequenced on a PromethION sequencer. Initially, five *de novo* assemblies were generated with varying sequence overlaps (5, 10, 15, 20 and 30kb) using Flye v2.7 and v2.8 [18]. Contigs from the five assemblies were combined into one assembly by merging contig ends overlapping with >20kb or more determined from

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

Inversion detection

We detected inversions in the long-read sequenced samples with both read-mapping and assembly comparisons. For the read-based SV-calling pipeline, see electronic supplementary material. Custom scripts can be found at https://github.com/kristinastenlokk/long_read_SV. Assembly alignments were made with Minimap2 v2.23 [22] and were used to verify the candidate inversions from the read-based SV-detection pipeline, limited to inversions >= 10kb that are visible in assembly alignments, and to detect additional large inversions that read-based methods have low power to detect. This provided a set of 11 high confidence inversions (for details see Data S1). For inspection of repeat blocks in inversion breakpoints, we created self-alignments with LASTZ v1.0.4 [19], (Figure S1). These 11 inversions were genotyped by manual inspection of plots where contigs and nanopore reads were mapped to inversion breakpoints using Minimap2 v2.23 [22]. To validate heterozygous inversions in the reference AQGE, ultra-long reads were created with PromethION using the Ultra-Long DNA Sequencing Kit and protocol (SQK-ULK001, v.ULK_9124_v110_revA_24Mar2021). Figure S2B demonstrates how a AQGE ultra-long read mapped to the AQGE assembly reveals that both orientations of the inversion are present and resolving that AQGE is heterozygous for the inversion. Figure S2C shows mapping of a contig from the OULU assembly spanning the repeat structure of the upstream inversion breakpoint of chr9inv, validating the alternative state of the inversion in this sample.

Illumina short-read mapping and variant calling

For the short-read mapping and variant calling we used Illumina data from the whole genome re-sequencing of 482 Atlantic salmon sampled from a broad phylogeographic distribution [23] (Figure 1, Data S2). The Illumina reads were mapped to the Atlantic salmon genome (Ssal_v3.1; GCA_905237065.2) using the bcbio-nextgen v1.2.3 pipeline [24] with the bwa-mem aligner v.0.7.17 [25]. Aligned reads were sorted with Samtools v1.9 [26] and duplicate reads were marked with Sambamba v0.7.1 [27]. Genomic variation was identified using Google's DeepVariant pipeline v1.1.2 with default parameters [28] and the individual genotypes were merged using Glnexus v1.2.2 with the 'DeepVariantWGS' configuration [29]. Variants were then filtered for depth >4 and <40, genotype quality >10 followed by missingness <30% in vcftools (v0.1.16).

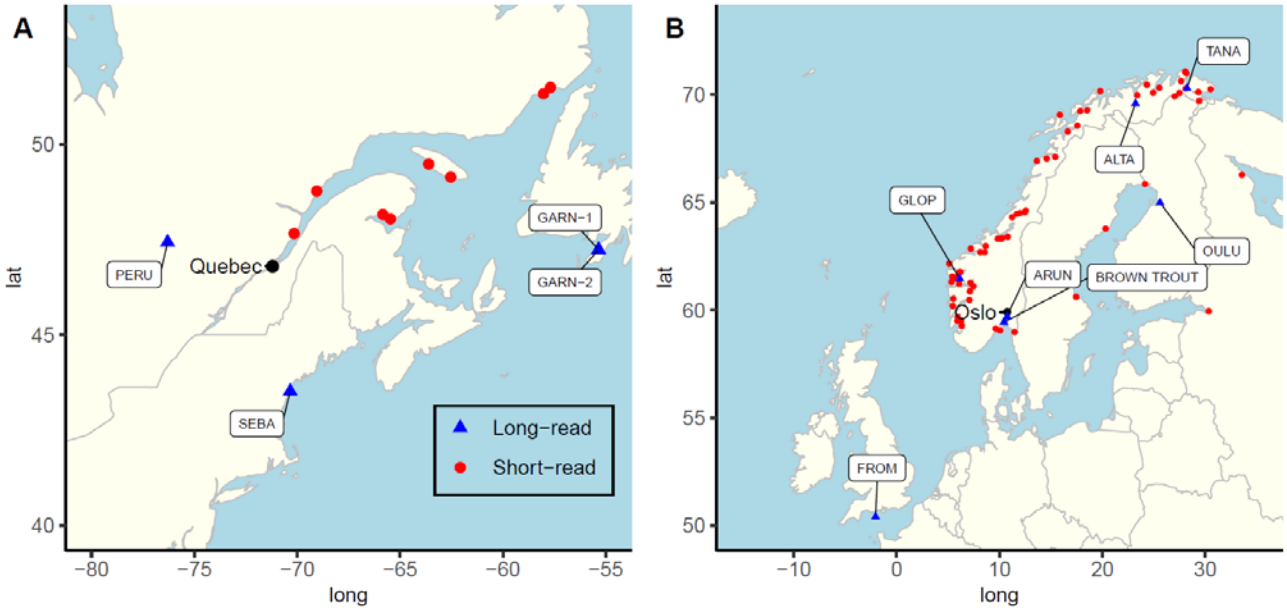

Figure 1: Map of sampling sites. Location of wild Atlantic salmon and brown trout samples used in study. Blue triangles designate nanopore long-read sequenced samples used for inversion detection and red dots indicate populations sampled and sequenced with Illumina short-read in North America **(A)**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

Identifying tag-SNPs and inversions type in population samples

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

Inversion dating

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

Genotype-Environment associations

To test if inversions were associated with adaptive variation, we tested for genotype-environment associations (GEA) determined using the Latent Factor Mixed Model (LFMM) approach [34] in the R package lea (R v4.1)[35]. 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 admixture (v1.23) [36]. Environment associations were tested on the pooled European (n=402) and the North American (n=80) samples separately because the strong differentiation between these lineages would confound associations and some inversions were only polymorphic in one group. Environment associations were tested for all SNPs on chromosomes containing an inversion (8 chromosomes, 1.07-1.23 million SNPs). False discovery control was employed using the Benjamini–Hochberg procedure with alpha thresholds of 0.05 and 0.01 across all tests. Variants were phased and imputed using Beagle v5.2 [37] (burnin 3, interactions 12, phase states 280) and then filtered for minor allele frequency >5%. Environment variables tested related to thermal, precipitation and river size conditions in the spawning and juvenile habitat expected to exert selection pressures on salmon. The individual river parameters were obtained from the WorldClim database for an arc of 30 translating to 1 square km 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 [38], 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 these influence the overwinter survival and growth potential respectively. Inversions were inferred to have adaptive potential where they overlap with multiple variant associations suggesting that the inversion has the potential to link different adaptive variants and is capable of becoming a supergene. The frequency of associated loci was calculated for inversion homozygotes, to avoid any influence of phasing errors, by summing the allele count and dividing by twice the number of homozygous individuals for each arrangement.

Mutation load

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

[40]) were computed to assess the impact of the detected missense variants for each protein using the Ensembl Rapid Release annotation of GCA_905237065.2. PROVEAN scores >= |2.5| were defined as deleterious. We compared the density of deleterious mutations within inversions to the genome wide level by dividing the number of significant PROVEAN scores by the number of genes per megabase (Mb) to obtain the mutation load per gene and Mb. We used the Wilcoxon Rank Sum test to test for significant enrichment inside inversions.

Detection of indels within inversions

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

Results and Discussion

Detection and characterization of inversions

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

Further, alignment of chromosome sequences in the Atlantic salmon reference (AQGE; GCA_905237065.2) with syntenic regions in the sister species brown trout, shows that for chr4inv, chr11inv3, chr16inv, chr18inv, chr22inv and chr26inv the reference has the standard configuration, whereas for chr3inv, chr9inv, chr10inv, chr11inv1 and chr11inv2 it has the inverted orientation (Figure S4).

Characterization of inversion breakpoints

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

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

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

Accumulation of deleterious mutations and indel enrichment

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

Haplotype structure within inversions

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

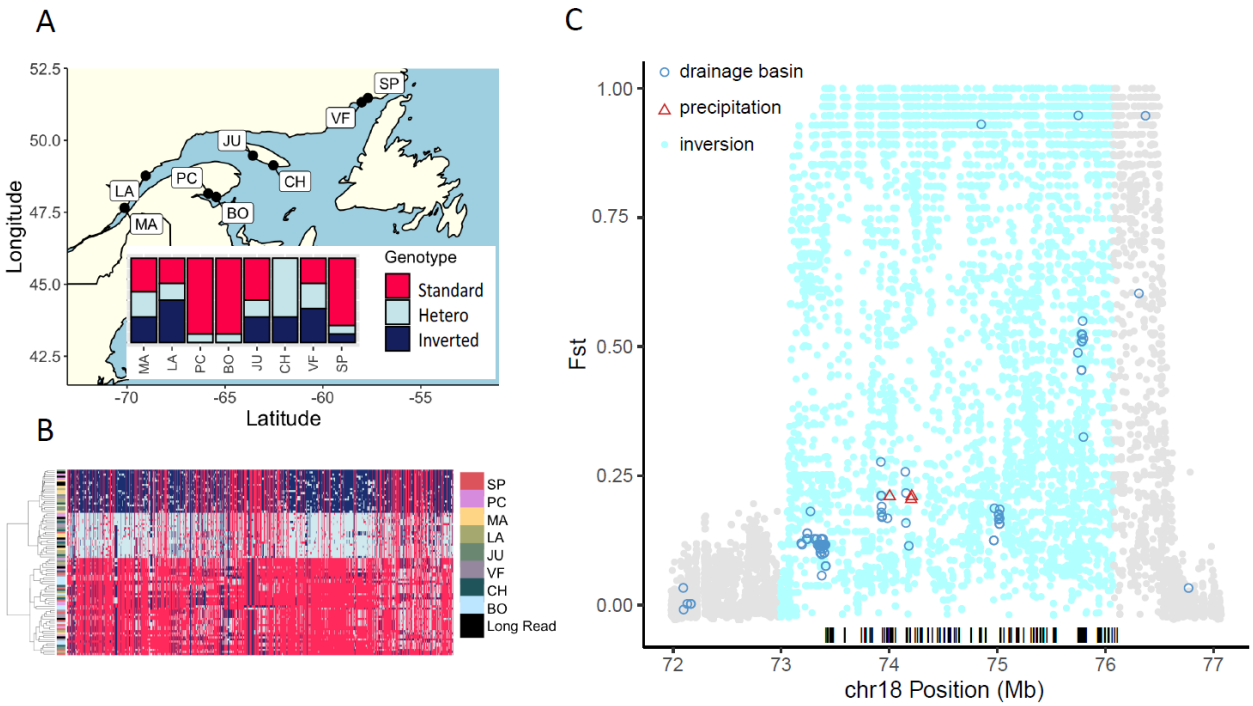

Figure 2. Chr18inv haplotype structure and F_{ST} between inverted and standard arrangements. A. Location of samples and estimated chr18inv haplotype frequencies in different Canadian rivers based on short-read sequence data. BO: Bonaventure, CH, De_la_Chaloupe, JU: Jupiter, LA: Laval, MA: Malbaie_(Charlevoix), PC: Petite_riviere_Cascapedia, SP: Saint-Paul, VF, Du_Vieux_Fort. B. Red: reference homozygous type, Light blue: heterozygous type, Navy blue: inversion homozygous type. **B**. Haplotype structure within the chr18inv region (The 1000th to the 2000th variants were selected from the 4828 variants to reduce the computational load for the effective visualization) based on short-read sequence data in North American populations. Individuals with long-read 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 and Cockerham) between homozygotes with alternative inversion orientations showing environment associated SNPs with p<0.05, black bars under plot show the positions of tag-SNPs used to genotype the inversion.

Dating of the inversions

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

Genotype-environment associations

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

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

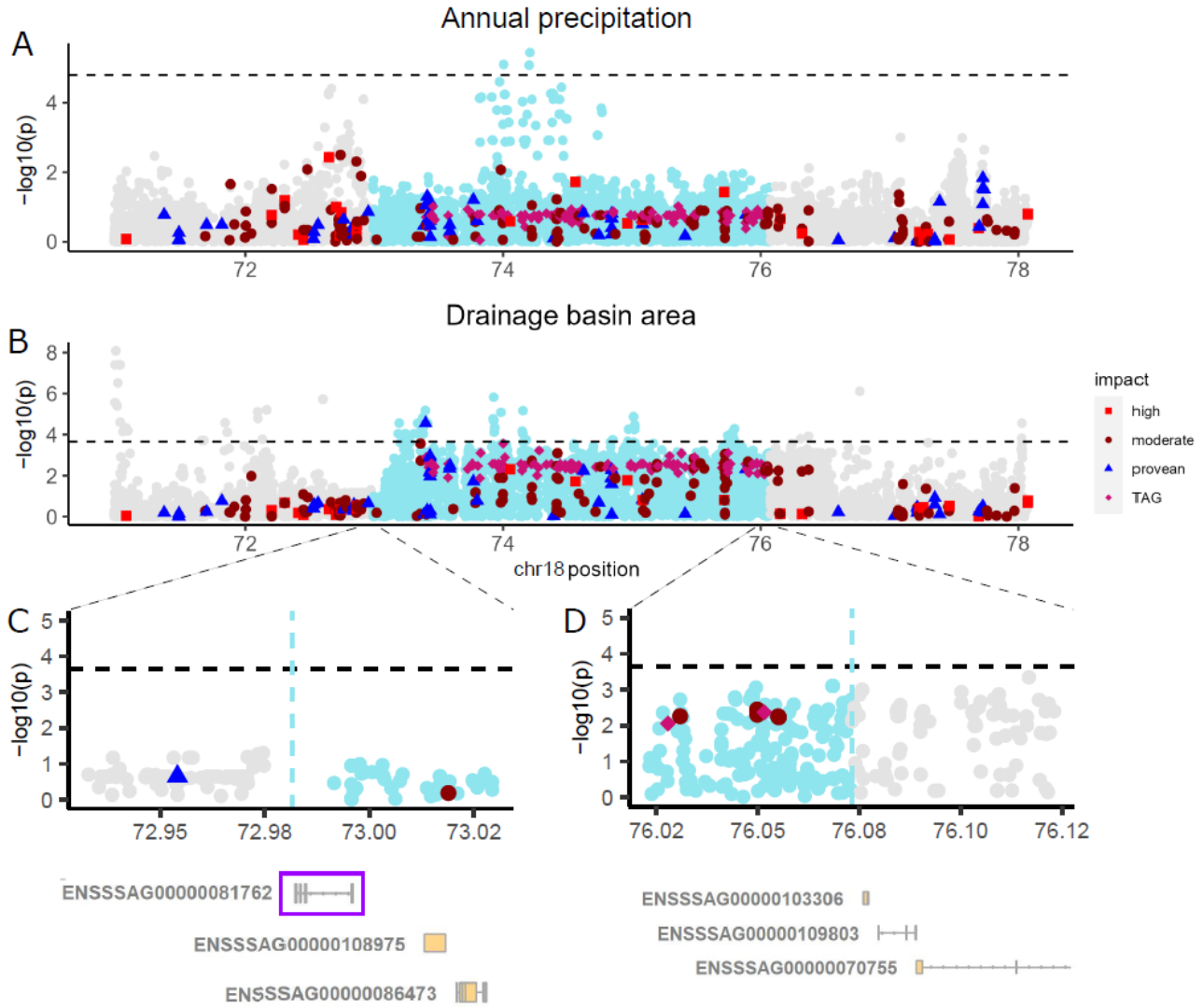

Figure 3. **Genotype-environment associations for chr18inv**. Associations with **A** annual precipitation and **B** drainage basin area, functional- and tag-SNPs are highlighted for the inversion (blue) and 2Mb flanking (grey). Significant associations, dashed horizontal line, indicates significance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

Capture and accumulation of adaptive variation

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

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

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

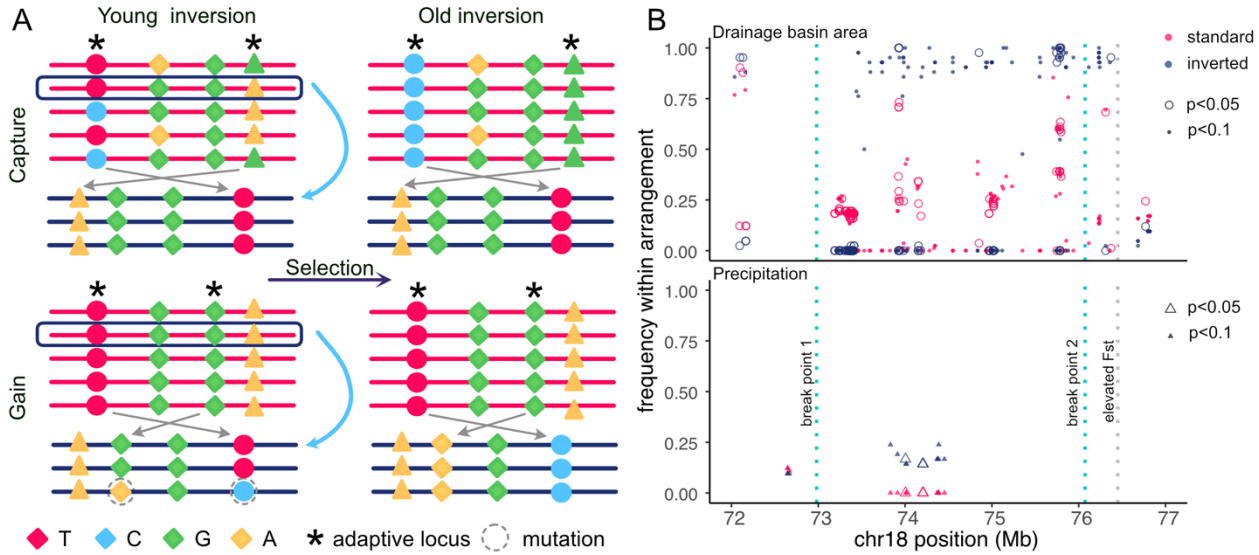

Figure 4. Capture and accumulation of adaptive variation by inversions. A. When an inversion first occurs the inverted arrangement captures a single invariant haplotype, while the standard arrangement will still have multiple variable haplotypes including pre-existing adaptive variation (upper left). 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.

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

supergene in rainbow trout. Nature Ecology & Evolution 3, 1731-1742. (DOI:https://doi.org/10.1038/s41559-019-1044-6).
[4] Jay, P., Chouteau, M., Whibley, A., Bastide, H., Parrinello, H., Llaurens, V. & Joron, M. 2021 Mutation load at a mimicry supergene sheds new light on the evolution of inversion polymorphisms. Nature Genetics 53, 288-293. (DOI:https://doi.org/10.1038/s41588-020-00771-1).
[5] Lamichhaney, S., Fan, G., Widemo, F., Gunnarsson, U., Thalmann, D. S., Hoeppner, M. P., Kerje, S., Gustafson, U., Shi, C., Zhang, H., et al. 2016 Structural genomic changes underlie alternative reproductive strategies in the ruff (Philomachus pugnax). Nature Genetics 48, 84-88. (DOI:https://doi.org/10.1038/ng.3430).
[6] Lee, C.-R., Wang, B., Mojica, J. P., Mandáková, T., Prasad, K. V. S. K., Goicoechea, J. L., Perera, N., Hellsten, U., Hundley, H. N., Johnson, J., et al. 2017 Young inversion with multiple linked QTLs under selection in a hybrid zone. Nature Ecology & Evolution 1, 0119. (DOI:https://doi.org/10.1038/s41559-017-0119).
[7] Villoutreix, R., Ayala, D., Joron, M., Gompert, Z., Feder, J. L. & Nosil, P. J. M. E. 2021 Inversion breakpoints and the evolution of supergenes. Molecular Ecology. (DOI:https://doi.org/10.1111/mec.15907).
[8] Küpper, C., Stocks, M., Risse, J. E., dos Remedios, N., Farrell, L. L., McRae, S. B., Morgan, T. C., Karlionova, N., Pinchuk, P., Verkuil, Y. I., et al. 2016 A supergene determines highly divergent male reproductive morphs in the ruff. Nature Genetics 48, 79-83. (DOI:https://doi.org/10.1038/ng.3443).
[9] Giraldo-Deck, L. M., Loveland, J. L., Goymann, W., Tschirren, B., Burke, T., Kempenaers, B., Lank, D. B. & Küpper, C. 2022 Intralocus conflicts associated with a supergene. Nature Communications 13, 1384. (DOI:https://doi.org/10.1038/s41467-022-29033-w).
[10] Faria, R., Johannesson, K., Butlin, R. K. & Westram, A. M. 2019 Evolving Inversions. Trends in Ecology & Evolution 34, 239-248. (DOI:https://doi.org/10.1016/j.tree.2018.12.005).
[11] Kirkpatrick, M. & Barton, N. 2006 Chromosome Inversions, Local Adaptation and Speciation. Genetics 173, 419-434. (DOI:https://doi.org/10.1534/genetics.117.300572).
[12] Charlesworth, B. & Barton, N. H. 2018 The Spread of an Inversion with Migration and Selection. Genetics 208, 377-382. (DOI:https://doi.org/10.1534/genetics.117.300426).
[13] Jay P, Leroy M, Le Poul Y, Whibley A, Arias M, Chouteau M, Joron M. 2022 Association mapping of colour variations in a butterfly provides evidences that a supergene locks together a cluster of adaptive loci. Phil. Trans. R. Soc. B 377, 20210193. (doi:10.1098/rstb.2021.0193)
[14] Hutchings, J. A. & Jones, M. E. B. 1998 Life history variation and growth rate thresholds for maturity in Atlantic salmon, Salmo salar. Canadian Journal of Fisheries and Aquatic Sciences 55, 22-47. (DOI:https://doi.org/10.1139/d98-004).
[15] 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, 405-408. (DOI:http://dx.doi.org/10.5061/dryad.23h4q).
[16] 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, 200-205. (DOI:https://doi.org/10.1038/nature17164).
[17] 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, 532-551. (DOI:https://doi.org/10.1111/mec.12003).
[18] Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. J. N. b. 2019 Assembly of long, error-prone reads using repeat graphs. Nature biotechnology 37, 540-546. (DOI:https://doi.org/10.1038/s41587-019-0072-8).
[19] Harris, R. S. 2007 Improved pairwise alignment of genomic DNA, The Pennsylvania State University.
[20] Shafin, K., Pesout, T., Chang, P.-C., Nattestad, M., Kolesnikov, A., Goel, S., Baid, G., Kolmogorov, M., Eizenga, J. M. & Miga, K. H. J. N. m. 2021 Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. Nature methods 18, 1322-1332. (DOI:https://doi.org/10.1038/s41592-021-01299-w).
[21] Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., 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., Røsæg, 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-κB 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, Árnyasi 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