Norwegian University of Life Sciences
Faculty of Biosciences
Department of Animal and Aquaculture sciences

Philosophiae Doctor (PhD)
Thesis 2026:20

Structural variants and whole genome duplication in Atlantic salmon evolution

Strukturelle varianter og helgenomduplisering i atlanterhavslaksens evolusjon

Célian Diblasi

Structural variants and whole genome duplication in Atlantic salmon evolution

Strukturelle varianter og helgenomduplisering i atlanterhavslaksens evolusjon

Philosophiae Doctor (PhD) Thesis

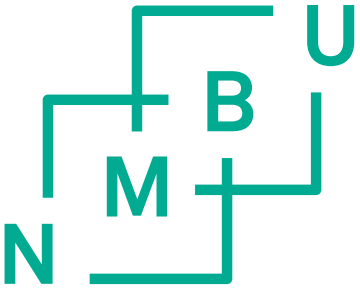
Célian Diblasi

Norwegian University of Life Sciences

Faculty of Biosciences (Biovit)

Department of Animal and Aquaculture sciences

Ås 2026



Thesis number 2026:20

ISSN 1894-6402

ISBN 978-82-575-2335-0

Supervisors and evaluation committee

PhD supervision team

Dr. Marie Saitou
Faculty of Biosciences
Department of Animal
And Aquacultural sciences, CIGENE,
Norwegian University of Life Sciences,
P.0. Box 5003 NMBU,
1432 Ås, Norway.
marie.saitou@nmbu.no

Prof. Simen Rød Sandve
Faculty of Biosciences
Department of Animal
And Aquacultural sciences, CIGENE,
Norwegian University of Life Sciences,
P.0. Box 5003 NMBU,
1432 Ås, Norway.
simen.sandve@nmbu.no

Dr. Nicola Jane Barson
Faculty of Biosciences
Department of Animal
And Aquacultural sciences, CIGENE,
Norwegian University of Life Sciences,
P.0. Box 5003 NMBU,
1432 Ås, Norway.
nicola.barson@nmbu.no

PhD Evaluation committee

Dr. Camille Berthelot
Pasteur Institute
Comparative Functional Genomics
25-28 Rue du Dr Roux,
75015 Paris
camille.berthelot@pasteur.fr

Dr. Tuomas Hämälä
Natural Resources Institute Finland
Tietotie 4
FI-31600, Jokioinen
Finland
tuomas.hamala@luke.fi

Prof. Dag Inge Våge
Faculty of Biosciences
Department of Animal
And Aquacultural sciences, CIGENE,
Norwegian University of Life Sciences,
P.0. Box 5003 NMBU,
1432 Ås, Norway.
daginge.vage@nmbu.no

Acknowledgments

I would first like to thank my main supervisor, Marie Saitou. You gave me the opportunity to work with you, and I hope you enjoyed it as I did. I am really glad to have had you as a supervisor, you were always here when I needed your help, stimulating my curiosity with relevant questions and ideas, and listening to my ideas. You also gave me many opportunities aligned with my interests, for instance teaching, for which I am thankful. You also provided a very nice working environment, making sure everyone feels comfortable, even sometimes more concerned about my wellness and mental health than I was.

I also want to thank my co-supervisors Nicolas Barson and Simen Sandve. You gave me many comments, ideas, and helped me to develop skills, and you had a great impact on me and my work. Nicky, I particularly liked the times when I had a small question to ask you, which often turned into a long discussion and brought me many new perspectives. Simen, I appreciated your honest comments and feedback, either on the scientific results or on the ways to visualize them, which were very helpful to bring more quality to my work.

I also want to sincerely thank the MSlab group, Jun and Domniki, as well as all the interns that worked with us. It has been a motivating environment working with you all.

Of course, this journey wouldn't have been the same without the people from Cigene around, thank you very much for your kindness, your feedback, your help when I had questions, and most importantly your creativity during lunch breaks to come up with far-fetched questions and discussions.

Since this work is the culmination of all my studies, I also want to have a word for all my former supervisors during my previous projects as an intern. You all nourished my passion for research, and you taught me a lot. If I have made it that far, it's in part thanks to you, and I am certain that, one way or another, some of your teachings are there in my PhD work.

These three years would not have been the same with the support of my friends, with our discussions, meaningful or not, with our travels and all the time we shared. I want to give a particular thanks to all my friends who listened to me for my various rehearsals and gave me much useful feedback, so thank you Côme, Pauline, Alice, Manel, Caroline, Célia.

I think I managed to maintain good mental health through this intense journey, and for that I owe a lot to my "chaos group", so thank you Alice, Côme, Thomas, Manon, Lorn, Tom, for your presence and for giving me opportunities to crack jokes in French in order not to lose my sense of humor.

Additionally, I want to thank Pauline, Clémentine, Avnik, Frederick, Disha, Amritha, whom I crossed paths with in Norway, and with whom I enjoyed this country.

I finally want to give special thanks to my family, who have been here for me since the beginning. I was attracted very early by research, and it's been a long way since, and you always supported me, even if my stories about genomes and evolution may have been difficult to follow.

Table of content

1 List of papers

Paper 1

Célian Diblasi, Domniki Manousi, David Hazlerigg, Lars Grønvold, Nicola Jane Barson, Simen Rød Sandve, Marie Saitou. **Large-scale eQTL analyses in Atlantic salmon reveal persistent dosage compensation 100 million years after genome duplication**. Submitted to Sciences Advances.

Paper 2

Pauline Buso, Célian Diblasi, Domniki Manousi, Jun Soung Kwak, Arturo Vera-Ponce de Leon, Kristina Stenløkk, Nicola Jane Barson, Marie Saitou (2025). **Parallel Selection in Domesticated Atlantic Salmon from Divergent Founders Including on Whole-Genome Duplication-derived Homeologous Regions**. Genome Biology and Evolution, Volume 17, Issue 4, April 2025, evaf063.
https://doi.org/10.1093/gbe/evaf063

Paper 3

Célian Diblasi, Nicola Jane Barson, Marie Saitou. **Recurrent emergence and ancient persistence of structural variants across continents in Atlantic salmon**. Manuscript.

2 Abstract

One of the central premises of evolution is the presence of phenotypic variation, principally caused by genetic mutations. Mutations occur in different forms, ranging from changes in a single base pair to more complex mutations that affect larger genomic regions. This is for instance the case of Structural variants (SVs), such as deletions, insertions, or inversions of a sequence. An even more extreme mutational event is whole genome duplication (WGD), in which the entire genome is doubled. Such complex variants are difficult to study, and their evolutionary significance remains partially answered. Atlantic salmon (*Salmo salar*) constitute an interesting model for studying the importance of complex variants in evolutionary processes. Salmonids experienced an autopolyploidization event 100 million years ago, and present-day salmonids harbor many SVs and repeated sequences that vary among wild and domesticated populations.

By doubling the genetic material, a WGD event also doubles the regulatory genes and their associated targets. This doubling then causes many challenges in terms of evolution of gene regulation by affecting the proper dosage balance of some genes. To investigate the evolution of duplicated gene regulation and expression, we analyzed a large eQTL dataset. Among the 230 duplicated gene pairs we found to share an eQTL, 16 of them showed compensatory gene expression, indicating a dosage compensation effect which has been maintained over 100 million years of evolution. We also found that trans-regulatory connections were more likely to appear within chromosomes after the WGD event, which may be linked to constraints imposed by 3D chromatin architecture.

The multiple gene copies generated through whole genome duplications can also facilitate adaptation, for instance during selection and subsequent selective breeding. Since Atlantic salmon has been independently domesticated in Europe and North America, we were able to use salmon as a model system to explore general effects of selection during domestication and breeding on genetic variation in duplicated genes. We performed a genome wide selection scan using several metrics (Tajima'D, FST, XP-

EHH) and identified 18 genes under selection, most of them being associated with immunity. We noticed that 2 genes (*daxx*, *tapbp*) have been selected in parallel in both domesticated lineages. For 3 other genes (*brd2a*, *psmb8*, *psmb13*), one gene copy was selected on one continent, while its WGD-derived duplicated copy was selected in the other.

We also investigated the evolutionary origins of structural variants segregating in both wild and domesticated populations in Europe and North America. We applied two methods to detect SVs which we combined with population genomics analysis and a simulation-based approach. We found that a majority of shared SVs were ancient SVs maintained by balancing selection since the split between European and Norh American Atlantic salmon. We also found some SVs occurring repeatedly at the same location, thus being confounded as a single SV. We also identified an immune-gene rich region on chromosome 3, showing signs of selection in domesticated salmon.

Overall, these results provide a better understanding of how complex variants shape phenotypic diversity in Atlantic salmon. This thesis highlights the role of genetic redundancy in the maintenance of important functions and their opportunity to be selected in different environments. It also provides elements to evaluate the possible evolutionary trajectories of SVs. Finally, this thesis identifies candidate immune genes under selection in domesticated Atlantic salmon and improves our comprehension of the role of complex variants in domestication.

3 Norsk sammendrag

Et av de sentrale premissene for evolusjon er tilstedeværelsen av fenotypisk variasjon, hovedsakelig forårsaket av genetiske mutasjoner. Mutasjoner kommer i mange ulike former, fra enkle endringer i ett enkelt basepar til mer komplekse mutasjoner som påvirker store genomiske regioner. Eksempler på mer komplekse mutasjonstyper er strukturelle varianter (SV-er), slik som delesjoner, inversjoner eller inversjoner av en sekvens. En enda mer ekstrem mutasjonshendelse er helgenomduplisering, der hele genomet dobles. Slike komplekse genomiske mutasjoner er vanskelige å studere, og betydningen av komplekse genetiske varianter for evolusjon er derfor også mangelfull. I denne konteksten utgjør Atlanterhavslaks (*Salmo salar*) en interessant modell for å studere betydningen av komplekse varianter for evolusjonære prosesser. Laksefisk har gjennomgått en autopolyploidiseringshendelse for omtrent 100 millioner år siden og dagens laksefisk har et stort antall strukturelle mutasjoner som varierer mellom ulike domestiserte- og ville populasjoner.

Genomduplikasjoner er kjent for å føre til evolusjon av nye genregulatoriske nettverk, bla ved selesksjon for ny gendosebalanse. For å undersøke evolusjonen av transregulatorer etter en genomduplikasjon analyserte vi et omfattende eQTL-datasett fra gjellevev i laks. Blant de 230 dupliserte genparene som delte en eQTL, viste 16 av dem kompenserende genuttrykk, noe som indikerer en gendosekompensasjon som har blitt opprettholdt gjennom 100 millioner år. Vi fant også at transregulatoriske forbindelser oftere oppstod innenfor kromosomer etter genomduplikasjonen, noe som kan være knyttet til 3D-kromatinarkitektur.

De mange genkopiene som skapes gjennom genomduplikasjoner kan også føre til nye tilpasninger, for eksempel under domestiseringsprosessen og videre avl. Siden atlanterhavslaks har blitt domestisert uavhengig i Europa og Nord-Amerika, kunne vi bruke laks som modellsystem for å utforske generelle effekter av seleksjon under

domestisering og avl på genetisk variasjon i dupliserte gener. Vi gjennomførte genomvide søk etter signaturer for seleksjon med ulike metoder (Tajima's D, FST, XP-EHH) og identifiserte 18 gener, hvorav de fleste er assosiert med immunfunksjoner. Vi observerte at to gener (*daxx* og *tapbp*) har blitt selektert parallelt i begge domestiserte linjer. For tre andre gener (*brd2a*, *psmb8* og *psmb13*) var én kopi av et duplikatpar under seleksjon på det ene kontinentet, mens den andre kopien var under seleksjon på det andre kontinentet.

Vi studerte også den evolusjonære opprinnelsen til strukturelle varianter som segregerte i både ville og domestiserte populasjoner i Europa og Nord-Amerika. Vi benyttet to metoder for å detektere strukturelle varianter, som vi kombinerte med populasjonsgenomiske analyser og en simuleringsbasert tilnærming. Vi fant at majoriteten av de strukturelle variantene var gamle og hadde blitt bevart av balanserende seleksjon siden før splitten mellom den Nord-Amerikanske og Europeiske Atlantiske laksen. Vi identifiserte også enkelte SV-er som har oppstått gjentatte ganger på samme genomiske lokasjon, og som derfor kan feiltolkes som én enkelt SV. I tillegg identifiserte vi en region rik på immunrelaterte gener på kromosom 3, som viser tegn til seleksjon i domestisert laks.

Samlet sett gir disse resultatene en bedre forståelse av hvordan komplekse varianter former fenotypisk mangfold hos atlanterhavslaks. Arbeidet fremhever betydningen av genetisk redundans for opprettholdelsen av viktige funksjoner, samt dens potensial for seleksjon i ulike miljøer. Videre bidrar avhandlingen med innsikt i mulige evolusjonære utviklingsbaner for strukturelle varianter. Til slutt identifiserer dette avhandlingen kandidatgener knyttet til immunitet som er under seleksjon i domestisert atlanterhavslaks, og forbedrer vår forståelse av rollen komplekse genetiske varianter spiller i domestisering.

4 Synopsis:

4.1 Introduction

Living organisms harbor large phenotypic variations. Many biological traits such as body size, number of offspring, life span, environmental tolerance, or social behavior display large variations across species. These phenotypic variations are remarkable because all phenotypes are encoded by DNA, a molecular structure that is shared between all living organisms. Despite this common base, organisms vary a lot regarding their genomic content and structure. For instance, the number of pairs of chromosomes can vary from 1 to more than a thousand (Sinha et al. 1979; Crosland and Crozier 1986) and the genome size from a few mega base pairs (mbp) to more than 150 Giga base pairs (Gbp) (Elliott and Gregory 2015; Fernández et al. 2024).

The phenotypic and genomic diversity are not independent and influence each other, playing an important role in the evolution of species (Sundberg and Pulkkinen 2015; Kapusta et al. 2017; Bowers and Paterson 2021). The genomic diversity observed is the result of different mutational processes. The most studied types of mutations are Single Nucleotides Polymorphisms (SNPs), where one nucleotide is substituted with another (e.g. A to G mutation). However, these mutations may not represent the majority of mutations present in the genome (Catanach et al. 2019).

Many other mutation types can be more complex and affect a larger genomic region, from several base pairs to the entire genome (Robertson 1916; Ohno 1970; Escaramís et al. 2015). These complex variants play important roles in evolution as they can result in genomic incompatibility, reproductive isolation, and speciation, but also be an important source for phenotypic variation that selection can act on (Sémon and Wolfe 2007; Mérot et al. 2020). Nevertheless, these complex variants are more difficult to quantify than single base changes, and hence their impact on trait- and species evolution is still lacking for most species. In this thesis, I investigated the link between different types of complex genome variants and their impact on the evolution of Atlantic salmon at different temporal levels to better understand how they can shape phenotypic diversity.

4.1.1 Structural variants and their evolution

4.1.1.1 Structural variant formation

Structural variants (SVs) are mutations that affect several consecutive bp, usually described as affecting at least 50 bp (Mahmoud et al. 2019). The most common SVs are deletions or insertions of a region, but the duplication, translocation of a region to another location, or inversion of a region, can also occur (**Figure 1**). SVs are often formed following errors during recombination or DNA repair. Indeed, when a double strand break occurs, different molecular pathways are involved in repairing the DNA. While the repair usually results in retrieving the original sequence (or alternatively the homologous sequence) without any mutation, some errors can occur and for instance lead to deletion, insertion or duplication of a fragment (Escaramís et al. 2015; Carvalho and Lupski 2016). Thus, the formation of SVs is dependent on local recombination rates, where some regions might be more favorable for SVs to appear (Carvalho and Lupski 2016). In addition, other elements are associated with SV formation. For example, some transposable elements families (sequences that can relocate themselves in the genome) can increase the proportion of recombination events and the formation of SVs, in addition to the SV they form themselves (Kidd et al. 2010; Carvalho and Lupski 2016).

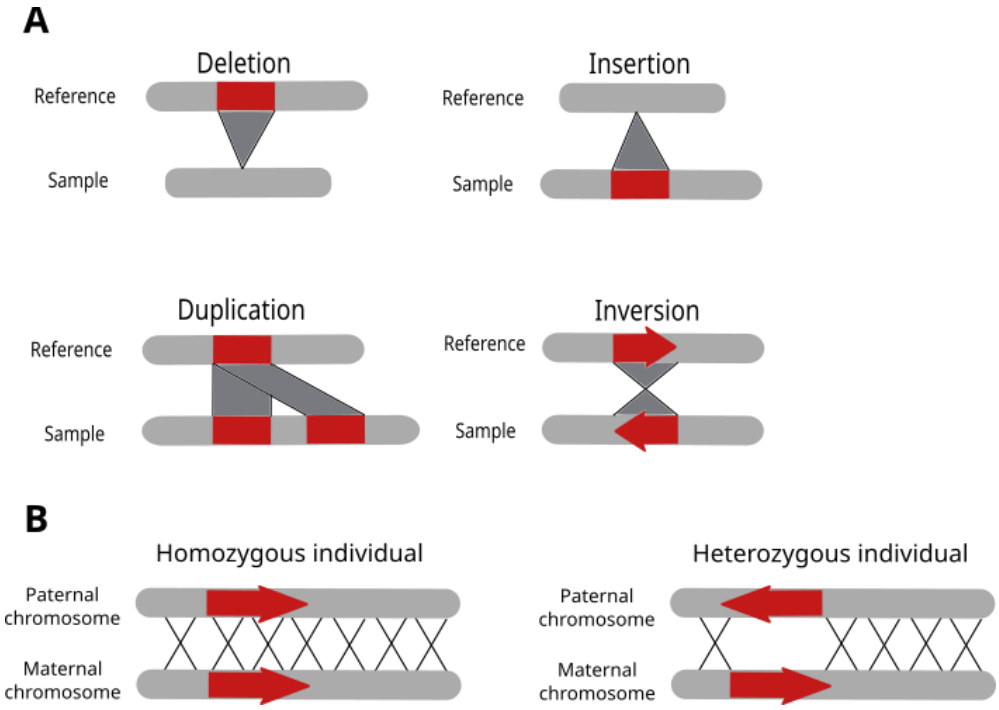


Figure 1: Schematic representation of different SV classes

A: Different classes of SVs are represented with a hypothetic reference genome and a newly sampled individual. **B:** Recombination (indicated by crosses) is reduced or absent in individuals heterozygous for the SVs, here exemplified with an inversion.

4.1.1.2 Structural variant detection

Due to their size that can span more than hundreds of bp, SVs have long been difficult to detect (Lappalainen et al. 2019; Mahmoud et al. 2019). Indeed, the most used sequence technology these past decades has been short reads sequencing, which is a method that cuts the genome into multiple small pieces and produces short sequences, usually <150bp, of all fragmented DNA (**Figure 2**). Even if most SVs remain short (<150bp), for larger SVs, the reads might not span the full length of the SV, which will then be difficult to detect and characterize (**Figure 2**). In addition, regions with high sequence similarity or repetitive sequences (duplications) would cause mapping issues, with unique reads being able to map at several locations, making it harder to detect SVs (Mahmoud et al. 2019; Gong et al. 2021). Despite these limitations, several

SV calling software packages have been developed to identify breakpoints of SVs using short reads sequencing data, mainly based on inference from mapping of reads to a reference genome (Cameron et al. 2019; Kosugi et al. 2019). Recently, the development of long-read sequencing technology, which produces much larger sequences than short reads, is beginning to change the approaches to studying SVs. Indeed, long reads do not suffer the same problem as short reads when it comes to detecting large SVs or mapping issues (**Figure 2**) (Zhao et al. 2021; Ahsan et al. 2023). Nevertheless, due to the high cost of long-read sequencing, short read sequencing is still a useful and practical approach to detect and study SVs (Byrska-Bishop et al.). Several recent studies on SVs have made use of the combination of short-read and long-read sequencing in order to increase the quality of SV detected at a reasonable cost (Yan et al. 2021; Lemay et al. 2022; Mérot et al. 2023; Wold et al. 2025).

In addition to these SVs detection methods that use read-mapping patterns to detect SVs, some other detection methods rely on using SNPs information. Indeed, in individuals heterozygous for the SV, recombination will be reduced or suppressed in the SV region if no homologous template can be found, either because it has been deleted, moved away or inverted (**Figure 1B**) (Mérot et al. 2020). Thus, the linkage disequilibrium (LD) between the SNPs inside the SV and the SV itself will generally be very high. Genome wide, it is then possible to detect large LD blocks, potentially indicating the presence of SVs (**Supplementary Figure 1**). This method is especially relevant for large SVs that are difficult to detect with SV calling software, but that can harbor LD over a large region.

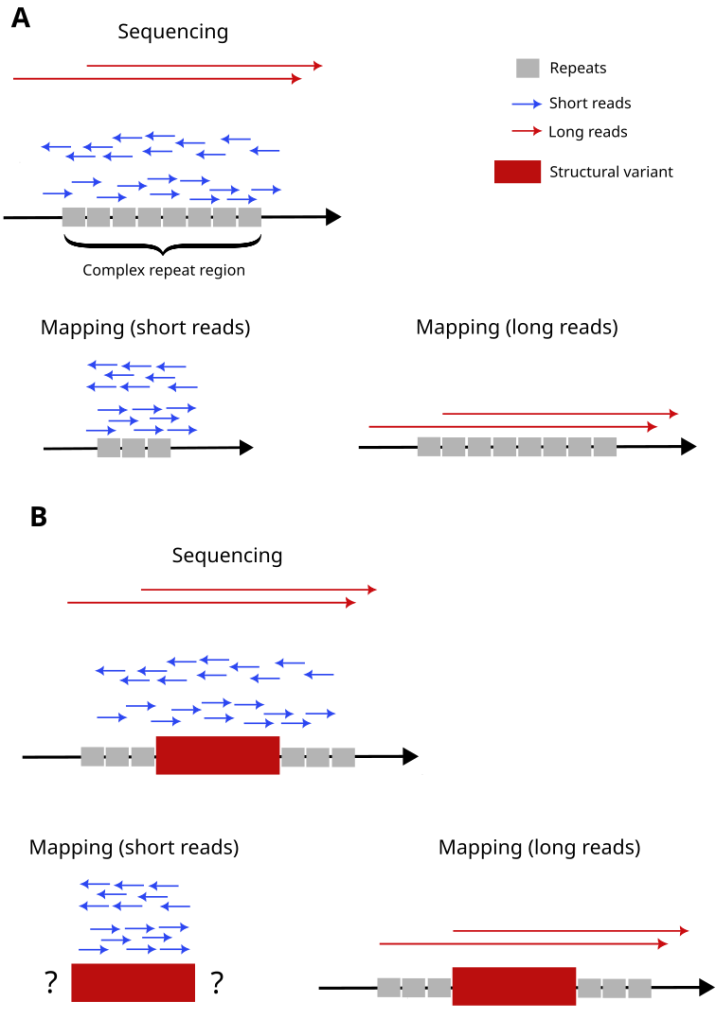


Figure 2: Sequencing technologies and their ability to detect SVs

A: Example of sequencing and mapping of a complex repeat region, using paired end shorts reads and long reads sequencing. Due to their lengths (<150bp), short reads will not spawn the entire region, so only a few repeats will be detected, whereas with long reads, the full region will be covered by a single read, which permits the detection of all the repeats. **B:** When repeats are flanking an SV, short reads sequencing will not inform about the location of the SV even if the SV sequence itself is detected, whereas long reads will cover breakpoints, allowing for mapping of the SV.

4.1.1.3 Mechanisms by which SVs drive evolution

4.1.1.3.1 SVs impacts on gene expression

SVs affect multiple base pairs and, thus, can have dramatic effects on gene expression. Deletion or insertion can disrupt genes or their regulatory elements, causing important changes in gene expression (Feuk et al. 2006; Scott et al. 2021). In addition, duplications can affect the number of copies of a gene, therefore increasing (or decreasing) its expression (Kliebenstein 2008; Gamazon and Stranger 2015). Alternatively, SVs can induce fusion of genes, for instance by deleting intergenic regions or by inserting gene elements near an existing gene (Nattestad et al. 2018; Gawroński et al. 2019).

Within genomes, genes are often located in close proximity in genomic regions within which they physically interact together, and this can play a crucial role in gene expression regulation (Luppino and Joyce 2020; Okhovat et al. 2023). These local gene interactions can be disrupted by SVs, for instance when they affect the boundaries of physically interacting regions and affect chromosomal folding, leading to changes in gene expression (Spielmann et al. 2018).

4.1.1.3.2 SVs affect the genome structure

A hallmark of SVs is the aforementioned recombination suppression that they induce, which is responsible for various evolutionary phenomena. For instance, if some co-adapted alleles happen to be located in a region affected by an SV, the SV will then harbor an evolutionary advantage because this combination won't be reshuffled by recombination (Kirkpatrick and Barton 2006; Mérot et al. 2020). Eventually, this strong linkage between co-adapted alleles can even lead to the formation of a supergene, which is a region containing several genes but acting as an allele of a defined phenotype (Thompson and Jiggins 2014). For instance, in the fire ant *Solenopsis invicta*, a large non-recombining region made up of several inversions is responsible for the colony organization (Wang et al. 2013; Gutiérrez-Valencia et al. 2021).

4.1.1.3.3 SVs play a role in selection dynamics

Another consequence of co-adapted alleles with high LD is that they can favor local adaptation if alleles are adapted to the same environment (Faria et al. 2019; Hämälä et al. 2024). Eventually, this will lead to population divergence and in some cases, reproductive isolation that will precede speciation (Zhang et al. 2023; Berdan et al. 2024). Alternatively, large SVs can cause issues during meiosis that can lead to unbalanced gametes (Searle 1993; Berdan et al. 2024). This situation of underdominance (lower fitness of heterozygous individuals) can then also favor reproductive isolation between populations harboring such SV at different frequencies (Walsh 1982; Berdan et al. 2024). Finally, the effect of SVs on gene expression mentioned earlier can also drive adaptation to new environments (Hämälä et al. 2021; Li et al. 2023).

The long-term dynamics of SVs evolution is another area with uncertainties. This comes in part from the different types of SVs having received different levels of interest, with for instance, the evolutionary fate of inversions being covered by both theoretical models and empirical observations (Kirkpatrick 2010; Faria et al. 2019; Berdan et al. 2021; Berdan et al. 2023). It has, for example, been shown that due to the reduction in recombination rate, inversions can accumulate deleterious mutations, leading to a decrease in the fitness associated with this SV (Jay et al. 2019; Berdan et al. 2021; Huang et al. 2022). However, less is known about SVs persistence and mode of selection over evolutionary time.

In addition to their structural complexity, another reason for the difficulties of investigating SVs persistence, is that when positively selected, they are likely to be under different form of selections than the classical stabilizing and directional selection (Hedrick 2007; Loewe 2008). It is indeed known that SVs can be subject to balancing selection, which include different selection types where selection intensity varies in time and space (Hedrick 2007; Lindtke et al. 2017; Aqil et al. 2023). For example, one type of balancing selection is heterozygote advantage, where heterozygous individuals will be subject to a stronger positive selection than

homozygous individuals (Robertson 1962; Gemmell and Slate 2006). Thus, the selective value of the haplotype depends on its frequency in a population at a given time. This type of balancing selection is especially relevant in the case of SVs because they can cause or accumulate deleterious mutations, which would then provide an advantage to heterozygotes individuals where these deleterious mutations will be masked (Berdan et al. 2021).

Another phenomenon that can play a role in SV retention over time is introgression between species, where an SV is passed to another lineage through hybridization. This phenomenon can even fuel adaptive evolution; it has for instance been shown that several introgressed SVs have played a role in adaptation in mammals (Quan et al. 2021; Dai et al. 2023; Xia et al. 2023).

In some cases, SVs can occur repeatedly at the same genomic locations (Boettger et al. 2016; Cantsilieris et al. 2018; Porubsky et al. 2022). In this case, several similar SVs could be considered a single occurrence, leading to incorrect inference about their evolutionary history. For instance, different alleles of the haptoglobin gene in humans have been formed by recurrent deletions (Boettger et al. 2016).

Together, these findings indicate that although some mechanisms involved in persistence of SVs over time are known, their relative contribution is still unknown. Being able to infer which of these processes have influenced the retention of a present SV is therefore challenging, but doing so would help to predict future SV evolutionary trajectories.

4.1.1.3.4 SVs role during domestication process

The numerous effects of SVs could also have an impact on the domestication process. Domestication often involves an increase in productivity, which can often be achieved by important changes in gene expression (Liu et al. 2019; Díaz-Valenzuela et al. 2023). Since SVs are known to have large effect on gene expression, the domestication context makes them interesting mutations to investigate. Many studies in various organisms, from plants to animals, have investigated the effect of SVs in domestication

contexts. Some studies found that indeed changes in gene expression induced by SVs were selected during domestication (Alonge et al. 2020; Zhou et al. 2021). Other studies identified SVs showing signature of selection during domestication (Low et al. 2020; Cumer et al. 2021; Jia et al. 2024). Nevertheless, some studies revealed that most SVs were deleterious, even in the context of domestication (Zhou et al. 2019; Kou et al. 2020). Together, these studies indicate that SVs can contribute to the shaping of the evolution of species during domestication, but their contrasting roles require more investigation (Hämälä et al. 2024).

4.1.2 Whole genome duplications and their evolution

4.1.2.1 Whole genome duplication formation

Whole Genome Duplication (WGD) describes the process where the genomic content is doubled; that is, every chromosome (and so every gene) ends up having twice as many copies as they had before the WGD. In the case of an ancestral diploid (2n) genome, after WGD this genome will be tetraploid (4n). This rare event can happen because of several processes. One possibility is that the duplication happens in a single species (or even in a single individual), which is called autopolyploidization. This can, for example, happen during the first zygotic division if an additional division occurs, thus doubling the genetic material (Dyban and Baranov 1987; Vasil'ev 2009). Alternatively, a WGD event can result from hybridization between two different species, which we call allopolyploidization (Vasil'ev 2009; Feldman and Levy 2012) (**Supplementary figure 2**). The differences between these two types of WGD go beyond their mechanistic differences. Indeed, both sub genomes in the case of allopolyploidization will remain relatively isolated compared to autopolyploidization, where all parts of the newly duplicated genome are free to recombine (**Supplementary figure 2**).

4.1.2.2 Long term persistence of whole genome duplication

These events remain very rare and are often evolutionary dead ends purged by natural selection, but through evolutionary time some may persist and give rise to

successful lineages (Van De Peer et al. 2009). Consequently, WGD events have been described in many lineages across the tree of life (Dehal and Boore 2005; Albertin and Marullo 2012; Glasauer and Neuhauss 2014; Clark and Donoghue 2018). Examples of well described WGD events include the two rounds of WGD that occurred in the ancestor of all vertebrates (Ohno 1970; Lundin 1993; Garcia-Fernàndez and Holland 1994; Dehal and Boore 2005), the one that occurred before the ancestor of all teleost fishes (Jaillon et al. 2004; Woods et al. 2005; Glasauer and Neuhauss 2014), or the hybridization that occurred in wheat (Mcfadden and Sears 1946; Matsuoka 2011; Feldman and Levy 2012).

The duplication of each gene after a WGD results in genetic redundancy. This redundancy allows mutations to accumulate in one copy without disrupting essential gene functions (Nowak et al. 1997; Tautz 2000; Hanada et al. 2009). These new mutations can in turn lead to the appearance of new functions and facilitate speciation. It has indeed been shown that WGD can increase speciation rates and favor adaptive radiation (Sémon and Wolfe 2007; Tank et al. 2015; Novikova et al. 2020). Interestingly, several domesticated species have undergone WGD events, either before or during the domestication process (Lien et al. 2016; Salman-Minkov et al. 2016; Li et al. 2019; Wang et al. 2024). WGD could potentially favor domestication due to the induced genetic redundancy and allelic diversity (Baduel et al. 2019; Mihajlovic et al. 2025). In addition, some immediate physical consequences of WGD might facilitate domestication. Indeed, by doubling the genetic material, the quantities of molecules in the nucleus are double, which often results in increased cell size (Mable et al. 2011; Doyle and Coate 2019; Tong et al. 2025). An increase in cell size can then be associated with more production; a trait often favored in domesticated species (Milla and Matesanz 2017; Vicente et al. 2023). However, the debate on the importance of WGD in the domestication process is still contrasted, and more studies are needed to shed light on this association (Salman-Minkov et al. 2016).

4.1.2.3 Fates of duplicated genes and their expression after a whole genome duplication

Once the WGD event has occurred, ohnologs will be interdependent due to recombination between sub genomes. However, this state will not last, and ohnologs will progressively start to diverge by accumulating mutations, as recombination is reduced between sub genomes. This divergence process, leading to ohnologs evolving and recombining as if they were diploid again, is called rediploidization (Parey et al. 2022; Redmond et al. 2023). For example, in the case of the two WGD events that happened in the ancestor of vertebrates, nowadays most of the genes have been rediploidized and vertebrates are not considered as "octoploid" (Skrabanek and Wolfe 1998; Hughes 2000).

While they accumulate mutations and diverge, ohnologs can follow several evolutionary trajectories. For instance, this divergence between ohnologs can lead to the emergence of a new function, different from the ancestral one, a process called neofunctionalization (Moriyama and Koshiba-Takeuchi 2018; Birchler and Yang 2022; Kuzmin et al. 2022) (**Figure 3**). This new function can eventually be selected if it's beneficial. Alternatively, the redundancy induced by WGD can cause ohnologs to specialize in different parts of the initial function, which is called subfunctionalization (Moriyama and Koshiba-Takeuchi 2018; Birchler and Yang 2022; Kuzmin et al. 2022) (**Figure 3**). In that case, either the gene product(s) can be subdivided between ohnologs, or the copies could be expressed in different tissues, or at different times in the life cycle. Subfunctionalization can happen because of genetic drift, where mutations accumulations will progressively reduce the efficiency of the gene in a part of the initial function (Force et al. 1999; Lynch and Force 2000). Subfunctionalization can also be positively selected if having two specialized genes increases fitness compared to having a single gene (Des Marais and Rausher 2008; Sikosek et al. 2012). Finally, the accumulation of mutations can cause some copies to lose their function without this being deleterious thanks to the genetic redundancy, a process called pseudogenization (Moriyama and Koshiba-Takeuchi 2018; Kuzmin et al. 2022; Parey et al. 2022) (**Figure 3**).

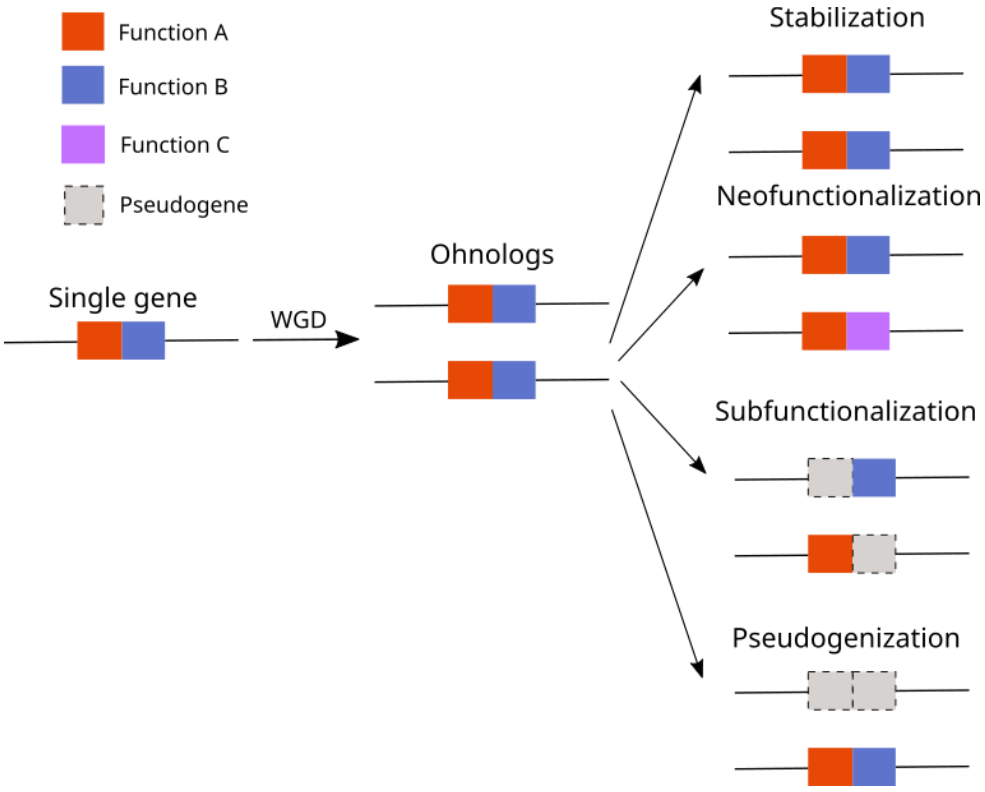


Figure 3: Genes fates after a whole genome duplication

An initial gene can be responsible for several biological functions. Once it gets duplicated (ohnolog), it can either get stabilized and not acquire further mutations, or acquire mutations and be neofunctionalized, subfunctionalized of pseudogenized.

The gene expression of ohnologs will be under additional evolutionary constraints after WGD. Gene expression generally evolves to reach an optimum value, with respect to their interaction with other proteins (Rohlfs and Nielsen 2015; Veitia and Birchler 2022; Dimayacyac et al. 2023). While stoichiometry is preserved immediately after the WGD, because every gene is expressed twice as much as it was before, deviations can start to occur as soon as some genes start to diverge or get pseudogenized. Some genes might then not be expressed at their optimal values, which can affect protein complex formation, or gene regulation. Therefore, the maintenance of proper gene expression balance after WGD is also an important evolutionary aspect (Veitia and

Birchler 2022; Almeida-Silva and Van De Peer 2023). A common observation after WGD events is the evolution of asymmetric regulation of ohnologs, where one copy can become dominant in expression, and the other copy retains a lower expression level (Marlétaz et al. 2018; Gillard et al. 2021). This asymmetric regulation highlights the importance of the sum of the ohnolog expression as being the optimal value of interest for selection. However, there are still some uncertainties about the genetic mechanisms responsible for this divergence and about the rewiring of regulatory connections following WGD.

4.1.3 Atlantic salmon: A model to study complex variants

4.1.3.1 Atlantic salmon recent evolutionary history and domestication

Salmonids, and particularly Atlantic salmon (*Salmo salar*) provide an interesting frame to study the evolution of complex variants, possessing a polyploid genome, as well as abundant transposable elements and SVs. Atlantic salmon is an anadromous fish, which starts its life in rivers, remaining for a variable number of years (1-6), before undergoing a transformation (smoltification) and migrating to the marine environment (sea or ocean). The fish will then feed in the marine environment and after a few years, come back to its natal river to reproduce and females will lay eggs (Stabell 1984; Verspoor et al. 2007; Davidsen et al. 2013).

Between 1.56-1.76 mya ago, Atlantic salmon split into two lineages, with one lineage in Europe and the other in North America, genetically isolated due to distance and the homing behavior of Atlantic salmon (Rougemont and Bernatchez 2018). This separation has progressively led to accumulation of divergence between European and American Atlantic salmon. Some major genetic differences include chromosomes fusions between chromosomes 1 and 23; 26 and 28; and 8 and 29, which are polymorphic in American lineage, but absent in European lineage (Brenna-Hansen et al. 2012; Lehnert et al. 2019; Wellband et al. 2019).

Atlantic salmon is also a species maintained in aquaculture settings. Indeed, in the last 50 years, Atlantic salmon was independently domesticated within the two continents due to an increase in consumer demands (Gjederm 1979). Many breeding programs have started since then, but in this work, we will focus on a subset of them. In North

America, three domesticated strains originated from Penobscot river (Gao et al. 2023), St. John's river (Liu et al. 2017) and Grand Cascapedia river (Nash and Waknitz 2003). In Europe, the MOWI strain was derived from a mixture of Norwegian river populations, primarily from the Vosso drainage basin and the river Årøy, in the 1960s.

These recent domestication events allow for comparison between four distinct lineages: the wild European and American Atlantic salmon lineages, and their domesticated counterpart. This context provides a suitable context for population genomic analysis and comparative genomics, allowing to disentangle geographical effects from the effects of domestication.

4.1.3.2 Salmonid whole genome duplication

Early in their evolutionary history, salmonids underwent a WGD event (autopolyploidisation), which occurred approximately 100 million years ago (Lien et al. 2016; Gundappa et al. 2022) (**Figure 4**). This is a relatively recent WGD event, compared to those that have occurred in the ancestor of vertebrates or teleost (respectively ~450 mya and ~320 mya), and the genome of salmonids is not completely rediploidized yet (Lien et al. 2016; Robertson et al. 2017). Interestingly, the rediploidization following WGD in salmonids has occurred in two waves, with some genes being rediploidized earlier (Ancestral ohnolog resolution – AORe) and some genes rediploidized more recently, in the last 50 mya (Lineage specific ohnolog resolution – LORe) (Robertson et al. 2017). This relatively recent WGD provides an interesting frame to study genomic evolution post WGD, in a case where the rediploidization is advanced but not complete. In addition, the two waves of rediploidization allow the comparison of phenomena occurring after WGD at different tempos, potentially highlighting differences related to rediploidization.

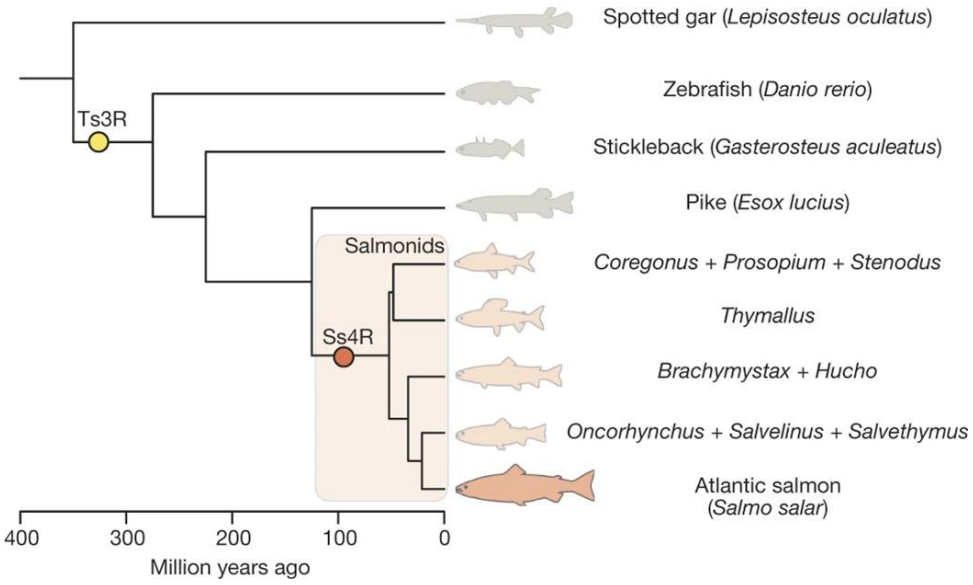


Figure 4: Phylogeny of salmonids and related teleost lineages.

The yellow dot represents the WGD ancestral to all teleosts (Ts3R), and the orange dot the WGD ancestral to salmonids (Ss4R). Figure from Lien et al., 2016.

4.1.3.3 Structural variant evolution in Atlantic salmon

Previous studies on SVs in Atlantic salmon in different lineages have revealed the presence of thousands of SVs (Bertolotti et al. 2020; Stenløkk et al. 2022; Lecomte et al. 2024). A notable observation is the abundance of SVs corresponding to a piggyBac-like transposable element, identifiable by the overrepresentation of SVs of a size between 1432-1436bp (Bertolotti et al. 2020). In addition, evidence for different selection pressure levels between lineages has been found for some SVs. These SVs could have played a role in Atlantic salmon domestication, by impacting behavioral traits (Bertolotti et al. 2020). Others show signatures of local adaptation, being associated with nervous system functions or environmental conditions (Kess et al. 2021; Stenløkk et al. 2022; Lecomte et al. 2024). Together, these studies indicate that SVs have played a role in the evolutionary history of Atlantic salmon, but more studies on these impacts are needed.

Overall, Atlantic salmon has a complex duplicated genome associated with the presence and evolution of complex variants, from WGD to SVs, in addition to chromosome fusions and a documented evolutionary history of transposable elements (Lien et al. 2016; Lehnert et al. 2019; Bertolotti et al. 2020; Stenløkk et al. 2022; Sahlström et al. 2023). This rich genomic structure, coupled with the recent evolutionary history of the species, provides a suitable framework to study the evolution of complex genomic variants. This includes the impact of such variants on the genome after they occur, their retention, and how different selection can influence these phenomena. In addition, this context encourages the investigation of complex variants potentially involved in the domestication process, thus contributing to a better understanding of the link between complex variants and domestication.

4.2 Aims and objectives

The objective of this thesis is to understand the impact of complex genomic variants on evolution. This is highlighted by approaches investigating the long-term effect of such variants, as well as the short-term effects, especially through the prism of domestication.

To these ends, we:

- Investigate how gene expression is affected following a WGD event and subsequent chromosomal rearrangements, using eQTL analysis (**paper 1**)
- Study how selection can act on these complex variants in the frame of recent domestication in Atlantic salmon (**paper 2**)
- Study the formation and retention of SV using multi-methods SV detection (**paper 3**)

4.3 Brief Paper summaries

Paper 1: Large-scale eQTL analyses in Atlantic salmon reveal persistent dosage compensation 100 million years after genome duplication

Whole genome duplication (WGD) events have major consequences on genome structure and subsequent gene evolution. While the patterns and processes of retention of duplicated genes (ohnologs) have been extensively studied, our understanding of their regulatory evolution lags behind. To this end we used a large eQTL (locus associated with gene expression) dataset from Atlantic salmon to investigate the evolution of trans-regulatory connections, following the WGD in a salmonid ancestor 100my ago.

We found a strong enrichment of trans-regulatory connection within chromosomes, highlighting the role of 3D chromatin conformation in shaping connections rewiring after WGD. Gene expression of ohnologs was highly correlated but more pronounced for ohnologs sharing eQTLs. Furthermore, we found that 16 ohnologs pairs displayed a genotype-dependent negative expression level correlation. These 16 pairs were housekeeping genes, and most were in genomic regions that underwent rediploidization recently. This finding is indicative of ohnolog gene dosage compensation and implies strong selection to maintain optimal expression at the level of the sum of ohnolog pair expression. Furthermore, genotype-dependent dosage compensation ohnologs pairs were enriched in late rediploized regions, which we hypothesized to be a consequence of the strong stabilizing selection for dosage compensation combined with the different tempo of rediploidization. We also aimed to experimentally edit some of these ohnologs pairs to validate the dosage compensation in cells, using CRIPR-cas9. We were only able to edit 2 pairs that were initially detected as being in dosage compensation. Unfortunately, after closer investigation on their gene annotation, we revealed that the initial dosage compensation association was due to an error in the automated gene annotation.

Paper 2: Parallel Selection in Domesticated Atlantic Salmon from Divergent Founders Including on Whole-Genome Duplication-derived Homeologous Regions

Adaptation to domestication often targets the same phenotypes. Thus, it's possible that independent events of domestication will give rise to the same genetic change underlying similar targeted phenotypes. In closely related species, this genetic change can theoretically target the same gene(s), referred to as "parallel selection". In the case of WGD, this parallel selection could target duplicated genes. We used the context of Atlantic salmon and its WGD, as well as the independent domestications that have occurred in Europe and North America, to investigate the genomic responses to domestication.

Using a combination of approaches to detect selection (Tajima'D, FST, XP-EHH), we identified 18 genes putatively under selection. Of particular interest are the genes *daxx* and *tapbp*, which are found to be putatively under selection in both European and North American lineages, potentially indicating parallel selection. Moreover, for 3 of those 18 genes (*brd2a, psmb8, psmb13*), one copy originating from the WGD was selected on one continent, while its duplicate copy was selected on the other continent. While not directly targeting the same genes, here selection seems to still favor the exact same protein or at least biological function. Interestingly, most of these genes are known to be involved in immunity-related functions, with 12 genes being present in the major histocompatibility complex region of the genome. These results indicate that selection on immunity targets duplicates genes and highlights the importance of the maintenance of high levels of immune diversity through gene duplications. Together, these results emphasize the role of WGD in shaping genome evolution and the importance of genetic redundancy in the context of a sudden change in selective pressure.

Paper 3 Recurrent emergence and ancient persistence of structural variants across continents in Atlantic salmon

Structural variants (SVs) play an important role in generating and maintaining phenotypic diversity. While we know they are more likely to be subject to some specific evolutionary forces, such as balancing selection and recurrent mutational hotspot, the relative contribution of these forces remains unstudied. We used two different detection methods to detect high-quality short SVs and large SVs, using short reads-based detection methods and local PCA. We found that SVs were enriched in repetitive regions, particularly segmental duplications and LTR retrotransposons, indicating mutational hotspots shaped by genome architecture. We found that a majority of SVs were shared across the four investigated lineages with both detection methods. Using population genomic analysis and evolutionary simulations, we found that these shared SVs mostly originated before the divergence between European and American lineages, with some cases rather attributed to recurrent origin of the SV. Finally, we identified a region on chromosome 3 showing traces of selection for domestication. Together, these results suggest that ancient maintenance and recurrent mutation shaped the evolutionary trajectories of SVs.

4.4 Paper contribution:

Paper 1

I investigated the trans regulatory-gene connections after WGD from the eQTL analysis data produced. I also compared ohnologs gene expression correlation. I investigated the negative gene expression correlation and extracted the 16 genes pairs I found to be in dosage compensation. I took part in drafting and writing the manuscript.

Paper 2

I took part in the investigation of the overlap of the different methods used to find selected genes, as well as comparing the results to null hypotheses. I also participated

in the analysis of selection on the MHC locus on chromomosome 14. I took part in drafting and writing the manuscript.

Paper 3

I worked and optimized the 2 methods used to detect SVs, as well as the filtering process downstream. I investigated the general statistics of SVs identified by both methods. I identified transposable elements (TE) and I compared the location of SVs with the identified TEs. Using slim, I simulated the evolutionary history of SVs, and I analyzed the results obtained by the simulations. I took part in drafting and writing the manuscript.

4.5 Discussion and futures perspectives:

While understanding how genomic variation shapes phenotypic diversity and evolutionary trajectories is a central question in biology, the contribution of complex variants remains insufficiently understood. In the present work, we made use of the Atlantic salmon genome, which contains many complex variants such as an ancestral whole genome duplication and various structural variants, to investigate their evolutionary impact. We studied these variants at several timescales, from a 100-million-year-old WGD to recent continental divergence and domestication of Atlantic salmon lineages by integrating different analytical approaches.

In this work, we found that ohnologs tend to evolve asymmetric gene expression after WGD (**see Paper 1**), aligning with previous observations (Marlétaz et al. 2018; Gillard et al. 2021). We also revealed the presence of dosage compensation between ohnologs, another manifestation of the strong interdependence between ohnologs. We also observed more dosage compensated ohnologs pairs in LORe than in AORe. We hypothesized that this difference was due to a combination of the strong selection for dosage compensation and of the different tempo of evolution of LORe and AORe. Comparatives studies between different species would help to investigate the influence of the tempo of rediploidization on the retention of ohnologs and their dosage compensation. Overall, these results highlight the importance of genetic redundancy to maintain a proper dosage balance after WGD.

Genetic redundancy is often seen as a transitory state allowing for deleterious mutations to accumulate in one copy without affecting the organism's fitness (Nowak et al. 1997; Tautz 2000; Hanada et al. 2009). However, there have been several examples of genetic redundancy maintained over time, sometimes over 100 million years (**Figure 3**) (Dean et al. 2008; Vavouri et al. 2008). Several hypotheses exist to account for positive selection of genetic redundancy. For instance, different models indicate that mutations rates can affect the retention of duplicate genes (Nowak et al. 1997; Krakauer and Nowak 1999; Wagner 2000; Láruson et al. 2020). For instance, high embryonic developmental error rate might allow for redundant genes to persist, ensuring the maintenance of the gene's function (Nowak et al. 1997). Another scenario leading to the maintenance of redundant genes occurs when both genes are

pleiotropic, and mutations can then affect several functions (Nowak et al. 1997; Krakauer and Nowak 1999; Wagner 2000; Vavouri et al. 2008). Previous studies also mentioned the possibility that dosage balance can drive the maintenance of redundant genes (Dean et al. 2008; Láruson et al. 2020), of which the present work provide an empirical example.

In addition to the immediate benefits, redundancy might also provide potential for adaptation in new environmental contexts. This is exemplified by our findings of parallel selection of immune genes in American and European domesticated Atlantic salmon lineages (**see Paper 2**). In this case, the existing genetic redundancy allowed for the selection of beneficial genes under selection for domestication. This phenomenon is often seen in immune gene families, which are known to have high duplications rates, creating many duplicate copies on which selection can act on (Hughes and Yeager 1997; Darbo et al. 2008). This has possibly occurred in the SV we found on chromosome 3 (**see Paper 3**), which contains sentimentally duplicated immune genes, and which also showed signatures of selection during domestication.

These two examples of positive selection of immune genes also highlight the role of complex variants such as SV and WGD in the process of domestication. These complex variants are thought to have much severe impacts on gene function compared to SNPs (Sémon and Wolfe 2007; Stranger et al. 2007; GTEx Consortium et al. 2017). This characteristic could be advantageous in the context of domestication, which often involve very strong and specific selective pressure. Nevertheless, the association between complex variants and domestication might only be limited to immune genes due to the specific importance of genetic redundancy for these genes, highlighting the needs of further investigation to clarify the role of these complex variants in domestication (Salman-Minkov et al. 2016; Alonge et al. 2020; Kou et al. 2020; Hämälä et al. 2024).

During this work, we focused on shared SVs between lineages, with the aim to understand the origin and retention of these SVs. It is generally assumed that most SVs presents in a given population are short lived, recent mutations, due to their most

likely deleterious effect (Belyeu et al. 2021; Berdan et al. 2024). However, in our study, we found that most SVs were shared across several lineages which have been isolated since more than 1 million years (**see Paper 3**). We found that these SVs are most likely ancestral SVs maintained under balancing selection. SVs are known to be prone to some types of balancing selection such as heterozygote advantage (Hedrick 2007; Lindtke et al. 2017; Aqil et al. 2023). The prevalence of such selection types in SVs could be explained by the accumulation of deleterious mutations, which would cause homozygotes to be less fit than heterozygotes (Berdan et al. 2021; Huang et al. 2022). Other types of balancing selection could also be more prevalent in SVs, such as fluctuating selection, where the optimum for a trait changes, which may maintain SVs polymorphisms advantageous for different values of the considered trait (Berdan et al. 2023)

We also revealed that some of the SVs shared between lineages arbored signatures of recurrent mutations. This phenomenon has also been observed in other organisms (Boettger et al. 2016; Cantsilieris et al. 2018; Porubsky et al. 2022), and highlight the need to consider this possibility when making inference of the evolutionary history of SVs. Importantly, such recurring SVs are likely to occur in regions containing repeats (Porubsky et al. 2022), which would then be difficult to identify (**Figure 2**), supporting the needs for better SV detection methods.

During this work, hybridization, which can also be a cause of shared SVs between populations, was not considered due to limited time and resources. Further investigation, including populations genomics analysis to detection genomic introgression, would be needed to determine the relative contribution of hybridization among the observed shared SVs. An introgression scenario could also be included in simulations approach, to compare this possibility with balancing selection and recurrent mutations.

One aspect of the present work is the combination of several methodologies to investigate complex variants. For instance, to detect SVs, two detection methods were

used (**see Paper 3**). Since SVs are difficult to detect, using different methods is a way to improve their detection. The combination of short reads sequencing and long read sequencing, or more recently the development of the pangenome strategy, are others way to combine methods to improve SV detection (Hickey et al. 2020; Lemay et al. 2022; Mérot et al. 2023; Rice et al. 2023; Jayakodi et al. 2024). We also used a simulations approach to investigate the evolutionary history of SVs (**see Paper 3**). This method can complement or confirm the results observed with population genomics analysis, and it also favors the exploration of scenarios that are empirically difficult to confirm, such as recurrent mutations.

Another example of the combination of different methodological approach lies in our tentative to experimentally edit the identified dosage compensated genes in laboratory conditions, in order to confirm the presence of dosage compensation and eventually bring insight into the underlying mechanism(s) (**see Paper 1**). Unfortunately, the high similarity between ohnologs limited our possibilities. This method was, however, first promising, before realizing that some automated gene annotations were incorrect. This annotation issue highlights a more widespread problem concerning automated annotations.

For most genomes, gene annotation is currently done by automated pipelines, based on sequence features and homology with other genes (Kasif and Steffen 2010; Mudge and Harrow 2016; Salzberg 2019). In some cases, this process can lead to unreliable annotations, for instance in regions involving tandemly duplicated genes; some genes may be merged in one annotation, or some may be referred to as unknown. This is for instance the case for the SV region investigated during this thesis on chromosome 3 (**see Paper 3**). While it's very likely that this region contains duplicated immune genes, it's very hard to know which genes are there exactly, therefore limiting possible conclusions. In other cases, the automated annotation can lead to discrepancies between method used, for instance between NCBI and Ensembl (Yen et al. 2017; Chisanga et al. 2022; Manousi 2025). In the present work, it led to pursuing analysis that was revealed to be unreliable due to these discrepancies.

Together, the issue encountered here and in other studies supports that more attention should be given to improve annotation pipelines and make use of existing manual annotations (Salzberg 2019; Yen et al. 2017; Cheng et al. 2017). Recent progress in deep learning methods could also enhance the automated gene annotations (Chen et al. 2024), although they can be sensitive to current errors in existing annotations.

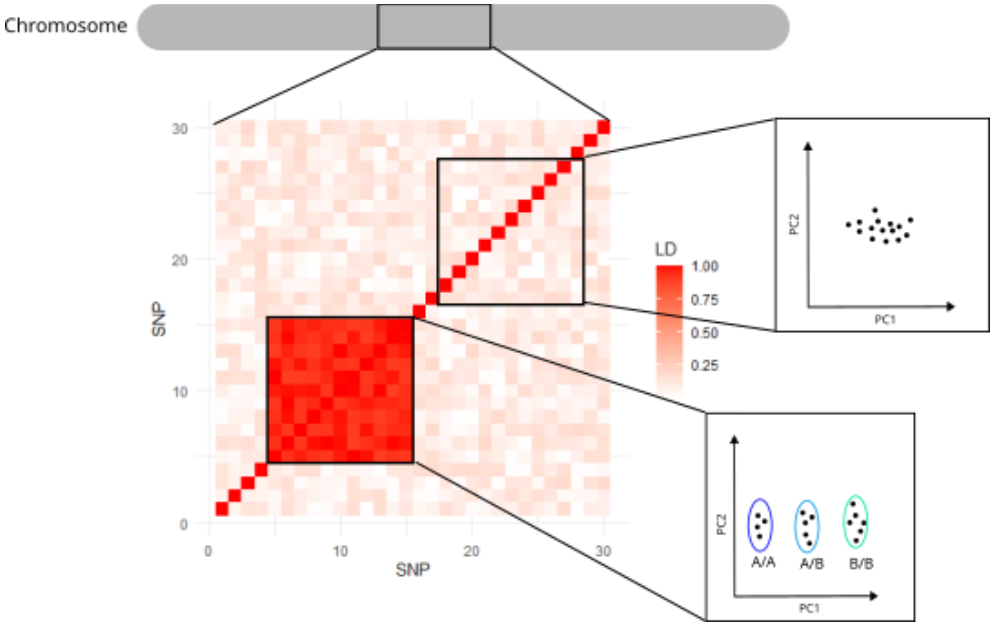
4.6 Conclusion:

This work aimed to investigate the evolution of complex variants and their role in shaping phenotypic diversity in Atlantic Salmon. One finding of this work was to highlight the importance of genetic redundancy, induced by some complex variants, on the retention and selection of beneficial genes. This redundancy is much present in immune genes, which can then be selected under sudden change of selective pressure. The present work also documents examples of a combination of several analytical approaches to detect complex variants and analyze their evolutionary history. Future works combining several analytical approaches or aiming to improve genome annotation would be useful to strengthen our understanding of complex genomic variants.

5 AI Statement

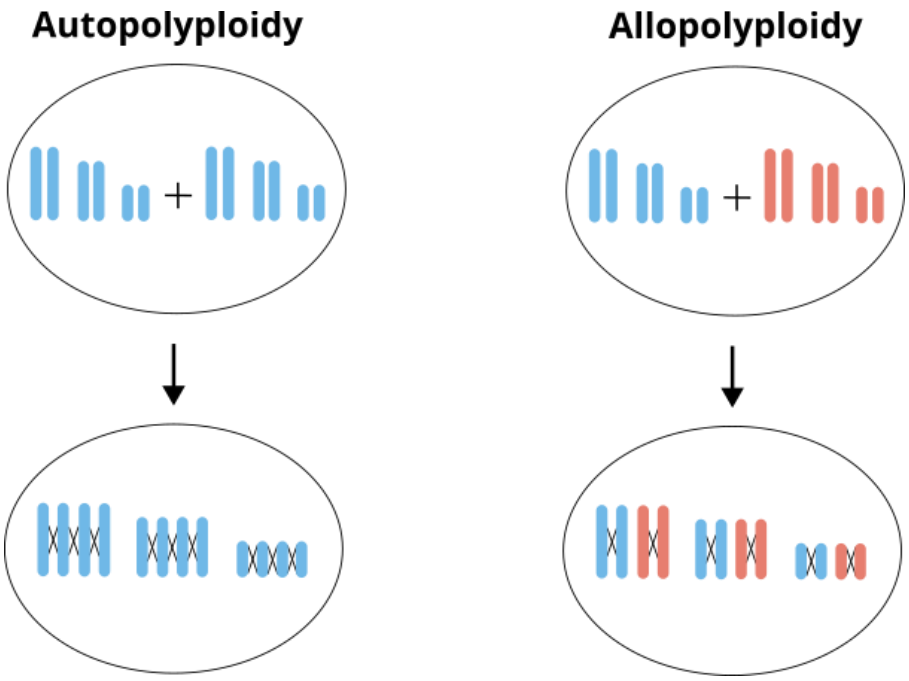
In this thesis, the model ChatGPT 5.2 was used to help for the translation from english to Norwegian (title and abstract).

6 Appendix



Supplementary figure 1: SV identification via LD block

By computing LD between SNPs in a chromosomal region, it's possible to identify LD blocks (left). When clustering SNPs in regions with LD blocks, it's possible that a grouping structure appears, potentially revealing SV. In that case, the genotypes of the putative SV are indicated.



Supplementary figure 2: Different origins of polyploidy

Polyploidy can appear within a single species (autopolyploidy – left) or results from hybridization between species (allopolyploidy – right). We can notice that where all resulting chromosomes in autopolyploids will be able to recombine (indicating by crosses), in allopolyploids due to the preexistent divergence between species, the recombination will only occur in chromosomes originating from the same species.

7 References

Ahsan MU et al. 2023. A survey of algorithms for the detection of genomic structural variants from long-read sequencing data. Nat Methods. 20(8):1143–1158. https://doi.org/10.1038/s41592-023-01932-w

Albertin W, Marullo P. 2012. Polyploidy in fungi: evolution after whole-genome duplication. Proc R Soc B Biol Sci. 279(1738):2497–2509. https://doi.org/10.1098/rspb.2012.0434

Almeida-Silva F, Van De Peer Y. 2023. Whole-genome Duplications and the Long-term Evolution of Gene Regulatory Networks in Angiosperms Purugganan M, editor. Mol Biol Evol. 40(7):msad141. https://doi.org/10.1093/molbev/msad141

Alonge M et al. 2020. Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. Cell. 182(1):145-161.e23. https://doi.org/10.1016/j.cell.2020.05.021

Aqil A, Speidel L, Pavlidis P, Gokcumen O. 2023. Balancing selection on genomic deletion polymorphisms in humans. eLife. 12:e79111. https://doi.org/10.7554/eLife.79111

Baduel P et al. 2019. Relaxed purifying selection in autopolyploids drives transposable element over-accumulation which provides variants for local adaptation. Nat Commun. 10(1):5818. https://doi.org/10.1038/s41467-019-13730-0

Belyeu JR et al. 2021. De novo structural mutation rates and gamete-of-origin biases revealed through genome sequencing of 2,396 families. Am J Hum Genet. 108(4):597–607. https://doi.org/10.1016/j.ajhg.2021.02.012

Berdan EL et al. 2023. How chromosomal inversions reorient the evolutionary process. J Evol Biol. 36(12):1761–1782. https://doi.org/10.1111/jeb.14242

Berdan EL et al. 2024. Structural Variants and Speciation: Multiple Processes at Play. Cold Spring Harb Perspect Biol. 16(3):a041446. https://doi.org/10.1101/cshperspect.a041446

Berdan EL, Blanckaert A, Butlin RK, Bank C. 2021. Deleterious mutation accumulation and the long-term fate of chromosomal inversions Buerkle A, editor. PLOS Genet. 17(3):e1009411. https://doi.org/10.1371/journal.pgen.1009411

Bertolotti AC et al. 2020. The structural variation landscape in 492 Atlantic salmon genomes. Nat Commun. 11(1):5176. https://doi.org/10.1038/s41467-020-18972-x

Birchler JA, Yang H. 2022. The multiple fates of gene duplications: Deletion, hypofunctionalization, subfunctionalization, neofunctionalization, dosage balance constraints, and neutral variation. Plant Cell. 34(7):2466–2474. https://doi.org/10.1093/plcell/koac076

Boettger LM et al. 2016. Recurring exon deletions in the HP (haptoglobin) gene contribute to lower blood cholesterol levels. Nat Genet. 48(4):359–366. https://doi.org/10.1038/ng.3510

Bowers JE, Paterson AH. 2021. Chromosome number is key to longevity of polyploid lineages. New Phytol. 231(1):19–28. https://doi.org/10.1111/nph.17361

Brenna-Hansen S et al. 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. https://doi.org/10.1186/1471-2164-13-432

Byrska-Bishop M et al. High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. 51

Cameron DL, Di Stefano L, Papenfuss AT. 2019. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. Nat Commun. 10(1):3240. https://doi.org/10.1038/s41467-019-11146-4

Cantsilieris S et al. 2018. Recurrent structural variation, clustered sites of selection, and disease risk for the complement factor H (*CFH*) gene family. Proc Natl Acad Sci. 115(19) [accessed 2025 Aug 22]. https://pnas.org/doi/full/10.1073/pnas.1717600115. https://doi.org/10.1073/pnas.1717600115

Carvalho CMB, Lupski JR. 2016. Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet. 17(4):224–238. https://doi.org/10.1038/nrg.2015.25

Catanach A et al. 2019. The genomic pool of standing structural variation outnumbers single nucleotide polymorphism by threefold in the marine teleost *Chrysophrys auratus*. Mol Ecol. 28(6):1210–1223. https://doi.org/10.1111/mec.15051

Chen Z, Ain NU, Zhao Q, Zhang X. 2024. From tradition to innovation: conventional and deep learning frameworks in genome annotation. Brief Bioinform. 25(3):bbae138. https://doi.org/10.1093/bib/bbae138

Cheng C et al. 2017. Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. Plant J. 89(4):789–804. https://doi.org/10.1111/tpj.13415

Chisanga D, Liao Y, Shi W. 2022. Impact of gene annotation choice on the quantification of RNA-seq data. BMC Bioinformatics. 23(1):107. https://doi.org/10.1186/s12859-022-04644-8

Clark JW, Donoghue PCJ. 2018. Whole-Genome Duplication and Plant Macroevolution. Trends Plant Sci. 23(10):933–945. https://doi.org/10.1016/j.tplants.2018.07.006

Crosland MWJ, Crozier RH. 1986. *Myrmecia pilosula* , an Ant with Only One Pair of Chromosomes. Science. 231(4743):1278–1278. https://doi.org/10.1126/science.231.4743.1278

Cumer T, Boyer F, Pompanon F. 2021. Genome-Wide Detection of Structural Variations Reveals New Regions Associated with Domestication in Small Ruminants Enard D, editor. Genome Biol Evol. 13(8):evab165. https://doi.org/10.1093/gbe/evab165

Dai X et al. 2023. A Chinese indicine pangenome reveals a wealth of novel structural variants introgressed from other *Bos* species. Genome Res. 33(8):1284–1298. https://doi.org/10.1101/gr.277481.122

Darbo E, Danchin EGJ, Mc Dermott MFP, Pontarotti P. 2008. Evolution of major histocompatibility complex by “en bloc” duplication before mammalian radiation. Immunogenetics. 60(8):423. https://doi.org/10.1007/s00251-008-0301-7

Davidsen JG et al. 2013. Homing behaviour of Atlantic salmon (S *almo salar*) during final phase of marine migration and river entry Trudel M, editor. Can J Fish Aquat Sci. 70(5):794–802. https://doi.org/10.1139/cjfas-2012-0352

Dean EJ, Davis JC, Davis RW, Petrov DA. 2008. Pervasive and Persistent Redundancy among Duplicated Genes in Yeast McVean G, editor. PLoS Genet. 4(7):e1000113. https://doi.org/10.1371/journal.pgen.1000113

Dehal P, Boore JL. 2005. Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate Holland P, editor. PLoS Biol. 3(10):e314. https://doi.org/10.1371/journal.pbio.0030314

Des Marais DL, Rausher MD. 2008. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. Nature. 454(7205):762–765. https://doi.org/10.1038/nature07092

Díaz-Valenzuela E, Hernández-Ríos D, Cibrián-Jaramillo A. 2023. The role of non-additive gene action on gene expression variation in plant domestication. EvoDevo. 14(1):3. https://doi.org/10.1186/s13227-022-00206-4

Dimayacyac JR, Wu S, Jiang D, Pennell M. 2023. Evaluating the Performance of Widely Used Phylogenetic Models for Gene Expression Evolution Eyre-Walker A, editor. Genome Biol Evol. 15(12):evad211. https://doi.org/10.1093/gbe/evad211

Doyle JJ, Coate JE. 2019. Polyploidy, the Nucleotype, and Novelty: The Impact of Genome Doubling on the Biology of the Cell. Int J Plant Sci. 180(1):1–52. https://doi.org/10.1086/700636

Dyban AP, Baranov VS. 1987. Cytogenetics of mammalian embryonic development. Clarendon Press ; Oxford University Press. (Oxford science publications).

Elliott TA, Gregory TR. 2015. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. Philos Trans R Soc B Biol Sci. 370(1678):20140331. https://doi.org/10.1098/rstb.2014.0331

Escaramís G, Docampo E, Rabionet R. 2015. A decade of structural variants: description, history and methods to detect structural variation. Brief Funct Genomics. 14(5):305–314. https://doi.org/10.1093/bfgp/elv014

Faria R, Johannesson K, Butlin RK, Westram AM. 2019. Evolving Inversions. Trends Ecol Evol. 34(3):239–248. https://doi.org/10.1016/j.tree.2018.12.005

Feldman M, Levy AA. 2012. Genome Evolution Due to Allopolyploidization in Wheat. Genetics. 192(3):763–774. https://doi.org/10.1534/genetics.112.146316

Fernández P et al. 2024. A 160 Gbp fork fern genome shatters size record for eukaryotes. iScience. 27(6):109889. https://doi.org/10.1016/j.isci.2024.109889

Feuk L, Carson AR, Scherer SW. 2006. Structural variation in the human genome. Nat Rev Genet. 7(2):85–97. https://doi.org/10.1038/nrg1767

Force A et al. 1999. Preservation of Duplicate Genes by Complementary, Degenerative Mutations. Genetics. 151(4):1531–1545. https://doi.org/10.1093/genetics/151.4.1531

Gamazon ER, Stranger BE. 2015. The impact of human copy number variation on gene expression: Figure 1. Brief Funct Genomics. 14(5):352–357. https://doi.org/10.1093/bfgp/elv017

Gao G et al. 2023. The generation of the first chromosome-level de novo genome assembly and the development and validation of a 50K SNP array for the St. John River aquaculture strain of North American Atlantic salmon Yáñez J, editor. G3 Genes Genomes Genet. 13(9):jkad138. https://doi.org/10.1093/g3journal/jkad138

Garcia-Fernàndez J, Holland PWH. 1994. Archetypal organization of the amphioxus Hox gene cluster. Nature. 370(6490):563–566. https://doi.org/10.1038/370563a0

Gawroński AR et al. 2019. Structural variation and fusion detection using targeted sequencing data from circulating cell free DNA. Nucleic Acids Res. 47(7):e38–e38. https://doi.org/10.1093/nar/gkz067

Gemmell NJ, Slate J. 2006. Heterozygote Advantage for Fecundity Scheffler K, editor. PLoS ONE. 1(1):e125. https://doi.org/10.1371/journal.pone.0000125

Gillard GB et al. 2021. Comparative regulomics supports pervasive selection on gene dosage following whole genome duplication. Genome Biol. 22(1):103. https://doi.org/10.1186/s13059-021-02323-0

Gjederm T. 1979. Selection for growth rate and domestication in Atlantic salmon. Z Für Tierz Zücht. 96(1–4):56–59. https://doi.org/10.1111/j.1439-0388.1979.tb00199.x

Glasauer SMK, Neuhauss SCF. 2014. Whole-genome duplication in teleost fishes and its evolutionary consequences. Mol Genet Genomics. 289(6):1045–1060. https://doi.org/10.1007/s00438-014-0889-2

Gong T, Hayes VM, Chan EKF. 2021. Detection of somatic structural variants from short-read next-generation sequencing data. Brief Bioinform. 22(3):bbaa056. https://doi.org/10.1093/bib/bbaa056

GTEx Consortium et al. 2017. The impact of structural variation on human gene expression. Nat Genet. 49(5):692–699. https://doi.org/10.1038/ng.3834

Gundappa MK et al. 2022. Genome-Wide Reconstruction of Rediploidization Following Autopolyploidization across One Hundred Million Years of Salmonid Evolution O'Connell M, editor. Mol Biol Evol. 39(1):msab310. https://doi.org/10.1093/molbev/msab310

Gutiérrez-Valencia J, Hughes PW, Berdan EL, Slotte T. 2021. The Genomic Architecture and Evolutionary Fates of Supergenes Betancourt A, editor. Genome Biol Evol. 13(5):evab057. https://doi.org/10.1093/gbe/evab057

Hämälä T et al. 2021. Genomic structural variants constrain and facilitate adaptation in natural populations of *Theobroma cacao* , the chocolate tree. Proc Natl Acad Sci. 118(35):e2102914118. https://doi.org/10.1073/pnas.2102914118

Hämälä T et al. 2024. Impact of whole-genome duplications on structural variant evolution in Cochlearia. Nat Commun. 15(1):5377. https://doi.org/10.1038/s41467-024-49679-y

Hanada K et al. 2009. Evolutionary Persistence of Functional Compensation by Duplicate Genes in Arabidopsis. Genome Biol Evol. 1:409–414. https://doi.org/10.1093/gbe/evp043

Hedrick PW. 2007. Balancing selection. Curr Biol. 17(7):R230–R231. https://doi.org/10.1016/j.cub.2007.01.012

Hickey G et al. 2020. Genotyping structural variants in pangenome graphs using the vg toolkit. Genome Biol. 21(1):35. https://doi.org/10.1186/s13059-020-1941-7

Huang K et al. 2022. Mutation Load in Sunflower Inversions Is Negatively Correlated with Inversion Heterozygosity Takahashi A, editor. Mol Biol Evol. 39(5):msac101. https://doi.org/10.1093/molbev/msac101

Hughes AL. 2000. Polyploidization and Vertebrate Origins: A Review of the Evidence. In: Sankoff D, Nadeau JH, editors. Comparative Genomics. Vol. 1. Springer Netherlands; p 493–502 (Dress A, editor. Computational Biology). [accessed 2025 Sep 3]. http://link.springer.com/10.1007/978-94-011-4309-7_42. https://doi.org/10.1007/978-94-011-4309-7_42

Hughes AL, Yeager M. 1997. Molecular evolution of the vertebrate immune system. BioEssays. 19(9):777–786. https://doi.org/10.1002/bies.950190907

Jaillon O et al. 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature. 431(7011):946–957. https://doi.org/10.1038/nature03025

Jay P et al. 2019. Mutation accumulation in chromosomal inversions maintains wing pattern polymorphism in a butterfly. [accessed 2025 May 13]. http://biorxiv.org/lookup/doi/10.1101/736504. https://doi.org/10.1101/736504

Jayakodi M et al. 2024. Structural variation in the pangenome of wild and domesticated barley. Nature. 636(8043):654–662. https://doi.org/10.1038/s41586-024-08187-1

Jia K-H et al. 2024. Telomere-to-telomere genome assemblies of cultivated and wild soybean provide insights into evolution and domestication under structural variation. Plant Commun. 5(8):100919. https://doi.org/10.1016/j.xplc.2024.100919

Kapusta A, Suh A, Feschotte C. 2017. Dynamics of genome size evolution in birds and mammals. Proc Natl Acad Sci. 114(8) [accessed 2025 May 8]. https://pnas.org/doi/full/10.1073/pnas.1616702114. https://doi.org/10.1073/pnas.1616702114

Kasif S, Steffen M. 2010. Biochemical networks: The evolution of gene annotation. Nat Chem Biol. 6(1):4–5. https://doi.org/10.1038/nchembio.288

Kess T et al. 2021. A putative structural variant and environmental variation associated with genomic divergence across the Northwest Atlantic in Atlantic

Halibut Hauser L, editor. ICES J Mar Sci. 78(7):2371–2384. https://doi.org/10.1093/icesjms/fsab061

Kidd JM et al. 2010. A Human Genome Structural Variation Sequencing Resource Reveals Insights into Mutational Mechanisms. Cell. 143(5):837–847. https://doi.org/10.1016/j.cell.2010.10.027

Kirkpatrick M. 2010. How and Why Chromosome Inversions Evolve. PLoS Biol. 8(9):e1000501. https://doi.org/10.1371/journal.pbio.1000501

Kirkpatrick M, Barton N. 2006. Chromosome Inversions, Local Adaptation and Speciation. Genetics. 173(1):419–434. https://doi.org/10.1534/genetics.105.047985

Kliebenstein DJ. 2008. A Role for Gene Duplication and Natural Variation of Gene Expression in the Evolution of Metabolism Butler G, editor. PLoS ONE. 3(3):e1838. https://doi.org/10.1371/journal.pone.0001838

Kosugi S et al. 2019. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biol. 20(1):117. https://doi.org/10.1186/s13059-019-1720-5

Kou Y et al. 2020. Evolutionary Genomics of Structural Variation in Asian Rice (*Oryza sativa*) Domestication Purugganan M, editor. Mol Biol Evol. 37(12):3507–3524. https://doi.org/10.1093/molbev/msaa185

Krakauer DC, Nowak MA. 1999. Evolutionary preservation of redundant duplicated genes. Semin Cell Dev Biol. 10(5):555–559. https://doi.org/10.1006/scdb.1999.0337

Kuzmin E, Taylor JS, Boone C. 2022. Retention of duplicated genes in evolution. Trends Genet. 38(1):59–72. https://doi.org/10.1016/j.tig.2021.06.016

Lappalainen T, Scott AJ, Brandt M, Hall IM. 2019. Genomic Analysis in the Age of Human Genome Sequencing. Cell. 177(1):70–84. https://doi.org/10.1016/j.cell.2019.02.032

Láruson ÁJ, Yeaman S, Lotterhos KE. 2020. The Importance of Genetic Redundancy in Evolution. Trends Ecol Evol. 35(9):809–822. https://doi.org/10.1016/j.tree.2020.04.009

Lecomte L et al. 2024. Investigating structural variant, indel and single nucleotide polymorphism differentiation between locally adapted Atlantic salmon populations. Evol Appl. 17(3):e13653. https://doi.org/10.1111/eva.13653

Lehnert SJ et al. 2019. Chromosome polymorphisms track trans-Atlantic divergence and secondary contact in Atlantic salmon. Mol Ecol. 28(8):2074–2087. https://doi.org/10.1111/mec.15065

Lemay M-A et al. 2022. Combined use of Oxford Nanopore and Illumina sequencing yields insights into soybean structural variation biology. BMC Biol. 20(1):53. https://doi.org/10.1186/s12915-022-01255-w

Li H, Huang C-H, Ma H. 2019. Whole-Genome Duplications in Pear and Apple. In: Korban SS, editor. The Pear Genome. Springer International Publishing; p 279–299 (Compendium of Plant Genomes). [accessed 2025 May 22]. http://link.springer.com/10.1007/978-3-030-11048-2_15. https://doi.org/10.1007/978-3-030-11048-2_15

Li Y et al. 2023. Large-Scale Chromosomal Changes Lead to Genome-Level Expression Alterations, Environmental Adaptation, and Speciation in the Gayal (*Bos frontalis*) Wei F, editor. Mol Biol Evol. 40(1):msad006. https://doi.org/10.1093/molbev/msad006

Lien S et al. 2016. The Atlantic salmon genome provides insights into rediploidization. Nature. 533(7602):200–205. https://doi.org/10.1038/nature17164

Lindtke D et al. 2017. Long-term balancing selection on chromosomal variants associated with crypsis in a stick insect. Mol Ecol. 26(22):6189–6205. https://doi.org/10.1111/mec.14280

Liu L et al. 2017. A genome scan for selection signatures comparing farmed Atlantic salmon with two wild populations: Testing colocalization among outlier markers, candidate genes, and quantitative trait loci for production traits. Evol Appl. 10(3):276–296. https://doi.org/10.1111/eva.12450

Liu W et al. 2019. Decrease of gene expression diversity during domestication of animals and plants. BMC Evol Biol. 19(1):19. https://doi.org/10.1186/s12862-018-1340-9

Loewe L. 2008. Negative selection. Nat Educ. 1

Low WY et al. 2020. Haplotype-resolved genomes provide insights into structural variation and gene content in Angus and Brahman cattle. Nat Commun. 11(1):2071. https://doi.org/10.1038/s41467-020-15848-y

Lundin LG. 1993. Evolution of the Vertebrate Genome as Reflected in Paralogous Chromosomal Regions in Man and the House Mouse. Genomics. 16(1):1–19. https://doi.org/10.1006/geno.1993.1133

Luppino JM, Joyce EF. 2020. Single cell analysis pushes the boundaries of TAD formation and function. Curr Opin Genet Dev. 61:25–31. https://doi.org/10.1016/j.gde.2020.03.005

Lynch M, Force A. 2000. The Probability of Duplicate Gene Preservation by Subfunctionalization. Genetics. 154(1):459–473. https://doi.org/10.1093/genetics/154.1.459

Mable BK, Alexandrou MA, Taylor MI. 2011. Genome duplication in amphibians and fish: an extended synthesis. J Zool. 284(3):151–182. https://doi.org/10.1111/j.1469-7998.2011.00829.x

Mahmoud M et al. 2019. Structural variant calling: the long and the short of it. Genome Biol. 20(1):246. https://doi.org/10.1186/s13059-019-1828-7

Manousi D. 2025. Genomic basis of pathogen response and immune gene evolution in Atlantic salmon. Norwegian University of Life Sciences, Faculty of Biosciences, Department of Animal and Aquacultural Sciences. https://nva.sikt.no/registration/01997bb4f1e3-e0cdb612-c483-482c-a812-0ca9796301ba

Marlétaz F et al. 2018. Amphioxus functional genomics and the origins of vertebrate gene regulation. Nature. 564(7734):64–70. https://doi.org/10.1038/s41586-018-0734-6

Matsuoka Y. 2011. Evolution of Polyploid Triticum Wheats under Cultivation: The Role of Domestication, Natural Hybridization and Allopolyploid Speciation in their Diversification. Plant Cell Physiol. 52(5):750–764. https://doi.org/10.1093/pcp/pcr018

Mcfadden ES, Sears ER. 1946. THE ORIGIN OF TRITICUM SPELTA AND ITS FREE-THRESHING HEXAPLOID RELATIVES*. J Hered. 37(3):81–89. https://doi.org/10.1093/oxfordjournals.jhered.a105590

Mérot C et al. 2023. Genome assembly, structural variants, and genetic differentiation between lake whitefish young species pairs (*Coregonus* sp.) with long and short reads. Mol Ecol. 32(6):1458–1477. https://doi.org/10.1111/mec.16468

Mérot C, Oomen RA, Tigano A, Wellenreuther M. 2020. A Roadmap for Understanding the Evolutionary Significance of Structural Genomic Variation. Trends Ecol Evol. 35(7):561–572. https://doi.org/10.1016/j.tree.2020.03.002

Mihajlovic L et al. 2025. A direct experimental test of Ohno's hypothesis. eLife. 13:RP97216. https://doi.org/10.7554/eLife.97216.3

Milla R, Matesanz S. 2017. Growing larger with domestication: a matter of physiology, morphology or allocation? Byers D, editor. Plant Biol. 19(3):475–483. https://doi.org/10.1111/plb.12545

Moriyama Y, Koshiba-Takeuchi K. 2018. Significance of whole-genome duplications on the emergence of evolutionary novelties. Brief Funct Genomics. 17(5):329–338. https://doi.org/10.1093/bfgp/ely007

Mudge JM, Harrow J. 2016. The state of play in higher eukaryote gene annotation. Nat Rev Genet. 17(12):758–772. https://doi.org/10.1038/nrg.2016.119

Nash CE, Waknitz FW. 2003. Interactions of Atlantic salmon in the Pacific Northwest. Fish Res. 62(3):237–254. https://doi.org/10.1016/S0165-7836(03)00063-8

Nattestad M et al. 2018. Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. Genome Res. 28(8):1126–1135. https://doi.org/10.1101/gr.231100.117

Novikova PYu et al. 2020. Polyploidy breaks speciation barriers in Australian burrowing frogs Neobatrachus Mauricio R, editor. PLOS Genet. 16(5):e1008769. https://doi.org/10.1371/journal.pgen.1008769

Nowak MA, Boerlijst MC, Cooke J, Smith JM. 1997. Evolution of genetic redundancy. Nature. 388(6638):167–171. https://doi.org/10.1038/40618

Ohno S. 1970. Evolution by Gene Duplication. Springer Berlin Heidelberg. [accessed 2025 Mar 27]. http://link.springer.com/10.1007/978-3-642-86659-3. https://doi.org/10.1007/978-3-642-86659-3

Okhovat M et al. 2023. TAD evolutionary and functional characterization reveals diversity in mammalian TAD boundary properties and function. Nat Commun. 14(1):8111. https://doi.org/10.1038/s41467-023-43841-8

Parey E et al. 2022. An atlas of fish genome evolution reveals delayed rediploidization following the teleost whole-genome duplication. Genome Res. 32(9):1685–1697. https://doi.org/10.1101/gr.276953.122

Porubsky D et al. 2022. Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. Cell. 185(11):1986-2005.e26. https://doi.org/10.1016/j.cell.2022.04.017

Quan C et al. 2021. Characterization of structural variation in Tibetans reveals new evidence of high-altitude adaptation and introgression. Genome Biol. 22(1):159. https://doi.org/10.1186/s13059-021-02382-3

Redmond AK et al. 2023. Independent rediploidization masks shared whole genome duplication in the sturgeon-paddlefish ancestor. Nat Commun. 14(1):2879. https://doi.org/10.1038/s41467-023-38714-z

Rice ES et al. 2023. A pangenome graph reference of 30 chicken genomes allows genotyping of large and complex structural variants. BMC Biol. 21(1):267. https://doi.org/10.1186/s12915-023-01758-0

Robertson A. 1962. SELECTION FOR HETEROZYGOTES IN SMALL POPULATIONS. Genetics. 47(9):1291–1300. https://doi.org/10.1093/genetics/47.9.1291

Robertson FM et al. 2017. Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification. Genome Biol. 18(1):111. https://doi.org/10.1186/s13059-017-1241-z

Robertson WmRB. 1916. Chromosome studies. I. Taxonomic relationships shown in the chromosomes of tettigidae and acrididae: V-shaped chromosomes and their significance in acrididae, locustidae, and gryllidae: Chromosomes and variation. J Morphol. 27(2):179–331. https://doi.org/10.1002/jmor.1050270202

Rohlfs RV, Nielsen R. 2015. Phylogenetic ANOVA: The Expression Variance and Evolution Model for Quantitative Trait Evolution. Syst Biol. 64(5):695–708. https://doi.org/10.1093/sysbio/syv042

Rougemont Q, Bernatchez L. 2018. The demographic history of Atlantic salmon (*Salmo salar*) across its distribution range reconstructed from approximate Bayesian computations*: ATLANTIC SALMON HISTORY AND LINKED SELECTION. Evolution. 72(6):1261–1277. https://doi.org/10.1111/evo.13486

Sahlström HM et al. 2023. Functional validation of transposable element–derived *cis* -regulatory elements in Atlantic salmon Andrews B, editor. G3 Genes Genomes Genet. 13(4):jkad034. https://doi.org/10.1093/g3journal/jkad034

Salman-Minkov A, Sabath N, Mayrose I. 2016. Whole-genome duplication as a key factor in crop domestication. Nat Plants. 2(8):16115. https://doi.org/10.1038/nplants.2016.115

Salzberg SL. 2019. Next-generation genome annotation: we still struggle to get it right. Genome Biol. 20(1):92, s13059-019-1715–2. https://doi.org/10.1186/s13059-019-1715-2

Scott AJ, Chiang C, Hall IM. 2021. Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. Genome Res. 31(12):2249–2257. https://doi.org/10.1101/gr.275488.121

Searle JB. 1993. Chromosomal Hybrid Zones in Eutherian Mammals. In: Hybrid Zones and the Evolutionary Process. p 309–353

https://books.google.no/books/about/Hybrid_Zones_and_the_Evolutionary_Proces.html?id=aFJFkVKskYIC&redir_esc=y

Sémon M, Wolfe KH. 2007. Consequences of genome duplication. Curr Opin Genet Dev. 17(6):505–512. https://doi.org/10.1016/j.gde.2007.09.007

Sikosek T, Chan HS, Bornberg-Bauer E. 2012. Escape from Adaptive Conflict follows from weak functional trade-offs and mutational robustness. Proc Natl Acad Sci. 109(37):14888–14893. https://doi.org/10.1073/pnas.1115620109

Sinha BMB, Srivastava DP, Jha J. 1979. Occurrence of Various Cytotypes of *Ophioglossum Reticulatum* L. in a Population from N. E. India. Caryologia. 32(2):135–146. https://doi.org/10.1080/00087114.1979.10796781

Skrabanek L, Wolfe KH. 1998. Eukaryote genome duplication - where's the evidence? Curr Opin Genet Dev. 8(6):694–700. https://doi.org/10.1016/S0959-437X(98)80039-7

Spielmann M, Lupiáñez DG, Mundlos S. 2018. Structural variation in the 3D genome. Nat Rev Genet. 19(7):453–467. https://doi.org/10.1038/s41576-018-0007-0

Stabell OB. 1984. HOMING AND OLFACTION IN SALMONIDS: A CRITICAL REVIEW WITH SPECIAL REFERENCE TO THE ATLANTIC SALMON. Biol Rev. 59(3):333–388. https://doi.org/10.1111/j.1469-185X.1984.tb00709.x

Stenløkk K et al. 2022. The emergence of supergenes from inversions in Atlantic salmon. Philos Trans R Soc B Biol Sci. 377(1856):20210195. https://doi.org/10.1098/rstb.2021.0195

Stranger BE et al. 2007. Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes. Science. 315(5813):848–853. https://doi.org/10.1126/science.1136678

Sundberg L-R, Pulkkinen K. 2015. Genome size evolution in macroparasites. Int J Parasitol. 45(5):285–288. https://doi.org/10.1016/j.ijpara.2014.12.007

Tank DC et al. 2015. Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. New Phytol. 207(2):454–467. https://doi.org/10.1111/nph.13491

Tautz D. 2000. A genetic uncertainty problem. Trends Genet. 16(11):475–477. https://doi.org/10.1016/S0168-9525(00)02118-1

Thompson MJ, Jiggins CD. 2014. Supergenes and their role in evolution. Heredity. 113(1):1–8. https://doi.org/10.1038/hdy.2014.20

Tong K et al. 2025. Genome duplication in a long-term multicellularity evolution experiment. Nature. 639(8055):691–699. https://doi.org/10.1038/s41586-025-08689-6

Van De Peer Y, Maere S, Meyer A. 2009. The evolutionary significance of ancient genome duplications. Nat Rev Genet. 10(10):725–732. https://doi.org/10.1038/nrg2600

Vasil'ev VP. 2009. Mechanisms of Polyploid Evolution in Fish: Polyploidy in Sturgeons. In: Carmona R et al., editors. Biology, Conservation and Sustainable Development of Sturgeons. Vol. 29. Springer Netherlands; p 97–117 [accessed 2025 Mar 27]. http://link.springer.com/10.1007/978-1-4020-8437-9_6. https://doi.org/10.1007/978-1-4020-8437-9_6

Vavouri T, Semple JI, Lehner B. 2008. Widespread conservation of genetic redundancy during a billion years of eukaryotic evolution. Trends Genet. 24(10):485–488. https://doi.org/10.1016/j.tig.2008.08.005

Veitia RA, Birchler JA. 2022. Gene-dosage issues: a recurrent theme in whole genome duplication events. Trends Genet. 38(1):1–3. https://doi.org/10.1016/j.tig.2021.06.006

Verspoor E, Stradmeyer L, Nielsen J, editors. 2007. The Atlantic Salmon: Genetics, Conservation and Management. 1st ed. Wiley. [accessed 2025 Apr 16]. https://onlinelibrary.wiley.com/doi/book/10.1002/9780470995846. https://doi.org/10.1002/9780470995846

Vicente MH et al. 2023. The *ORGAN SIZE* (*ORG*) locus modulates both vegetative and reproductive gigantism in domesticated tomato. Ann Bot. 132(7):1233–1248. https://doi.org/10.1093/aob/mcad150

Wagner A. 2000. The Role of Population Size, Pleiotropy and Fitness Effects of Mutations in the Evolution of Overlapping Gene Functions. Genetics. 154(3):1389–1401. https://doi.org/10.1093/genetics/154.3.1389

Walsh JB. 1982. Rate of Accumulation of Reproductive Isolation by Chromosome Rearrangements. Am Nat. 120(4):510–532. https://doi.org/10.1086/284008

Wang B et al. 2024. Whole-genome Sequencing Reveals Autooctoploidy in Chinese Sturgeon and Its Evolutionary Trajectories Jiang Y, editor. Genomics Proteomics Bioinformatics. 22(1):qzad002. https://doi.org/10.1093/gpbjnl/qzad002

Wang J et al. 2013. A Y-like social chromosome causes alternative colony organization in fire ants. Nature. 493(7434):664–668. https://doi.org/10.1038/nature11832

Wellband K et al. 2019. Chromosomal fusion and life history-associated genomic variation contribute to within-river local adaptation of Atlantic salmon. Mol Ecol. 28(6):1439–1459. https://doi.org/10.1111/mec.14965

Wold JR et al. 2025. The promise and challenges of characterizing genome-wide structural variants: A case study in a critically endangered parrot. Mol Ecol Resour. 25(5):e13783. https://doi.org/10.1111/1755-0998.13783

Woods IG et al. 2005. The zebrafish gene map defines ancestral vertebrate chromosomes. Genome Res. 15(9):1307–1314. https://doi.org/10.1101/gr.4134305

Xia X et al. 2023. Structural variation and introgression from wild populations in East Asian cattle genomes confer adaptation to local environment. Genome Biol. 24(1):211. https://doi.org/10.1186/s13059-023-03052-2

Yan SM et al. 2021. Local adaptation and archaic introgression shape global diversity at human structural variant loci. eLife. 10:e67615. https://doi.org/10.7554/eLife.67615

Yen JL et al. 2017. A variant by any name: quantifying annotation discordance across tools and clinical databases. Genome Med. 9(1):7. https://doi.org/10.1186/s13073-016-0396-7

Zhang L et al. 2023. Population genomic evidence of selection on structural variants in a natural hybrid zone. Mol Ecol. 32(6):1497–1514. https://doi.org/10.1111/mec.16469

Zhao X et al. 2021. Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies. Am J Hum Genet. 108(5):919–928. https://doi.org/10.1016/j.ajhg.2021.03.014

Zhou R et al. 2021. The Meishan pig genome reveals structural variation-mediated gene expression and phenotypic divergence underlying Asian pig domestication. Mol Ecol Resour. 21(6):2077–2092. https://doi.org/10.1111/1755-0998.13396

Zhou Y et al. 2019. The population genetics of structural variants in grapevine domestication. Nat Plants. 5(9):965–979. https://doi.org/10.1038/s41477-019-0507-8

8 Papers

PAPER I

Long title: Large-scale eQTL analyses in Atlantic salmon reveal persistent dosage compensation 100 million years after genome duplication.

Authors:

Célian Diblasi[1], Domniki Manousi[1,2], David Hazlerigg[3], Lars Grønvold[1], Nicola Jane Barson[1], Simen Rød Sandve[1], Marie Saitou[1*]

[1] Department of Animal and Aquacultural Sciences, Section for Genome Biology, Faculty of Biosciences, Norwegian University of Life Sciences, 1433 Ås, Norway

[2]The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh Easter Bush, EH25 9RG, Midlothian, United Kingdom.

[3] Department of Arctic and Marine Biology, UiT The Arctic University of Norway, N-9037, Tromsø, Norway

Emails:

Célian Diblasi: celian.diblasi@nmbu.no

Domniki Manousi: dmanousi@ed.ac.uk

David Hazlerigg: david.hazlerigg@uit.no

Lars Grønvold: lars.gronvold@nmbu.no

Nicola J. Barson: nicola.barson@nmbu.no

Simen Rød Sandve: simen.sandve@nmbu.no

Marie Saitou: marie.saitou@nmbu.no

Abstract

Whole-genome duplication (WGD) through autopolyploidization has played a role in genome evolution across eukaryotes. A major consequence of WGD is the rewiring of gene regulatory networks, partly driven by selection on dosage balance. In multicellular organisms, evidence for dosage balance selection has relied on

comparative patterns of duplicate gene retention and expression, with few studies directly examining regulatory architecture after WGD.

Here, we analysed a large-scale eQTL dataset from Atlantic salmon (*Salmo salar*), which experienced a WGD 100 million years ago. We found that trans-regulatory connections were enriched between duplicated regions, indicating long-term conservation of ancestral interchromosomal regulatory interactions. Overall, 230 duplicated genes (5%) shared eQTLs, suggesting conserved regulatory control. Moreover, 16 gene pairs showed compensatory expression effects mediated by a common regulator, consistent with predictions of the dosage balance hypothesis. These gene pairs were significantly enriched in recently rediploidized regions. Our results indicate long-term maintenance of dosage balance after WGD.

Teaser:

Genetic regulation in Atlantic salmon shows that duplicated genes can remain dosage-balanced across 100 million years of evolution.

Introduction

Whole genome duplication (WGD) is a major driving force in genome evolution across the eukaryote tree of life (1–4). Doubling of the genetic material results in functional redundancy and reduced selective constraints. This allows for mutations to accumulate (1, 5) which impact trait evolution including gene regulatory phenotypes (6–8) . However, our understanding of how WGDs impact the gene regulatory architecture through rewiring of cis- and trans-regulatory connections is only addressed in a few studies in yeast (9, 10) and is still limited.

Gene expression levels evolve under selection for optimum 'trait' values (11, 12). These optima change depending on (i) external factors such as the local environment and (ii) internal genomic factors such as imbalances in gene dosage that upsets the stoichiometric balance between RNA and/or protein levels (13, 14). The genomic upheaval that follows WGDs, including rampant pseudogenization of redundant

duplicate copies, must therefore spark selection to minimize stoichiometric imbalances. This could be achieved by selection to retain a balanced number of gene copies involved in dosage-sensitive interactions (15, 16), selection on gene expression levels (17), or a combination. Studies of long-term evolutionary consequences of WGDs highlight that asymmetric evolution of ohnolog regulation is a common fate (17, 18) resulting in one copy dominating expression. This is in line with a model whereby optimal expression levels are mostly resolved through selection on the sum of ohnolog expression levels, blind to which copy contributes to the expression trait value.

While numerous comparative transcriptomics studies have provided valuable insights into ohnolog regulatory evolution, our understanding of the genetic mechanisms in play is limited. To address this knowledge gap, associations between genetic variation and variation in gene expression levels, i.e. expression Quantitative Trait Loci (eQTLs), offers a rare opportunity to study patterns of conservation and divergence in gene regulatory mechanisms across duplicated genes (19, 20). Here we leverage a large eQTL dataset from Atlantic salmon, a lineage which underwent autopolyploidization 100 Mya (21, 22), to characterize divergence in gene regulatory network connections inferred by trans-eQTL associations. This lineage is particularly interesting due to its two waves of rediploidization, resulting in ancestral ohnolog resolution regions (AORe; ~100mya) and lineage-specific ohnolog resolution regions (LORe; <50mya) (21, 22).

Results

Trans-regulatory connections across homeologous regions are retained following WGD

To identify trans-regulatory associations, we used a recently published dataset of bulk-RNAseq from gill samples of ~1000 Atlantic salmon raised on constant photoperiod, with matching genome wide SNP data for all individuals (23). Using this data, we identified 12,056 eQTL impacting the expression of 19,526 genes in gill

tissue. Of these, 52% (5,951) were cis-eQTLs and 43% (5,022) were trans-eQTLs, with the remaining 5% (556) showing both types of interaction (**Figure 1A, B**). Note that the numbers of eQTL in this study differ from (23) due to the different sample sizes in the two analyses.

We estimated the average effect size to be 0.358 for cis-eQTLs and 0.379 for trans eQTL. Gene expression levels were correlated with the proportion of genes with an eQTL, as well as the proportion of trans-eQTLs (**Figure 1C**), most likely due to increased power to detect eQTLs. Highly expressed genes without detectable eQTLs were predominantly involved in ribosomal function (GO:0005840, FDR=2.8E-155), in line with the regulation of these genes being under strong purifying selection (24–26) (**Figure 1C**). Distribution of trans- and cis-eQTLs showed that chromosome 14 was strongly enriched for trans-eQTLs (chi^2 = 29914, p-value < 2.2E-16), while there was no such chromosomal bias present in cis-eQTLs associations (**Figure 1D**). Among the most significant cis-eQTLs for each gene, 35.3% were located within 500kb (10% of the cis-definition range) of the transcription start site (TSS) (**Figure 1E**).

WGD and chromosomes fusions affect trans-eQTL connections

Immediately after a WGD occurs, all the gene regulatory connections are duplicated. However, during the subsequent rediploidization, regulatory connections may be lost or rewired independently among ohnolog copies (16, 27). Trans connections typically refer to interactions between proteins, such as transcription factors, and genomic elements that are not necessarily immediately adjacent such as enhancers or distant regulatory sequences (28, 29). It is thus reasonable to expect that when a genomic region is duplicated, these interacting elements may be able to bind to both copies of the sequence. Furthermore, following whole genome duplication (WGD), new trans connections can arise. These could in theory involve interactions between a duplicated gene and its 'original' cis-regulatory element landscape physically located on the duplicate copy chromosome. However, this is rather unlikely due to the inherent local nature and distance-dependent specificity of cis elements (30). New

trans connections can also arise between genes and regulatory elements that become physically linked on the same chromosome as a result of chromosomal fusions. To investigate the trans connections landscape after WGD, we classified trans-eQTLs into the following four groups (**Figure 2A**): (i) Inter-chromosomal homeologs where the eQTL and target gene reside in homeologous regions (regions duplicated by WGD) on different chromosomes, (ii) inter-chromosomal non-homeologs where the eQTL and the target gene are located on different chromosomes but not in homeologous regions, (iii) intra-chromosome syntenic where eQTL and the target gene reside on the same chromosome within a syntenic block originating from the WGD >5 Mbp apart, and (iv) intra-chromosome non-syntenic where the eQTL and the target gene reside on the same chromosome as a result of chromosome fusion following WGD (22). A schematic representation of evolutionary events leading to these four different types of trans regulatory connections before and after WGD is shown in **Figure 2B**.

Among the four trans-regulatory categories, most (61%) were inter-chromosomal non-homeolog, while 20% were intra-chromosomal non-syntenic, 17% were intra-chromosomal non-syntenic, and only 2% were inter-chromosomal homeologs (**Figure 2C**). Compared to randomized data (**see Materials and Methods**) this represents a 13.7-fold enrichment of intra-chromosomal syntenic trans-eQTL connections (category iii) and a 6-fold enrichment of intra-chromosomal non-syntenic trans-eQTL connections (category iv) (**Figure 2D**). These results suggest that trans-regulatory connections within the same chromosomes tend to be retained, and that new regulatory connections are being established between variants and genes within newly fused chromosomes. We observed that the distance between eQTLs and the target genes were higher for intra-chromosomal non-syntenic trans-eQTL connections (category iv) than for intra-chromosomal syntenic connections (category iii), even after correcting for differences in the maximum eQTL-gene distance between the two categories. This result confirmed that the enrichment of intra-chromosomal non-syntenic trans connections is not a technical artifact arising from the distance cutoff (**Figure S1**).

With respect to inter-chromosomal relationships, we observed two times more inter-chromosomal homeologous eQTLs target gene connections (category i) than expected by chance (**Figure 2D**), which suggest that regulatory elements, originally located on the same chromosome as their associated genes before WGD, continue to influence the expression of their original target genes after being separated into different chromosomes by WGD (Figure 2 A, B). It has previously been shown that transcription factor genes are preferentially retained after WGD (vertebrates: Kassahn et al. 2009, paramecia : McGrath et al. 2014, plants: Blanc & Wolfe 2004), however, we did not observe significant enrichment for any GO term, including transcription factors, for genes in inter-chromosomal homeologous trans-eQTL connections (FDR cutoff = 0.05).

To test the impact of rediploidization timing on trans-eQTL evolution we next compared early (LORe, <50 Mya) vs late (AORe, ~100 Mya) rediploidizing regions. Inter-chromosomal homeologous trans connections (category ii) were more frequently retained in LORe compared to AORe regions (chi^2 = 95.6, p-value = 2.2e-16) (**Figure S2**).

Shared trans-eQTLs are associated with higher gene expression correlation

Immediately following a WGD, newly duplicated ohnologs share identical "ancestral" regulatory networks. This includes the same regulatory molecules (both with indirect and direct effects) and identical cis-regulatory motifs where the direct-acting regulatory molecules can interact with promoters and enhancers. Through time, these duplicated regulatory networks will diverge, e.g. through divergence in the cis-regulatory landscape, driving gene regulatory divergence between ohnologs.

To assess the level of ohnolog regulatory divergence, we classified ohnolog pairs into three different 'ohnolog eQTL divergence classes' (**Figure 3A**); (i) ohnologs sharing an eQTL, (ii) ohnologs with distinct eQTLs, (iii) ohnologs where only one copy has an eQTL, and (iv) ohnologs without eQTLs. To reduce the risk of misclassifying ohnolog pairs due to statistical noise in lead SNP selection, we applied fine-mapping with functionally informed priors (**see Methods, Figure S3**). This method improves the

resolution of causal variant identification by estimating the likelihood that a variant directly affects gene regulation rather than merely reflecting statistical association. For each gene, we defined causal eQTLs as fine-mapped cis or trans variants with non-zero maximum posterior inclusion probability (PIP) (**Figure S4**).

Our analysis of the fine-mapped eQTLs revealed that 5% of the ohnolog gene pairs (230/4,267) share at least one fine-mapped eQTL. For those genes, we identified a GO enrichment in functions related to mitochondrial pathways and respiration (Figure S5). Among these 230 ohnolog gene pairs with shared eQTL, 229 pairs were regulated by trans-eQTLs, with only one pair sharing both cis- and trans-eQTLs. Conversely, 1,224 ohnolog pairs (29%) had distinct eQTLs for each gene copy (**Figure 3B**). If shared eQTLs actually reflect that ohnologs have more conserved (and similar) regulatory networks, one expectation is that these ohnologs should have higher gene expression correlation compared to those that do not share eQTLs. To test this, we used the absolute Spearman's correlation coefficient for each ohnolog pair across all samples, to only account for correlation strength. In line with this expectation, ohnolog pairs with shared eQTLs exhibited significantly higher correlation (Spearman's rho: 0.74 ± 0.01) compared to pairs with distinct eQTLs (rho: 0.49 ± 0.01), single-copy eQTLs (rho: 0.50 ± 0.01), or no eQTLs (rho: 0.56 ± 0.01) (**Figure 3B**). Moreover, the distribution of expression correlation for shared eQTL pairs was significantly ($p < 2E-16$) different from the other regulation categories (**Table S1**). These trends were still present when considering only genes within LORe (rho: 0.70; 0.52; 0,53; 0.56, respectively) or AORe (rho: 0.75; 0.49; 0,48; 0.56, respectively) regions. In addition, we found that the mean absolute correlation of observed ohnolog gene pairs (rho: 0.52 ± 0.004) was much higher than for randomly shuffled gene pairs (rho: 0.30 ± 0.003), confirming the strong positive correlation between ohnolog gene pairs.

Because divergence time is associated with regulatory network divergence of gene duplicates (19, 34) we asked if the proportion of shared trans-eQTLs was associated with the timing of rediploidization (22). Indeed, more ohnolog regulatory divergence (i.e. distinct and single classes) was observed in ohnologs residing in early

rediplodized regions (AORe) compared to late rediploidized regions (LORe) (chi^2 = 9.7, p-value = 0.02) (**Figure 3C**).

Taken together we find strong evidence for long term conservation of regulatory networks among ohnologs. However, the majority of ohnologs do not share eQTLs, consistent with previous findings of asymmetric ohnolog expression evolution dominating after the salmonid WGD (17).

Genotype-dependent dosage compensation in ohnolog gene pairs

The vast majority of ohnolog pairs with shared eQTLs had genetic variants impacting gene expression in the same direction in both ohnolog copies. However, for a small proportion of ohnologs (7 ohnologs gene pairs) the alleles of the shared eQTL had opposite effects on the gene expression (**Figure S6**), indicating dynamic dosage compensation mechanisms (**Figure 4A**).

To investigate this phenomenon in more detail we reanalysed the data using less stringent criteria for identifying shared eQTLs across ohnolog pairs showing negative expression correlation. Specifically, some shared eQTLs could have been misclassified as distinct eQTLs due to the selection of different eQTLs as fine mapped eQTL for each gene copy, despite the eQTLs being in linkage disequilibrium and tagging a shared, true causal variant. Out of 53 negatively correlated ohnolog pairs that were previously defined as having distinct trans-eQTLs, 9 pairs were redefined as having shared eQTLs (**Figure S3**), increasing the total number of negatively correlated gene pairs with shared eQTLs to 16 (0.4% of the 4,267 pairs). These are referred to as genotype-dependent dosage compensated ohnologs (**Figure 4A**).

For the majority (11/16 = 69%) of the dosage compensated ohnologs the expression sum of both copies did not differ across eQTL genotypes, while for the others the expression varied from 5% to 19% of the total expression range (**Table S2**).These genes include several highly conserved housekeeping genes such as Eukaryotic Translation Initiation Factor 3 Subunit M (*eif3m*), ribosome binding factor A (*rbfa*),

and N-Acyl Phosphatidylethanolamine Phospholipase D (*napepld*). In most genotype-dependent dosage compensated ohnologs (9/16 = 56%), one copy consistently showed much lower expression compared to the other (e.g tetraspanin8), while in four ohnolog pairs this was not the case (e.g. *skap2* in **Figure 4A**) (**Table S2**). Additionally, we observed a significant enrichment of dosage compensated ohnolog in LORe regions compared to AORe regions (chi^2=25, p-value=4.7E-7, 3.16-fold enrichment) (**Figure 4B**).

Discussion

Conserved trans-regulation across 100 million years

Here, we studied the evolution of gene regulation following WGD using a large eQTL dataset from Atlantic salmon gill. We found that most ohnologs have distinct eQTLs, in line with other studies demonstrating largely asymmetric gene expression following WGD in vertebrates (17, 18). However, we found 229 ohnolog pairs with conservation of trans-regulatory connections, across 100-50 million years. The ohnologs that sharing trans-eQTLs exhibited the highest expression correlation across individual gill samples (**Figure 3**), as has been shown in recently duplicated genes in allopolyploid plants (35). This elevated expression correlation also indicates a higher level of gene-regulatory network conservation compared to other ohnologs. This pattern cannot be exclusively explained by ohnolog divergence time (i.e. rediploidization history), as the difference persists within the divergence time category. Hence, we argue that the shared eQTLs and strong co-expression of these 229 ohnolog gene pairs reflects strong purifying selection on gene regulatory networks. The presence of an important proportion of genes involved in respiration and energy pathway, which are known to be under stabilizing selection (17, 36), further support this hypothesis.

Selection on gene dosage constraints is heavily biased towards LORe regions

A small subset of ohnolog pairs (16 out of 4,267), most of which encode housekeeping genes, shared trans-eQTLs but exhibited negatively correlated expression levels (**Figure 4**). In addition, the expression levels summed across ohnologs and genotypes were often identical or varied only little (**Figure 4**). We interpret this pattern as evidence for genotype-dependent dosage compensation, a pattern expected when duplicated genes are at least partly functionally redundant and there is stabilizing selection on their combined expression level.

Two gene-regulatory mechanisms could in theory give rise to genotype-dependent dosage compensation. One possibility is feedback regulation, in which excessive transcript levels from a gene directly or indirectly repress its own expression. In duplicated genes, such feedback could act asymmetrically, with one copy being more sensitive to repression, and the strength or direction of this effect varying with genotype (37, 38). Alternatively, ohnologs could have evolved opposite regulatory outputs from the same trans-acting factor. In this scenario, a transcription factor might function as a repressor and activator, depending on the cis-regulatory element context. This mechanism is in line with recent work in yeast showing that many transcription factors act as both repressors and activators depending on promoter context (39). Although this seems like a straightforward explanation, the high sequence similarity between duplicates makes it uncertain to occur.

Notably, almost all ohnolog pairs with genotype-dependent dosage compensation signatures were located in late rediploidizing (LORe) regions, where sequence divergence is minimal and the encoded proteins are likely to be identical or nearly identical (21, 40). The phenomenon of genotype-dependent dosage compensation itself is expected to arise when purifying selection acts on the sum of ohnolog gene expression and ohnologs are functionally redundant. However, the enrichment of this phenomenon in LORe is likely to be a consequence of the low “divergence time” of genes in those regions, thus having less chances to accumulate mutations that decouples their regulatory network interactions.

Impact of genome structure evolution on gene regulatory connections after WGD

During rediploidization following WGD, gene regulatory connections are gradually lost or rewired (41, 42). Here, we examined this process in the context of post-WGD karyotype evolution. Consistent with previous studies, we observed a strong enrichment of gene regulatory connections within chromosomes (43, 44), potentially reflecting the facilitation of regulatory interactions by 3D chromatin conformation (45). This enrichment was higher (~16-fold) in chromosomal regions with no karyotype reorganization since the WGD compared to those that have undergone more recent chromosome fusions (~7-fold). One possible explanation for this phenomenon is that post-WGD chromosome fusions disrupted ancestral long-range cis-regulatory interactions between enhancers and promoters. Alternatively, fusions could also interrupt trans-regulatory interactions mediated by transcription factor activity. In humans, interacting proteins, complexes, and pathways tend to be non-randomly co-located within chromosomes (46, 47), a pattern thought to arise from selection maintaining genes with shared functional properties in close physical proximity, and thereby maintain similar regulatory inputs from the surrounding enhancer landscapes. The lower enrichment of regulatory connections between recently fused chromosomal regions may therefore reflect the disruption of ancient hubs of co localized functionally related genes.

An additional pattern of interest is the enrichment of inter-chromosomal homeologous interactions. Although these connections constitute only about 2% of all observed trans-regulatory interactions (209 connections), they occur approximately two-fold more often than expected by chance. A possible explanation is that they represent physical contacts between duplicated chromosomes, potentially maintained by higher-order nuclear architecture since the WGD. Such contacts could preserve ancestral regulatory relationships even after the two homeologous regions were separated onto different chromosomes during rediploidization. The observed enrichment of these connections in LORe regions is consistent with reduced sequence

divergence allowing the retention of ancestral cis-regulatory compatibility. These interactions would then reflect ancient cis-regulatory elements that, after WGD, became positioned on different chromosomes yet continue to exert trans effects on their original target genes. Further investigation of 3D genome conformation data (e.g., Hi-C) to directly assess spatial proximity between homeologous chromosome pairs would be needed to confirm this hypothesis.

Our study provides insights into how WGD shapes gene regulatory evolution, linking genomic duplication events to long-term regulatory dynamics. We identified sixteen dosage-compensated genes under purifying selection, indicating that balanced expression levels have been maintained by evolutionary constraints for over 100 million years. We also found that newly formed regulatory connections are enriched in intra-chromosomal interactions, suggesting an important contribution of 3D chromatin conformation to the emergence of rewired regulatory links after WGD.

Building on these insights into regulatory rewiring, further studies could provide a more comprehensive understanding of these phenomena. The present study relied on RNAseq data from a single tissue, and because gene expression patterns can vary across tissues, incorporating additional tissue types in future work would help determine the degree of tissue or cell-type specificity or generality in these regulatory responses. Moreover, experimental validation of dosage compensation, for example through gene editing, would offer deeper mechanistic insight. This was not feasible in the present study due to the extremely high sequence similarity between LORe-embedded duplicates. The uses of different Atlantic salmon lineages having accumulated sequence divergence could eventually allow the targeting of duplicates for an experimental setup.

Taken together, our results contribute to the understanding of how WGD influences the evolution of trans-regulatory connections and the long-term maintenance of gene dosage compensation.

Material and methods:

Ethical statement

The experiment adhered to EU regulations for the protection of animals used in research (Directive 2010/63/EU). All necessary precautions were taken to reduce pain and discomfort. The Norwegian Food and Safety Authority approved the experiment (FOTS ID: 25658).

Fish samples

For a detailed explanation of the fish rearing process, refer to Gjerde et al. (2024). Briefly, three groups of approximately 3000 juvenile Atlantic salmon (*Salmo salar*) were subjected to different photoperiods: LL (continuous 24-hour light), 12:12 (12 hours light / 12 hours dark), and 8:16 (8 hours light / 16 hours dark). The LL group was maintained under continuous light from hatching until transfer to seawater (SW). The 8:16 group was reared under continuous light until they reached a body mass of approximately 50g, after which they were exposed to a short photoperiod (8L:16D) for 6 weeks, followed by 8 more weeks of continuous light before SW transfer at a body mass of around 100g. The 12:12 group experienced 12 hours of daylight and 12 hours of darkness for 6 weeks, followed by 8 weeks of continuous light prior to SW transfer. Gill tissues were collected one week before seawater transfer.

For this study, we focused on the 906 individuals reared under the LL condition. To perform imputation for these fish, we used genotyping data from a 66K SNP array and whole genome sequencing (WGS) data from their 112 parent fish.

Whole genome sequencing:

Initially, we identified 9.4 million SNP variants through whole genome resequencing of 112 parent fish using the GATK pipeline gatk4-spark:4.3.0.0 (49). After quality filtering with PLINK v1.9 (50) based on Hardy-Weinberg Equilibrium (HWE p-value < 1E-8), minor allele frequency (MAF < 0.01), and genotype missingness thresholds per individual and SNP variant (10% and 5%, respectively), we retained 108 individuals and 5.8 million SNPs for further analysis.

Genotyping array:

Fish were genotyped using a 66K SNP Affymetrix array [Salmow01]. Physical coordinates of array markers were re-assigned according to the newer Atlantic salmon genome reference (GenBank accession: GCA_905237065.2) and variants were quality-controlled using PLINK v1.9 (50, 51). Quality filtering for genotypes used the following thresholds: HWE p-value < 1E-6, MAF < 0.01. Samples with more than 10% of their genotypes missing were removed, as well as variants missing in more than 3% of the total sample. A total of 2970 samples and 48,563 SNP were retained after quality filtering.

Genotype imputation

To improve the SNP coverage for the 906 LL condition fish, we performed genotype imputation using the Beagle 5.2 software (52, 53) leveraging the WGS data from the 112 parent fish. In order to identify SNP with high inference confidence, a five-fold cross-validation method was applied, by dividing the parent fish into five groups. The genotypes missing from the SNP array were masked in 20% of the individuals and imputed using the remaining 80%. Imputation accuracy was assessed as the square of Pearson's correlation (R^2) between the true and imputed genotypes, averaged across the five validation results. This process resulted in the retention of 412,069 SNPs with an $R^2 > 0.80$. After further quality filtering of imputed SNP based on HWE deviations (HWE p-value < 1E-8) using PLINK v1.9, the final imputed dataset consisted of 366,437 SNP markers, in addition to the original 48,563 SNPs from the genotyping array.

RNA-sequencing

For detailed methods of RNA sequencing, see Grønvold et al. (2024). Briefly, gill biopsies were flash frozen, and RNA was extracted using the RNeasy Fibrous Tissue Mini kit (QIAGEN). Libraries were prepared and sequenced with 2x150 bp paired-end Illumina sequencing.

Before transcript quantification, adaptor sequences were removed using fastp version 0.23.2 (55). Read mapping and transcript quantification were conducted with Salmon version 1.1.0 (56), utilizing the Atlantic salmon transcriptome annotation from the Ssal_v3.1 genome assembly (Ssal_v3.1, GCA_905237065.2) from Ensembl. To prevent misalignment of reads to similar but unannotated regions, a transcriptome index file was used as a decoy sequence, applying the selective mapping mode (57). The --keepDuplicates and --gcBias options were used to address high sequence similarity between duplicated genes resulting from the salmonid genome duplication (21) and to correct for fragment-level GC biases, respectively. Gene-level mRNA expression was calculated by summing raw read counts and normalizing them into transcripts per million (TPM) using the R package tximport (58).

eQTL analysis:

We combined genotype and gene expression datasets using QTLTools (version 06 May 2020) (59) to identify eQTLs. Covariates including Father, Mother, Sex, Tank, Weight, Body length, Conditional factor, Smolt status, and ten genomic principal components (PCs) were regressed out to control for genetic relatedness and batch effects. This analysis resulted in the identification of cis-eQTLs (SNP-gene distance ≤ 5 Mb, FDR < 0.05) and trans-eQTLs (SNP-gene distance > 5 Mb or located on different chromosomes, FDR < 0.05) (**Figure 1**).

Categorization of eQTL- gene connections in the context of WGD

WGD-derived synteny regions in Atlantic salmon genome were retrieved from https://salmobase.org/. Each trans-eQTL-gene pair was categorized into one of four groups based on the location of the eQTL and the gene, as detailed in the results section (**Figure 2**). To assess the statistical significance of the observed trans-eQTL connections, we generated a random expectation by shuffling the eQTLs and their associated genes for all identified trans-eQTL pairs. This randomization process was repeated 1,000 times.

To further analyse the differences in retention of inter-chromosomal homeologous connections between AORe and LORe regions, we first calculated the observed connection density by dividing the number of connections by the number of base pairs (bp) covered by each region. We then computed the expected density through 1,000 iterations of random shuffling for both AORe and LORe regions. Finally, a chi-square test was conducted to compare the observed and expected densities of connections between these regions.

Fine-mapping with functional annotation

In the previous eQTL analysis on individual genes (**Figure 2**), we used the lead SNP for each gene. Next, we aimed to determine whether each ohnolog gene pair has a common eQTL, or distinct eQTLs (**Figure 3**). Using the lead SNPs naively could incorrectly suggest that a gene pair is associated with different eQTLs simply due to statistical noise, while they actually share a common genetic basis.

To address potential misclassification, we conducted fine-mapping using SparsePro (60) to identify putatively causal variants for cis and trans eQTLs separately. p-values from the QTLtools output, representing the association between genotypes and gene expression, were converted to Z-scores and used as input for SparsePro. For genes with either cis or trans eQTL SNPs, we allowed up to 10 credible sets (L = 10) in the fine-mapping process.

We applied fine-mapping with functionally informed priors, incorporating exonic regions and non-coding RNA locations from Ensembl Ssal_v3 annotation, and transcription factors retrieved from orthologous transcription factors from Esox lucius. For each gene, we defined its causal eQTL as the fine-mapped cis or trans variant with a non-zero maximum posterior inclusion probability (PIP), following an exploration of different PIP cutoffs for eQTL definition (**Figure S3 and S4**).

Ohnologs expression analysis

Information from ohnologs genes were obtained using synteny information available at https://gitlab.com/sandve-lab/defining_duplicates/. Some gene pairs

(ENSSSAG00000045167/ ENSSSAG00000001778 ; ENSSSAG00000044129/ ENSSSAG00000117896 ; ENSSSAG00000090152/ ENSSSAG00000110471) with no AORe/LORe information were manually assigned to LORe/AORe region using WGD-derived synteny region information. All ohnologs were classified among 4 categories depending on their regulator, as explained in the results section (Figure 3). For each gene duplicate, if the credible sets of two genes have at least one overlapping fine-mapped eQTL with non-zero posterior probabilities, we define the overlapping eQTL as the "shared" causal eQTL. Otherwise, they were classified into the three other categories, depending on if the two copies were regulated by two distinct eQTLs, or only one copy was regulated, or neither copy was regulated. Gene expression data was then used to compute spearman coefficients. We compared the rho of each distribution using absolute Spearman correlation values in order to investigate the strength of the association in expression of the pairs, regardless of its direction. Since not all gene pairs have recorded expression, these analyses were performed on 4267 pairs.

Revisiting the classification of eQTLs for gene pairs with negative expression correlation

We initially identified nine ohnolog gene pairs that shared eQTLs and exhibited negatively correlated expressions, where higher expression of one copy was linked to lower expression of the other. However, upon closer examination of Ensembl and NCBI gene annotations, we discovered some discrepancies in the two databases: in some cases, a single gene in Ensembl corresponded to two distinct genes in NCBI (**Figure S7**). To resolve this, we reassessed the expression correlation using NCBI annotations and found that these ohnologs pairs actually showed positive correlations. Consequently, we excluded these pairs from the set of negatively correlated ohnologs.

This finding prompted us to further investigate the mechanisms of regulatory divergence following whole genome duplication (WGD), seeking to capture additional gene pairs that might have been overlooked. We also hypothesized that our fine-mapping procedure may have been overly stringent in defining shared eQTLs. Specifically, gene pairs might have been misclassified as having distinct eQTLs if the selected credible sets differed, especially when eQTLs were in linkage disequilibrium and tagging a common causal variant. To address this, we revisited the QTLtools results for 53 negatively correlated ohnolog pairs initially categorized as having distinct eQTLs. We reviewed all significant gene-eQTL associations (nominal $p < 0.05$) prior to fine-mapping (59, 60). If the fine-mapped eQTL in one gene copy was also a significant eQTL for the other gene copy, we reclassified that pair as having a shared eQTL for the negative correlation analysis (**Figure S3**). This reanalysis recaptured 9 ohnolog pairs from the "distinct eQTL" category, bringing the total number of shared eQTL pairs with negative correlations to 16.

Gene ontology enrichment

Gene Ontology (GO) enrichment analyses were performed using ShinyGO (V 0.80, http://bioinformatics.sdstate.edu/go/). For each gene set of interest, enrichment was tested using the default hypergeometric framework with Benjamini–Hochberg correction for multiple testing. Terms with an FDR < 0.05 were considered significantly enriched. The background gene universe was defined all registered genes in the default setting, unless specified otherwise. GO enrichment analyses were performed for specific gene sets, including:

(i) highly expressed genes without detectable eQTLs,

(ii) gene overlapped by Inter-chromosomal-homeologous trans eQTL,

(iii) ohnolog pairs sharing at least one fine-mapped eQTL. (Gene set: the 4267 ohnologs pairs, singletons being ignored).

References:

1. M. Sémon, K. H. Wolfe, Consequences of genome duplication. Curr. Opin. Genet. Dev. 17, 505–512 (2007).

2. Y. Moriyama, K. Koshiba-Takeuchi, Significance of whole-genome duplications on the emergence of evolutionary novelties. Brief. Funct. Genomics 17, 329–338 (2018).

3. O. Jaillon, J.-M. Aury, P. Wincker, "Changing by doubling", the impact of Whole Genome Duplications in the evolution of eukaryotes. C. R. Biol. 332, 241–253 (2008).

4. T. E. Wood, N. Takebayashi, M. S. Barker, I. Mayrose, P. B. Greenspoon, L. H. Rieseberg, The frequency of polyploid speciation in vascular plants. Proc. Natl. Acad. Sci. 106, 13875–13879 (2009).

5. S. P. Otto, J. Whitton, POLYPLOID INCIDENCE AND EVOLUTION. Annu. Rev. Genet. 34, 401–437 (2000).

6. K. V. Krasileva, H. A. Vasquez-Gross, T. Howell, P. Bailey, F. Paraiso, L. Clissold, J. Simmonds, R. H. Ramirez-Gonzalez, X. Wang, P. Borrill, C. Fosker, S. Ayling, A. L. Phillips, C. Uauy, J. Dubcovsky, Uncovering hidden variation in polyploid wheat. Proc. Natl. Acad. Sci. 114 (2017).

7. L. Mihajlovic, B. R. Iyengar, F. Baier, I. Barbier, J. Iwaszkiewicz, V. Zoete, A. Wagner, Y. Schaerli, A direct experimental test of Ohno's hypothesis. eLife 13, RP97216 (2025).

8. A. M. Selmecki, Y. E. Maruvka, P. A. Richmond, M. Guillet, N. Shoresh, A. L. Sorenson, S. De, R. Kishony, F. Michor, R. Dowell, D. Pellman, Polyploidy can drive rapid adaptation in yeast. Nature 519, 349–352 (2015).

9. S. A. Teichmann, M. M. Babu, Gene regulatory network growth by duplication. Nat. Genet. 36, 492–496 (2004).

10. T. Hämälä, C. Moore, L. Cowan, M. Carlile, D. Gopaulchan, M. K. Brandrud, S. Birkeland, M. Loose, F. Kolář, M. A. Koch, L. Yant, Impact of whole-genome duplications on structural variant evolution in Cochlearia. Nat. Commun. 15, 5377 (2024).

11. R. V. Rohlfs, R. Nielsen, Phylogenetic ANOVA: The Expression Variance and Evolution Model for Quantitative Trait Evolution. Syst. Biol. 64, 695–708 (2015).

12. J. R. Dimayacyac, S. Wu, D. Jiang, M. Pennell, Evaluating the Performance of Widely Used Phylogenetic Models for Gene Expression Evolution. Genome Biol. Evol. 15, evad211 (2023).

13. R. A. Veitia, Exploring the etiology of haploinsufficiency. BioEssays 24, 175–184 (2002).

14. B. Papp, C. Pál, L. D. Hurst, Dosage sensitivity and the evolution of gene families in yeast. Nature 424, 194–197 (2003).

15. J. A. Birchler, R. A. Veitia, Gene balance hypothesis: Connecting issues of dosage sensitivity across biological disciplines. Proc. Natl. Acad. Sci. 109, 14746–14753 (2012).

16. F. Almeida-Silva, Y. Van De Peer, Whole-genome Duplications and the Long-term Evolution of Gene Regulatory Networks in Angiosperms. Mol. Biol. Evol. 40, msad141 (2023).

17. G. B. Gillard, L. Grønvold, L. L. Røsæg, M. M. Holen, Ø. Monsen, B. F. Koop, E. B. Rondeau, M. K. Gundappa, J. Mendoza, D. J. Macqueen, R. V. Rohlfs, S. R. Sandve, T. R. Hvidsten, Comparative regulomics supports pervasive selection on gene dosage following whole genome duplication. Genome Biol. 22, 103 (2021).

18. F. Marlétaz, P. N. Firbas, I. Maeso, J. J. Tena, O. Bogdanovic, M. Perry, C. D. R. Wyatt, E. De La Calle-Mustienes, S. Bertrand, D. Burguera, R. D. Acemel, S. J. Van Heeringen, S. Naranjo, C. Herrera-Ubeda, K. Skvortsova, S. Jimenez-Gancedo, D. Aldea, Y. Marquez, L. Buono, I. Kozmikova, J. Permanyer, A. Louis, B. Albuixech-Crespo, Y. Le Petillon, A. Leon, L. Subirana, P. J. Balwierz, P. E. Duckett, E. Farahani, J.-M. Aury, S. Mangenot, P. Wincker, R. Albalat, È. Benito-Gutiérrez, C. Cañestro, F. Castro, S. D'Aniello, D. E. K. Ferrier, S. Huang, V. Laudet, G. A. B. Marais, P. Pontarotti, M. Schubert, H. Seitz, I. Somorjai, T. Takahashi, O. Mirabeau, A. Xu, J.-K. Yu, P. Carninci, J. R. Martinez-Morales, H. R. Crollius, Z. Kozmik, M. T. Weirauch, J. Garcia-Fernàndez, R. Lister, B. Lenhard, P. W. H. Holland, H. Escriva, J. L. Gómez-Skarmeta, M. Irimia, Amphioxus functional genomics and the origins of vertebrate gene regulation. Nature 564, 64–70 (2018).

19. Y. Zou, Z. Su, J. Yang, Y. Zeng, X. Gu, Uncovering genetic regulatory network divergence between duplicate genes using yeast eQTL landscape. J. Exp. Zoolog. B Mol. Dev. Evol. 312B, 722–733 (2009).

20. J. You, Z. Liu, Z. Qi, Y. Ma, M. Sun, L. Su, H. Niu, Y. Peng, X. Luo, M. Zhu, Y. Huang, X. Chang, X. Hu, Y. Zhang, R. Pi, Y. Liu, Q. Meng, J. Li, Q. Zhang, L. Zhu, Z. Lin, L. Min, D. Yuan, C. E. Grover, D. D. Fang, K. Lindsey, J. F. Wendel, L. Tu, X. Zhang, M. Wang, Regulatory controls of duplicated gene expression during fiber development in allotetraploid cotton. Nat. Genet. 55, 1987–1997 (2023).

21. S. Lien, B. F. Koop, S. R. Sandve, J. R. Miller, M. P. Kent, T. Nome, T. R. Hvidsten, J. S. Leong, D. R. Minkley, A. Zimin, F. Grammes, H. Grove, A. Gjuvsland, B. Walenz, R.

A. Hermansen, K. von Schalburg, E. B. Rondeau, A. Di Genova, J. K. A. Samy, J. Olav Vik, M. D. Vigeland, L. Caler, U. Grimholt, S. Jentoft, D. Inge Våge, P. de Jong, T. Moen, M. Baranski, Y. Palti, D. R. Smith, J. A. Yorke, A. J. Nederbragt, A. Tooming-Klunderud, K. S. Jakobsen, X. Jiang, D. Fan, Y. Hu, D. A. Liberles, R. Vidal, P. Iturra, S. J. M. Jones, I. Jonassen, A. Maass, S. W. Omholt, W. S. Davidson, The Atlantic salmon genome provides insights into rediploidization. Nature 533, 200–205 (2016).

22. M. K. Gundappa, T.-H. To, L. Grønvold, S. A. M. Martin, S. Lien, J. Geist, D. Hazlerigg, S. R. Sandve, D. J. Macqueen, Genome-Wide Reconstruction of Rediploidization Following Autopolyploidization across One Hundred Million Years of Salmonid Evolution. Mol. Biol. Evol. 39, msab310 (2022).

23. M. Saitou, D. Manousi, J. Van Dalum, A. Striberny, L. Grønvold, C. Brekke, S. A. Boison, J. Kwak, A. V. Ponce De León, B. Gjerde, E. Jørgensen, D. Hazlerigg, S. R. Sandve, A major locus on chromosome 14 impacts developmental variation of Atlantic salmon smoltification. [Preprint] (2025). https://doi.org/10.1101/2025.10.16.682803.

24. H. Akashi, Gene expression and molecular evolution. Curr. Opin. Genet. Dev. 11, 660–666 (2001).

25. H. -c. Wang, Evidence for strong selective constraint acting on the nucleotide composition of 16S ribosomal RNA genes. Nucleic Acids Res. 30, 2501–2507 (2002).

26. C. Morris, D. Cluet, E. P. Ricci, Ribosome dynamics and MRNA turnover, a complex relationship under constant cellular scrutiny. WIREs RNA 12, e1658 (2021).

27. K. Voordeckers, K. Pougach, K. J. Verstrepen, How do regulatory networks evolve and expand throughout evolution? Curr. Opin. Biotechnol. 34, 180–188 (2015).

28. H. Göös, M. Kinnunen, K. Salokas, Z. Tan, X. Liu, L. Yadav, Q. Zhang, G.-H. Wei, M. Varjosalo, Human transcription factor protein interaction networks. Nat. Commun. 13, 766 (2022).

29. K. Van Dyke, S. Lutz, G. Mekonnen, C. L. Myers, F. W. Albert, Trans -acting genetic variation affects the expression of adjacent genes. Genetics 217, iyaa051 (2021).

30. S. Kim, J. Wysocka, Deciphering the multi-scale, quantitative cis-regulatory code. Mol. Cell 83, 373–392 (2023).

31. K. S. Kassahn, V. T. Dang, S. J. Wilkins, A. C. Perkins, M. A. Ragan, Evolution of gene function and regulatory control after whole-genome duplication: Comparative analyses in vertebrates. Genome Res. 19, 1404–1418 (2009).

32. C. L. McGrath, J.-F. Gout, P. Johri, T. G. Doak, M. Lynch, Differential retention and divergent resolution of duplicate genes following whole-genome duplication. Genome Res. 24, 1665–1675 (2014).

33. G. Blanc, K. H. Wolfe, Functional Divergence of Duplicated Genes Formed by Polyploidy during Arabidopsis Evolution[W]. Plant Cell 16, 1679–1691 (2004).

34. D. Dong, Z. Yuan, Z. Zhang, Evidences for increased expression variation of duplicate genes in budding yeast: from cis- to trans- regulation effects. Nucleic Acids Res. 39, 837–847 (2011).

35. F. He, W. Wang, W. B. Rutter, K. W. Jordan, J. Ren, E. Taagen, N. DeWitt, D. Sehgal, S. Sukumaran, S. Dreisigacker, M. Reynolds, J. Halder, S. K. Sehgal, S. Liu, J. Chen, A. Fritz, J. Cook, G. Brown-Guedira, M. Pumphrey, A. Carter, M. Sorrells, J. Dubcovsky, M. J. Hayden, A. Akhunova, P. L. Morrell, L. Szabo, M. Rouse, E. Akhunov, Genomic variants affecting homoeologous gene expression dosage contribute to agronomic trait variation in allopolyploid wheat. Nat. Commun. 13, 826 (2022).

36. F. Zhang, R. E. Broughton, Heterogeneous natural selection on oxidative phosphorylation genes among fishes with extreme high and low aerobic performance. BMC Evol. Biol. 15, 173 (2015).

37. J. A. Birchler, U. Bhadra, M. P. Bhadra, D. L. Auger, Dosage-Dependent Gene Regulation in Multicellular Eukaryotes: Implications for Dosage Compensation, Aneuploid Syndromes, and Quantitative Traits. Dev. Biol. 234, 275–288 (2001).

38. D. M. Bravo-Estupiñan, K. Aguilar-Guerrero, S. Quirós, M. Acón, C. Marín-Müller, M. Ibáñez-Hernández, R. A. Mora-Rodríguez, Gene dosage compensation: Origins, criteria to identify compensated genes, and mechanisms including sensor loops as an emerging systems-level property in cancer. Cancer Med. 12, 22130–22155 (2023).

39. L. Mahendrawada, L. Warfield, R. Donczew, S. Hahn, Low overlap of transcription factor DNA binding and regulatory targets. Nature 642, 796–804 (2025).

40. F. M. Robertson, M. K. Gundappa, F. Grammes, T. R. Hvidsten, A. K. Redmond, S. Lien, S. A. M. Martin, P. W. H. Holland, S. R. Sandve, D. J. Macqueen, Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification. Genome Biol. 18, 111 (2017).

41. S. Varadharajan, S. R. Sandve, G. B. Gillard, O. K. Tørresen, T. D. Mulugeta, T. R. Hvidsten, S. Lien, L. Asbjørn Vøllestad, S. Jentoft, A. J. Nederbragt, K. S. Jakobsen, The Grayling Genome Reveals Selection on Gene Expression Regulation after Whole-Genome Duplication. Genome Biol. Evol. 10, 2785–2800 (2018).

42. R. De Smet, Y. Van De Peer, Redundancy and rewiring of genetic networks following genome-wide duplication events. Curr. Opin. Plant Biol. 15, 168–176 (2012).

43. A. Dean, In the loop: long range chromatin interactions and gene regulation. Brief. Funct. Genomics 10, 3–10 (2011).

44. P. Quintero-Cadena, P. W. Sternberg, Enhancer Sharing Promotes Neighborhoods of Transcriptional Regulation Across Eukaryotes. G3 GenesGenomesGenetics 6, 4167–4174 (2016).

45. M. Di Stefano, F. Di Giovanni, V. Pozharskaia, M. Gomar-Alba, D. Baù, L. B. Carey, M. A. Marti-Renom, M. Mendoza, Impact of Chromosome Fusions on 3D Genome Organization and Gene Expression in Budding Yeast. Genetics 214, 651–667 (2020).

46. R. Cao, J. Cheng, Deciphering the association between gene function and spatial gene-gene interactions in 3D human genome conformation. BMC Genomics 16, 880 (2015).

47. A. Thévenin, L. Ein-Dor, M. Ozery-Flato, R. Shamir, Functional gene groups are concentrated within chromosomes, among chromosomes and in the nuclear space of the human genome. Nucleic Acids Res. 42, 9854–9861 (2014).

48. B. Gjerde, S. A. Boison, D. Hazlerigg, T. Ytrestøyl, T. Mørkøre, E. Jørgensen, A. Striberny, S. R. Sandve, Impact of three light smoltification regimes on performance and genetic parameters of traits in Atlantic salmon. [Preprint] (2024). https://doi.org/10.1101/2024.04.02.587721.

49. G. Van der Auwera, B. O'Connor, Genomics in the Cloud (O'Reilly Media, Inc., 1st edition., 2020).

50. C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, J. J. Lee, Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4, s13742-015-0047–8 (2015).

51. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. De Bakker, M. J. Daly, P. C. Sham, PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am. J. Hum. Genet. 81, 559–575 (2007).

52. B. L. Browning, Y. Zhou, S. R. Browning, A One-Penny Imputed Genome from Next-Generation Reference Panels. Am. J. Hum. Genet. 103, 338–348 (2018).

53. B. L. Browning, X. Tian, Y. Zhou, S. R. Browning, Fast two-stage phasing of large-scale sequence data. Am. J. Hum. Genet. 108, 1880–1890 (2021).

54. L. Grønvold, M. J. Van Dalum, A. Striberny, D. Manousi, T. Ytrestøyl, T. Mørkøre, S. Boison, B. Gjerde, E. Jørgensen, S. R. Sandve, D. G. Hazlerigg, Transcriptomic profiling of gill biopsies to define predictive markers for seawater survival in farmed Atlantic salmon. [Preprint] (2024). https://doi.org/10.1101/2024.08.20.608748.

55. S. Chen, Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. iMeta 2, e107 (2023).

56. R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, C. Kingsford, Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods 14, 417–419 (2017).

57. A. Srivastava, L. Malik, H. Sarkar, M. Zakeri, F. Almodaresi, C. Soneson, M. I. Love, C. Kingsford, R. Patro, Alignment and mapping methodology influence transcript abundance estimation. Genome Biol. 21, 239 (2020).

58. C. Soneson, M. I. Love, M. D. Robinson, Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research 4, 1521 (2015).

59. O. Delaneau, H. Ongen, A. A. Brown, A. Fort, N. I. Panousis, E. T. Dermitzakis, A complete tool set for molecular QTL discovery and analysis. Nat. Commun. 8, 15452 (2017).

60. W. Zhang, H. Najafabadi, Y. Li, SparsePro: An efficient fine-mapping method integrating summary statistics and functional annotations. PLOS Genet. 19, e1011104 (2023).

Acknowledgements

We acknowledge the use of the Orion computing cluster at the Norwegian University of Life Sciences (NMBU). We thank Torgeir Rhoden Hvidsten from NMBU- KBM for his advice on expression network analysis. We thank Jun Soung Kwak from Kongju National University for his help on investigating the possibility to design and experiment to test the dosage compensation effect.

Funding:

This study was supported by the Norwegian Seafood Research Fund (FHF) through project 901589 and Norwegian Research Council (NRC) through project 325874.

Author contribution:

The study was conceived by SRS, DH, CD, MS. Data was generated by DH and SRS. Bioinformatic and computational analyses were conducted by CD, DM, GG, LG and MS. Interpretation of the results were done by all authors. Drafting paper was done by CD, MS, LG, SRS, NJB. Manuscript editing was done by all authors.

Competing interests

The authors declare no competing interests.

Data and Code availability

The Illumina sequencing data from RNAseq and whole genome resequencing are available in the European Nucleotide Archive under the project PRJEB47441. The code and output from bioinformatics pipelines used to generate all the figures and downstream analyses of data in R is available at https://github.com/celiand/eQTL_WGD_salmo

Figures

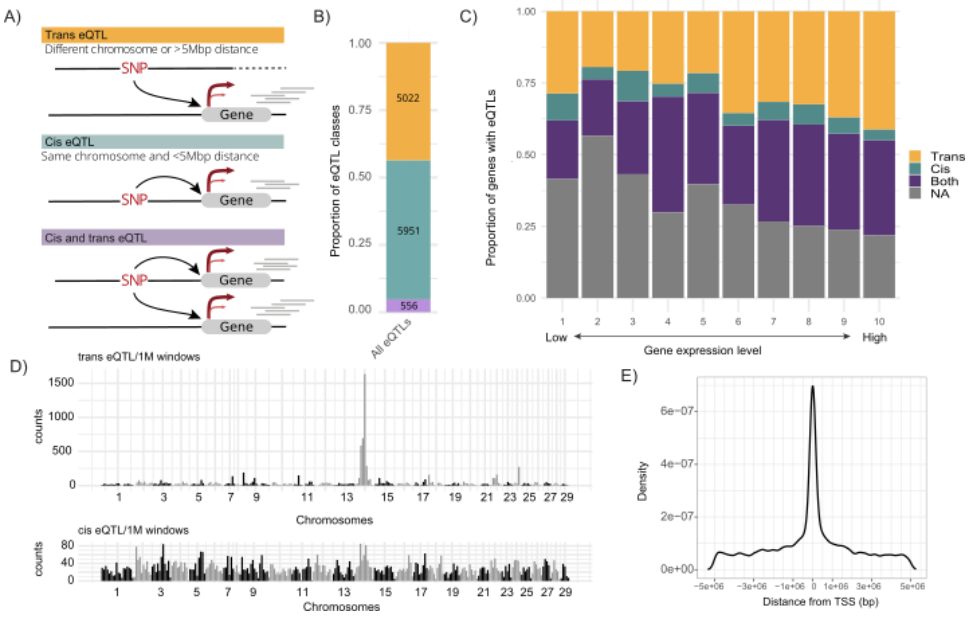


Figure 1: Overview of cis and trans eQTLs and their genomic distribution in Atlantic Salmon.

A: Schematic illustration of different eQTL-gene relationships. A: Cis-eQTLs are defined as when SNP is located within 5 Mb of the gene it regulates. Trans-eQTLs are those where the SNP and gene are located more than 5 Mb apart or on different chromosomes. The third category involves eQTLs that show both cis and trans interactions. **B:** The proportion of identified eQTLs: 49% (5,887) are cis-eQTLs, 46% (5,549) are trans-eQTLs, and 5% (620) show both cis and trans interactions. **C:** Proportion of genes with cis, trans, both, or no eQTLs across 10 groups of genes, ranked by expression levels from low to high. Each group represents 10% of the genes, with group 1 having the lowest expression and group 10 the highest. **D:** Genome-wide distribution of trans-eQTLs (top) and cis-eQTLs (bottom) across chromosomes, divided into 1 Mb windows. Chromosome 14 shows a marked enrichment for trans-eQTLs (chi^2 = 29914, p-value < 2.2E-16), while no such enrichment is seen for cis-eQTLs. E: Density plot of distances between cis-eQTLs and the transcription start site (TSS) of the genes they regulate.

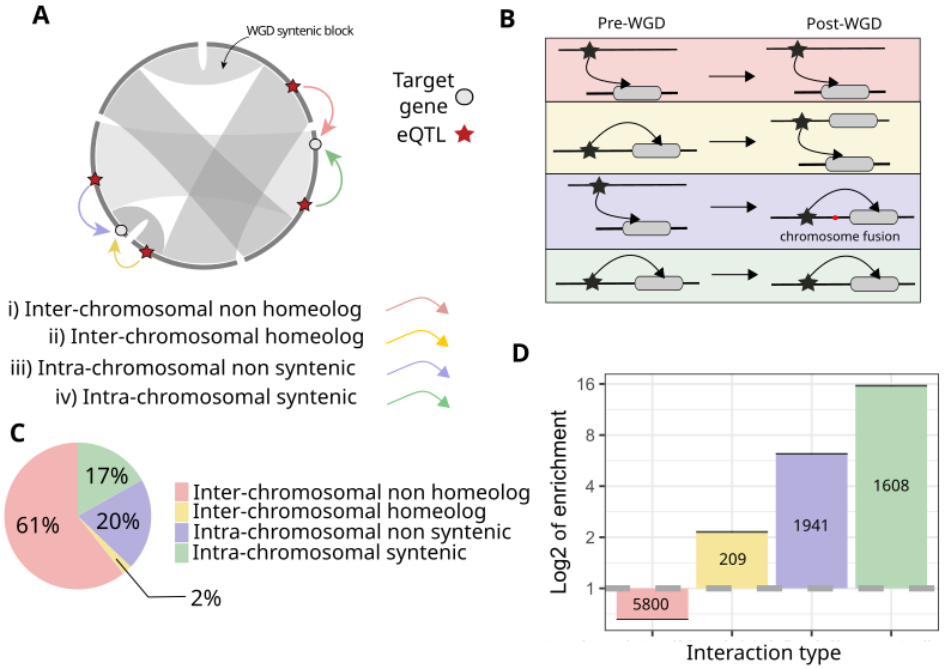


Figure 2. Types of trans-regulatory interactions in the context of WGD and their abundance

A) Four types of trans regulation: i: The eQTL SNP is associated with a gene on another chromosome that is not the homeologous region (inter-chromosome non-homeolog). ii: The eQTL SNP is associated with a gene on the homeologous region on a duplicate chromosome originating from WGD (inter-chromosome homeolog). iii: The eQTL SNP is associated with a gene on the same chromosome outside the syntenic region (intra-chromosome non-syntenic) iv: The eQTL SNP is associated with a gene in the same syntenic region (intra-chromosome syntenic). **B)**. Schematic representation of evolutionary events leading to different types of trans regulation (Left: before WGD, right: after WGD). **C)**. Proportion of the different types of trans regulation in observed connections. **D)**. Enrichment or depletion of each type of trans-regulatory interaction compared to randomly shuffled interactions. Enrichment values above 1 indicate that the interaction type is more frequent than expected by chance (**see Methods**), with the number of observed connections indicated within the bars. Standard deviation is shown in black.

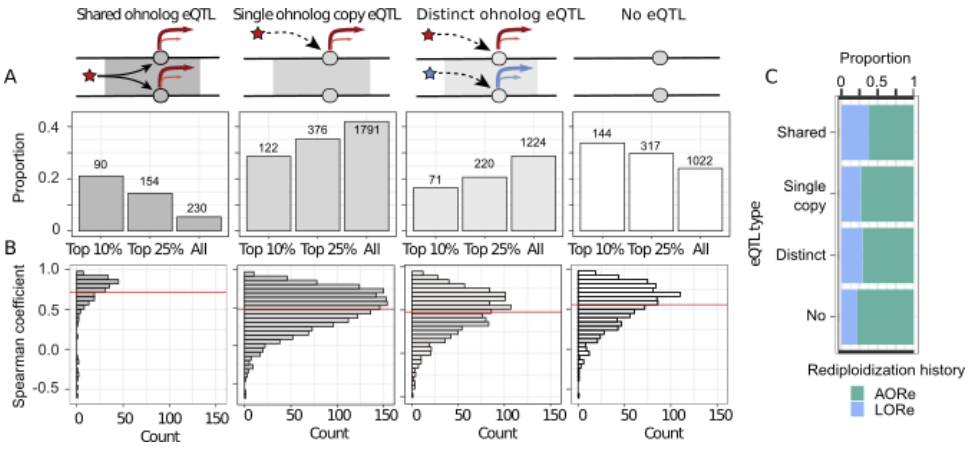


Figure 3: Ohnolog regulation patterns and expression correlations between ohnolog gene pairs

A. Four regulation states are defined: i: Both copies of a gene are regulated by the same eQTL (shared eQTL). ii: One of the two copies is regulated by an eQTL (Single copy eQTL. iii: Both copies are regulated by different eQTLs (distinct eQTL). iv: None of the copies are regulated by an eQTL (no eQTL). The proportion of ohnolog pairs in each regulatory category is shown for all pairs, as well as for the top 10% and 25% of pairs with the highest correlation in gene expression. **B:** Spearman correlation coefficients of gene expression between ohnolog pairs for each regulatory category. The horizontal red line represents the mean correlation. **C:** The proportion of pairs in the AORe (green) and LORe (blue) region depends on their regulatory context.

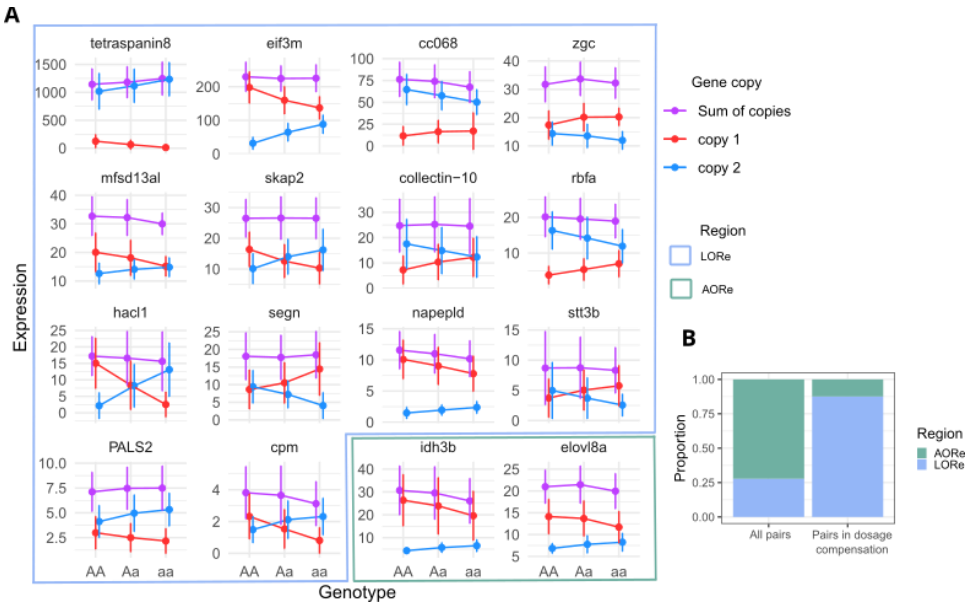


Figure 4: Gene expression levels of dosage compensated ohnolog according to their eQTL genotype

A: Mean expression levels (in TPM) for the 16 ohnolog pairs with shared eQTLs and negative expression correlations, with their standard deviation. The expression is grouped by the genotype of the shared eQTL, which are represented by aa, Aa, and AA. Red lines represent Copy 1, blue lines represent Copy 2, and purple lines represent the sum of both copies' expressions. **B:** Proportion of dosage compensated ohnolog (among the 16 found) in AORe and LORe region versus all gene pairs.

Supplementary materials

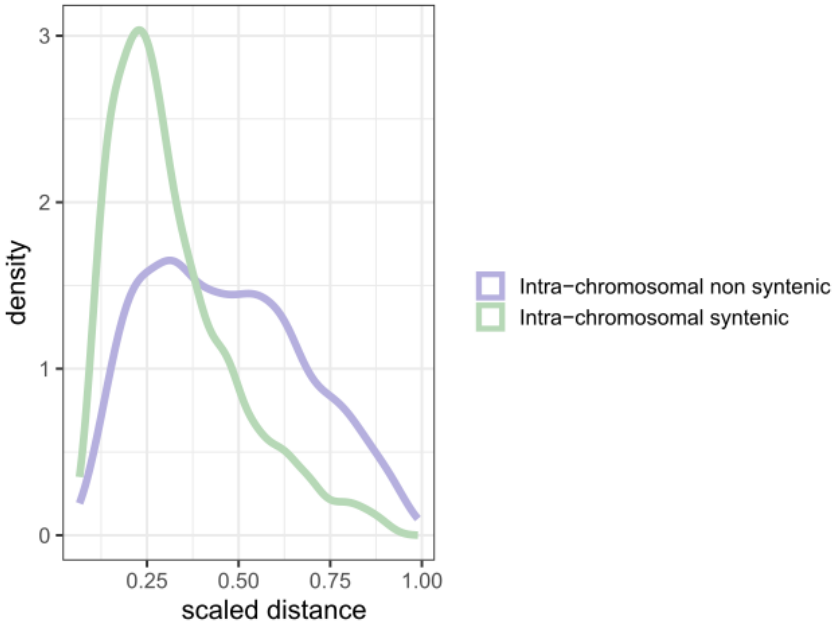


Figure S1: Scaled physical distance between the trans eQTL and the regulated gene

The distance between the eQTL and the gene is scaled by it, dividing by the maximum range possible for the eQTL-gene connection in the region where it's located. Since intra chromosome non syntenic connections can cover a greater range (mean=47 024 214 bp, max=161 086 148 bp) than intra chromosome syntenic connections (mean=34 021 492 bp, max=56 025 461 bp), an identical raw distance will generally be a higher scaled distance in intra chromosome syntenic connections than in intra chromosome non syntenic connections, which could bias our results. However, even with this potential bias, intra chromosome non syntenic connections are larger with scaling (mean=0.47) than intra chromosome syntenic connections (mean=0.28).

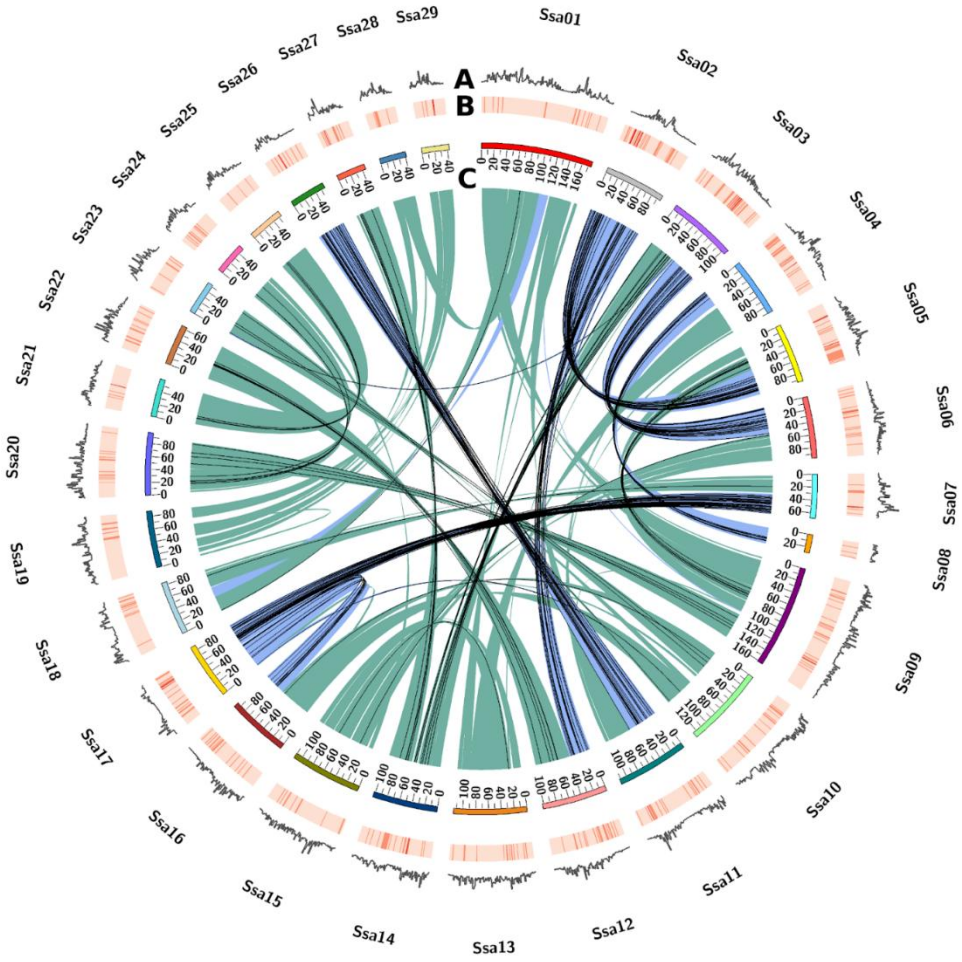


Figure S2: Trans regulatory connections across homeologous regions

This figure shows the distribution of trans-regulatory connections between eQTLs and genes in homeologous regions, regions that were duplicated as a result of WGD event of the Atlantic salmon genome. The outermost ring (**A**) represents the gene density across the genome, with darker areas indicating higher gene concentrations. The middle ring (**B**) shows the proportion of trans eQTLs identified within 1 Mb windows across each chromosome, where orange bars indicate eQTL density. In the inner circle (**C**), blue ribbons represent lineage-specific ohnolog resolution (LORe) regions, while green ribbons mark ancestral ohnolog resolution (AORe) regions. The black lines within the circle illustrate the trans-regulatory connections between

trans-eQTLs and their associated genes, focusing on those connected across homeologous regions. Trans-regulatory connections are more frequently retained in the more recently rediploidized LORe regions compared to the older AORe regions (chi-square test, p = 2.2e-16).

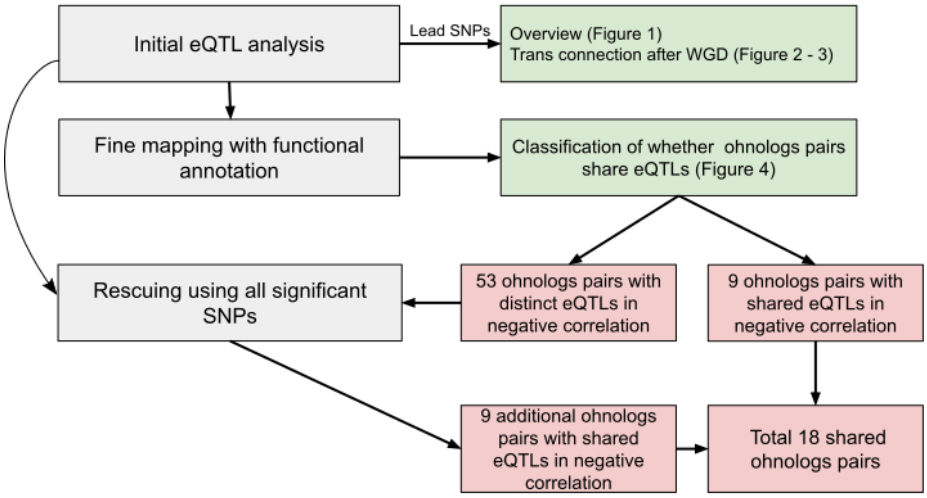


Figure S3: Flow chart of eQTLs analysis performed

In grey: methods used. In green: associated analysis. In red: results obtained.

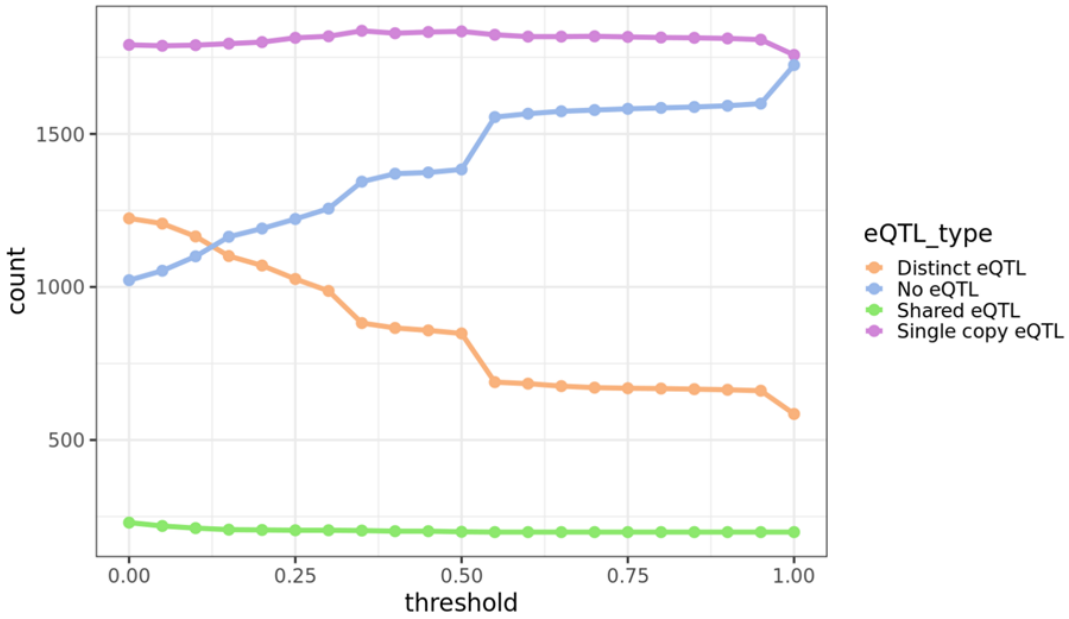


Figure S4: Change in number of pairs in each category depending on PIP threshold

Each dot correspond to a pip threshold tested.

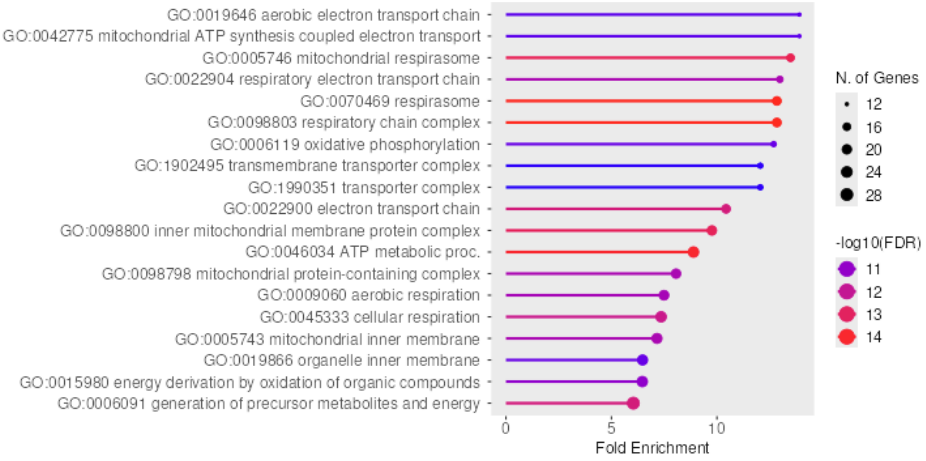


Figure S5: GO enrichment of the 230 ohnologs with shared eQTL

Enrichment of the 230 ohnologs using all gene set available (FDR cutoff=0.05). The 4267 ohnolog pairs were used as background, ignoring singletons.

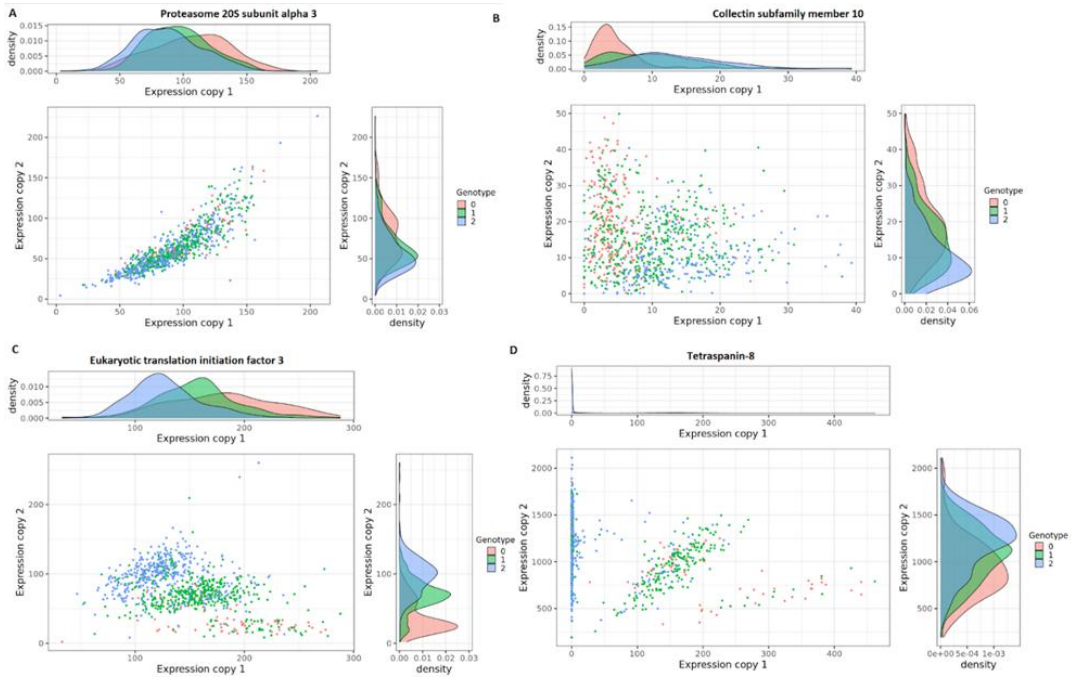


Figure S6: Ohnologs expression according to their shared eQTL genotype

Example of ohnologs pairs with shared eQTL, showing expression of both ohnolog copies for each individual according to their genotype for the eQTL. **A**: An example of an ohnolog pair with positive correlation, **BCD:** 3 ohnologs pairs with negative correlations. We can see 3 distinct groups of expression, representing each of the genotype.

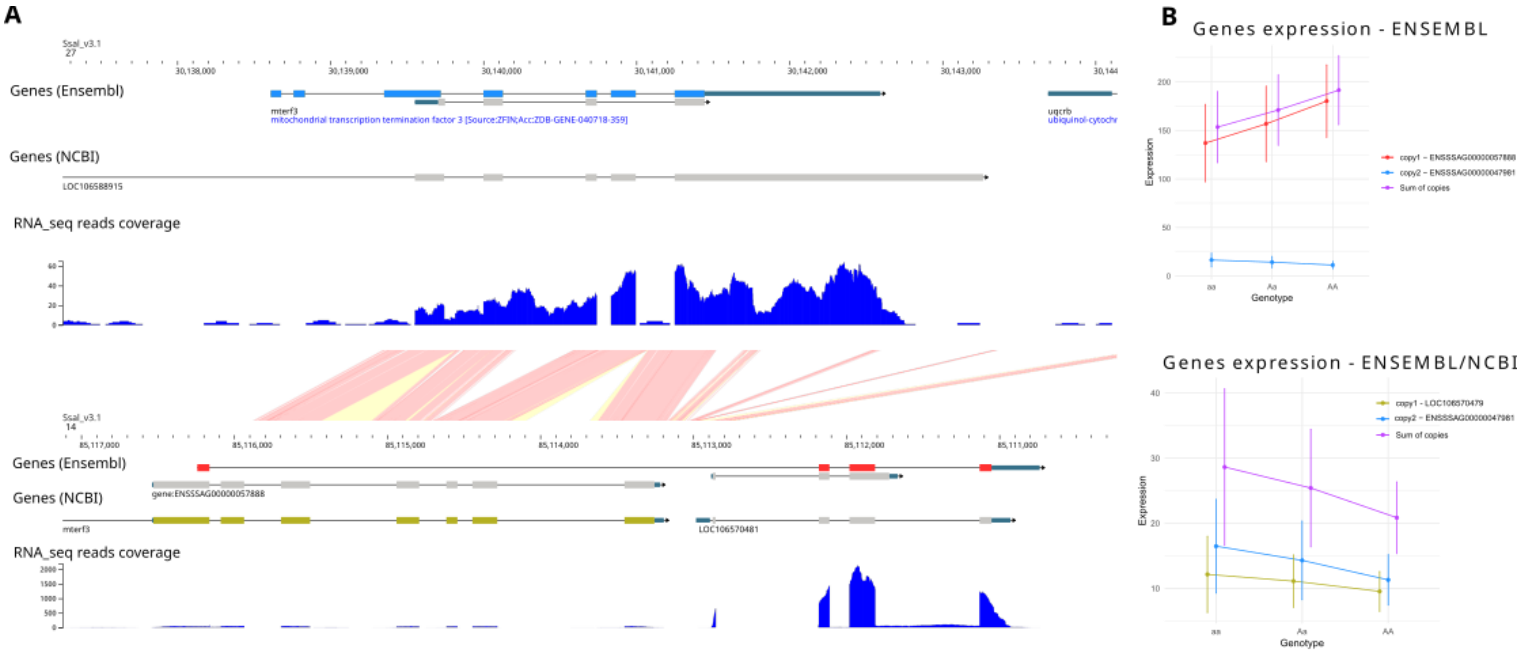


Figure S7: Genome view and corresponding gene expression of mterf3

A: Genome view of chr 27 (top) and chr14 (bottom) at the location of mterf3 ohnologs. **B:** Mean expression levels (in TPM) for mterf3 gene depending on the annotation choose for mterf3 on chr14. Mterf3 is a gene which was found to have a shared eQTL and negative correlation between ohnologs expression. We can observe that on chr14 (bottom), annotation is discordant between NCBI and ENSEMBL. On the expression panel, we can see that the negative correlation is not observed anymore when using NCBI annotation.

Table S1: Comparison of distribution of spearman coefficient

Kolmogorov-smirnov test used. Threshold p-value using Bonferroni correction is 0.0083.

	Shared eQTL	**Distinct eQTL**	**Single copy eQTL**	**No eQTL**
Shared eQTL				
Distinct eQTL	D=0.50 p-val < 2.2E-16			
Single copy eQTL	D=0.48 p-val < 2.2E-16	D=0.03 p-val =0.35		
No eQTL	D=0.37 p-val < 2.2E-16	D=0.15 p-val = 4.5E-12	D=0.14 p-val = 1.5E-11	

Table S2: Gene in dosage compensation with shared eQTL ohnologs

P-values indicate anova test for difference in sum of expression and are corrected for multiple testing using Bonferroni correction. Significant values indicate pairs where the sum of expression is different between genotypes. When significant, the percentage of variation in expression between AA and aa genotypes is indicated in parenthesis. Expression symmetry corresponds to pairs having one copy with consistently lower expression than the others (asymmetric), or pairs with a different

copy being the most expressed between each homozygous genotype (Opposite). Gene pairs that could not be classified in these categories (i.e non significantly different expression between copies in at least one genotype and no opposite pattern) are noted "NS".

Gene	Adjusted p value	Gene 1 ID	Gene 1 location	Gene 2 ID	Gene 2 location	eQTL location	Expression Symmetry
Isocitrate Dehydrogenase (NAD(+)) 3 Non-Catalytic Subunit Beta (*idh3b*)	5.91E-5 *** (15%)	ENSSSAG00000043068	ssa12:50,997,002-51,006,996	ENSSSAG00000054986	ssa22:44,294,869-44,307,005	ssa14:51,370,830	Asymmetric
Major facilitator Superfamily Domain Containing 13A (*mfsd13al*)	0.84	ENSSSAG00000001778	ssa03:77,488,542-77,494,418	ENSSSAG00000045167	ssa06:24,446,097-24,456,254	ssa14:51,354,372	NS
N-acyl phosphatidylethanolamine phospholipase D (*napepld*)	0.0016 *** (12%)	ENSSSAG00000054393	ssa07:63,480,345-63,491,382	ENSSSAG00000117250	ssa17:82,065,545-82,081,353	ssa17:62,041,220	Asymmetric
2-hydroxyacyl-CoA lyase 1 (*hacl1*)	1	ENSSSAG00000058006	ssa05:83,319,427-83,336,253	ENSSSAG00000093636	ssa02:9,545,825-9,559,847	ssa02:14,696,160	Opposite
tetraspanin-8	7.70E-3 *** (9%)	ENSSSAG00000086178	ssa17:81,256,811-81,261,845	ENSSSAG00000087231	ssa07:62,703,781-62,708,799	ssa07:56,592,523	Asymmetric
Carboxypeptidase M (*cpm*)	1.60E-4 *** (18%)	ENSSSAG00000092357	ssa17:77,867,721-77,886,108	ENSSSAG00000109189	ssa07:59,241,376-59,299,087	ssa17:62,357,348	Opposite
zgc	0.25	ENSSSAG00000001739	ssa03:88,263,660-88,266,629	ENSSSAG00000041942	ssa06:13,783,192-13,787,468	ssa06:12,926,921	Asymmetric

Oligosaccharyl transferase complex catalytic subunit B (*stt3b*)	1	ENSSSAG00000002173	ssa05:90,696,619-90,795,848	ENSSSAG00000114008	ssa02:1,960,032-2,083,478	ssa05:68,517,542	Opposite
Src kinase-associated phosphoprotein 2 (*skap2*)	1	ENSSSAG00000009565	ssa02:9,123,961-9,143,339	ENSSSAG00000121915	ssa05:83,717,162-83,735,864	ssa05:68,517,542	Opposite
collectin 10	1	ENSSSAG00000011938	ssa02:4,851,964-4,887,386	ENSSSAG00000108018	ssa05:87,782,557-87,817,196	ssa05:68,654,814	NS
Secretagogin (*segn*)	1	ENSSSAG00000036507	ssa05:88,240,966-88,290,399	ENSSSAG00000116949	ssa02:4,369,425-4,417,790	ssa05:67,379,201	NS
ELOVL fatty acid elongase 8b (*elovl8a*)	2.26E-5 *** (5%)	ENSSSAG00000044129	ssa16:47,926,373-47,953,310	ENSSSAG00000117896	ssa10:12,881,358-12,886,392	ssa14:48,488,679	Asymmetric
Ribosome Binding Factor A (*rbfa*)	1	ENSSSAG00000083156	ssa05:86,193,704-86,196,331	ENSSSAG00000116300	ssa02:6,628,362-6,630,877	ssa02:20,664,047	Asymmetric
cc068	1	ENSSSAG00000090152	ssa02:1,387,837-1,406,608	ENSSSAG00000110471	ssa05:91,371,106-91,389,936	ssa13:57,657,347	Asymmetric
Protein associated with LIN7 2, MAGUK p55 family member (*PALS2*)	0.93	ENSSSAG00000091200	ssa05:84,249,698-84,299,276	ENSSSAG00000121190	ssa02:8,502,544-8,550,092	ssa02:55,822,073	Asymmetric
Eukaryotic translation initiation factor 3, subunit M (*eif3m*)	1	ENSSSAG00000113858	ssa11:38,167,566-38,174,827	ENSSSAG00000120807	ssa26:38,895,331-38,901,886	ssa26:24,868,613	Asymmetric

PAPER II

Parallel Selection in Domesticated Atlantic Salmon from Divergent Founders Including on Whole-Genome Duplication-derived Homeologous Regions

Pauline Buso [1], Célian Diblasi [2], Domniki Manousi [2], Jun Soung Kwak [2], Arturo Vera-Ponce de Leon [2], Kristina Stenløkk [2,3], Nicola Barson [2,*], Marie Saitou [2,*]

[1]IHPE, Université de Montpellier, CNRS, Ifremer, Université de Perpignan Via Domitia, Montpellier, France

[2]Centre for Integrative Genetics (CIGENE), Department of Animal and Aquacultural Sciences, Faculty of Biosciences, Norwegian University of Life Sciences, Ås, Norway

[3]Present address: Department of Forensic Sciences, Oslo University Hospital, 0424 Oslo, Norway

*Corresponding authors: E-mails: nicola.barson@nmbu.no; marie.saitou@nmbu.no.

Accepted: March 26, 2025

Abstract

Domestication and artificial selection for desirable traits have driven significant phenotypic changes and left detectable genomic footprints in farmed animals. Since the 1960s, intensive breeding has led to the rapid domestication of Atlantic salmon (*Salmo salar*), with multiple independent events that make it a valuable model for studying early domestication stages and the parallel evolution of populations of different origins subjected to similar selection pressures. Some aquatic species, including Atlantic salmon, have undergone whole-genome duplication (WGD), raising the possibility that genetic redundancy resulting from WGD has contributed to adaptation in captive environments, as seen in plants. Here, we examined the genomic responses to domestication in Atlantic salmon, focusing on potential signatures of parallel selection, including those associated with WGD. Candidate genomic regions under selection were identified by comparing whole-genome sequences from aquaculture and wild populations across 2 independently domesticated lineages (Western Norway and North America) using a genome-wide scan that combined 3 statistical methods: allele frequencies (F_{ST}), site frequency (Tajima's D), and haplotype differentiation (XP-EHH). These analyses revealed shared selective sweeps on identical SNPs in major histocompatibility complex (MHC) genes across aquaculture populations. This suggests that a combination of long-term balancing selection and recent human-induced selection has shaped MHC gene evolution in domesticated salmon. Additionally, we observed selective sweeps on a small number of gene pairs in homeologous regions originating from WGD, offering insights into how historical genome duplication events may intersect with recent selection pressures in aquaculture species.

Key words: domestication, MHC, artificial selection, aquaculture, ohnolog.

Significance

Domestication rapidly alters the genomic architecture of farmed animals to enhance favorable traits, yet the early genetic responses to domestication are incompletely understood in many systems. By analyzing 2 independently domesticated Atlantic salmon lineages, this study identified shared genomic regions under selection across continents, particularly in immune-related genes. This suggests that independent domestication events can lead to similar genetic changes under artificial selection, highlighting the role of immune-related genes in Atlantic salmon evolution.

© The Author(s) 2025. Published by Oxford University Press on behalf of Society for Molecular Biology and Evolution.
This is an Open Access artide distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/gbe/article/17/4/evaf063/8115854 by guest on 08 January 2026

Introduction

Domestication marks one of the key shifts in human history, tracing back to the enduring bond between early hunter-gatherers and wolves over 15,000 year ago (Vigne 2011). Domesticated plants and animals became essential to human societies for social, nutritional, symbolic, or economic purposes, reflecting the development of multicultural societies that emerged at the end of the last ice age (Larson et al. 2014). Domestication refers to a small, selected fraction from a wild population adapting to a human-controlled environment. This process can change the genetic composition and phenotypes of farmed individuals in response to both intentional and unintentional selection (Wang et al. 2014). Small initial founder populations result in a strong bottleneck and limited genetic diversity. Consequently, the genomic evolution of domesticated species is influenced by strong genetic drift, as well as intentional artificial selection for favorable production traits (Innan and Kim 2004), and unintentional domestication selection resulting from adaptation to the captive environment and human contact (Mignon-Grasteau et al. 2005; Frantz et al. 2020; Huang et al. 2022). Thus, domesticated species carry signatures of these processes in their genomes, providing valuable insights into their domestication history and make them good models for genetic studies of evolution (Wiener and Wilkinson 2011).

Conventional investigations into the evolutionary changes that occurred after domestication began with archaeological records (Diamond 2002). Recent advances using Omics approaches have enabled researchers to trace the geographical and temporal origins of domestication and its early stages. These approaches have revealed the genomic basis of favored traits and selected loci in various species (Frantz et al. 2020), while also shedding light on the unintended consequences of these processes (Glover et al. 2017; Fernandez et al. 2021; Izawa 2022). Domestication processes are often associated with selection pressures on key production traits such as growth, immunity, and age of sexual maturity (Gjederm 1979; Milla et al. 2021; Podgorniak et al. 2022). Artificial selection for such traits and domestication selection may lead to parallel evolution between independent populations exposed to similar selection pressures. This refers to the independent development of genetic adaptations within different populations that affect the organism's phenotype in the same way (Wood et al. 2005; Bailey et al. 2017). For instance, high-altitude adaptation has been observed within independent human populations and their domesticated animals such as dogs, cows, horses, sheep, pigs, and chickens, highlighting a shared genetic response to environmental pressures (Witt and Huerta-Sánchez 2019). Similarly, mammals that have adapted to high-starch diets, including various domesticated animals, have an increased copy number of the amylase gene (*amy*) (Pajic et al. 2019).

Compared to the long history of terrestrial livestock domestication, the domestication of most aquaculture species is relatively recent. Finfish aquaculture expanded rapidly in the early 1980s, while domestication timelines vary across aquaculture species (Houston et al. 2020; Teletchea 2021). Unlike established terrestrial livestock species, aquaculture species present a varied landscape of domestication success, with some species achieving breakthroughs in captive breeding while others continue to rely on wild stock imports. Successful domestications in aquaculture systems are often observed in certain fish families, such as sturgeons, cyprinids (including carp), and salmonids. Notably, some of these families have undergone whole-genome duplication (WGD), which has been linked to increased genomic redundancy, allelic diversity, heterozygosity, and meiotic recombination (Lien et al. 2016; Gundappa et al. 2022). These genomic features may be one factor to enhance tolerance to new mutations and facilitate adaptation to human-mediated environments, while WGD is not essential for successful domestication (Salman-Minkov et al 2016; Baduel et al. 2019; Mihajlovic et al. 2025; Wang et al. 2024). Many successful aquaculture species, including catfish, sea bass, and flatfishes, among others, have achieved domestication without lineage-specific WGD events. Beyond fish, domesticated aquaculture species also include mollusks (e.g. oysters and mussels) and crustaceans (e.g. shrimp and crabs), which do not have known histories of lineage-specific WGD (Houston et al. 2020; Naylor et al. 2021). While WGD-derived genetic redundancy may still contribute to adaptation in some lineages, multiple genomic pathways facilitate domestication across diverse taxa.

The Atlantic salmon (*Salmo salar*) genome underwent a WGD event (Ss4R) ~100 million years ago, leading to the retention of a substantial fraction of duplicated genes (Macqueen and Johnston 2014; Gundappa et al. 2022). Unlike a random loss of duplicate genes, rediploidization followed a structured trajectory, resulting in a nonrandom spectrum of ohnolog divergence levels. These ohnologs are organized into collinear blocks across the genome, reflecting the evolutionary history of this process (Lien et al. 2016; Robertson et al. 2017). In this study, we use "homeolog" to refer to corresponding genomic regions and "ohnolog" to refer to duplicated genes arising from WGD events. While many duplicated genes were subsequently lost, a considerable proportion has been maintained as ohnolog pairs, contributing to functional diversification. Some ohnologs retained their ancestral function, while others underwent subfunctionalization or neofunctionalization, leading to functional differentiation and specialization. In addition to structural retention, the expression of Ss4R-derived ohnologs has evolved under selective pressures, leading to differential expression patterns. Some

Downloaded from https://academic.oup.com/gbe/article/17/4/evaf063/8115854 by guest on 08 January 2026

ohnolog pairs exhibit complementary expression across tissues, while others have diverged functionally (Lien et al. 2016). Additionally, understanding these expression dynamics is crucial for interpreting how Ss4R-derived variation contributes to domestication and aquaculture adaptation, particularly in traits related to metabolism, immune function, and stress response.

Among the aquaculture species with duplicated genomes, Atlantic salmon, has been subject to strong artificial selection and has seen growing economic importance since the 1960s (Cross and King 1983; Ford and Myers 2008; Glover et al. 2017). Atlantic salmon is a widely distributed anadromous species characterized by a strong tendency to reproduce in the rivers of their birth, which encourages genetic divergence among populations (Klemetsen et al. 2003; Lehnert et al. 2019). Wild populations of Atlantic salmon have been split into 4 main phylogeographical lineages, 3 in Europe (East Atlantic, Barents/White Sea, and Baltic) and 1 in North America (West Atlantic) (Bourret et al. 2013; Rougemont and Bernatchez 2018). These lineages exhibit karyotypic variation, with the North American populations displaying a polymorphism in chromosome numbers due to fusions (Brenna-Hansen et al. 2012). Atlantic salmon in Europe and North America have been independently domesticated multiple times from different source populations, with repeated artificial selection directed at similar traits under human-induced environments. This suggests that parallel evolution has likely occurred in domesticated Atlantic salmon populations across both continents, driven by comparable selective pressures.

Despite its recent domestication, studies of Atlantic salmon have revealed rapid genetic changes in response to artificial selection targeting metabolism, disease resistance, and faster growth (Debes et al. 2012; Bicskei et al. 2014; Harvey et al. 2018; Houston and Macqueen 2019; Jin et al. 2020; Bull et al. 2022). Moreover, signatures of domestication selection have also been observed in neurological genes (Bertolotti et al. 2020), causing behavioral alterations, such as increased sensitivity to predation and increased tolerance to stress compared with the wild (Solberg et al. 2020). Especially, the probability of parallel evolution is dependent on the richness of genetic variation, standing genetic variation when this occurs over short time periods (Wood et al. 2005; Bailey et al. 2017; Ebadi et al. 2023). Several studies have investigated parallel selection in Atlantic salmon across different geographic lineages. López et al. (2019) observed parallel evolution at 4 genes, e.g ubiquitin-conjugating enzyme E2F putative (*ube2f*), collagen alpha-1XIII chain (*coda1*), autism susceptibility 2 protein-like (*auts2-like*), and transient receptor potential cation channel subfamily M member 3-like (*trpm3-like*) between Scottish and North American populations (i.e. Canada). However, Naval-Sanchez et al. (2020) did not observe any overlap among selected loci between populations of European and North American origin, highlighting the need for further research to clarify the role of parallel selection in Atlantic salmon domestication.

Here, we investigated the genomic responses of independent Atlantic salmon lineages to recent domestication and artificial selection. By examining genomic divergence between wild and aquaculture populations in Western Norway and North America, we aimed to: (i) identify loci and pathways involved in the early stages of domestication, (ii) explore parallel evolution across lineages, and (iii) assess the impact of the salmonid-specific WGD on domestication responses. Using whole-genome sequencing and multiple selections scan methods (F_{ST}, Tajima's D, and XP-EHH), we identified loci under selection, including shared loci between the 2 lineages, with a few examples of parallel selection on ohnolog pairs (i.e. genes retained in duplicate from the salmonid-specific WGD).

Results

Population Genetic Structure

We first examined the genetic structure of the sequenced populations. We compared the genetic patterns of aquaculture samples from MOWI (Norwegian) with their wild counterparts from 17 rivers in West Norway, the region from where this domesticated lineage originates. Similarly, we analyzed aquaculture samples from 3 independently domesticated North American lineages and wild samples from 8 rivers in the Gulf of St. Lawrence for comparison (Fig. 1a and b). Principal component analysis (PCA) (Fig. 1c) based on 148,433 SNPs derived from whole-genome sequencing showed 4 distinct clusters: (i) Norwegian aquaculture, (ii) wild Norwegian, (iii) North American aquaculture from the St John and Penobscot rivers and wild North American, and (iv) North American aquaculture from Gaspe. The first 2 principal components (PCs) jointly accounted for 64.59% of the total variance, with PC1 (53.85% of variance) separating the East Atlantic (North American) and West Atlantic (West Norwegian) lineages, and PC2 (10.74%) distinguishing between wild and aquaculture populations in Norway, while we did not observe a clear separation between wild and aquaculture populations in North America (supplementary fig. S2, Supplementary Material online). In agreement with the PCA, ADMIXTURE analysis (Alexander et al. 2009; Alexander and Lange 2011) also highlighted a clear separation between the 2 geographical regions. At $K = 6$, where the cross-validation error plateaued (supplementary fig. S3, Supplementary Material online), each phylogeographical lineage showed distinct ancestry, with varying extents of shared ancestry among wild and aquaculture individuals from the same lineage (Fig. 1d). Despite recent domestication from wild populations (Gjedrem et al. 1991; Glover et al. 2009;

Downloaded from https://academic.oup.com/gbe/article/17/4/evaf063/8115854 by guest on 08 January 2026

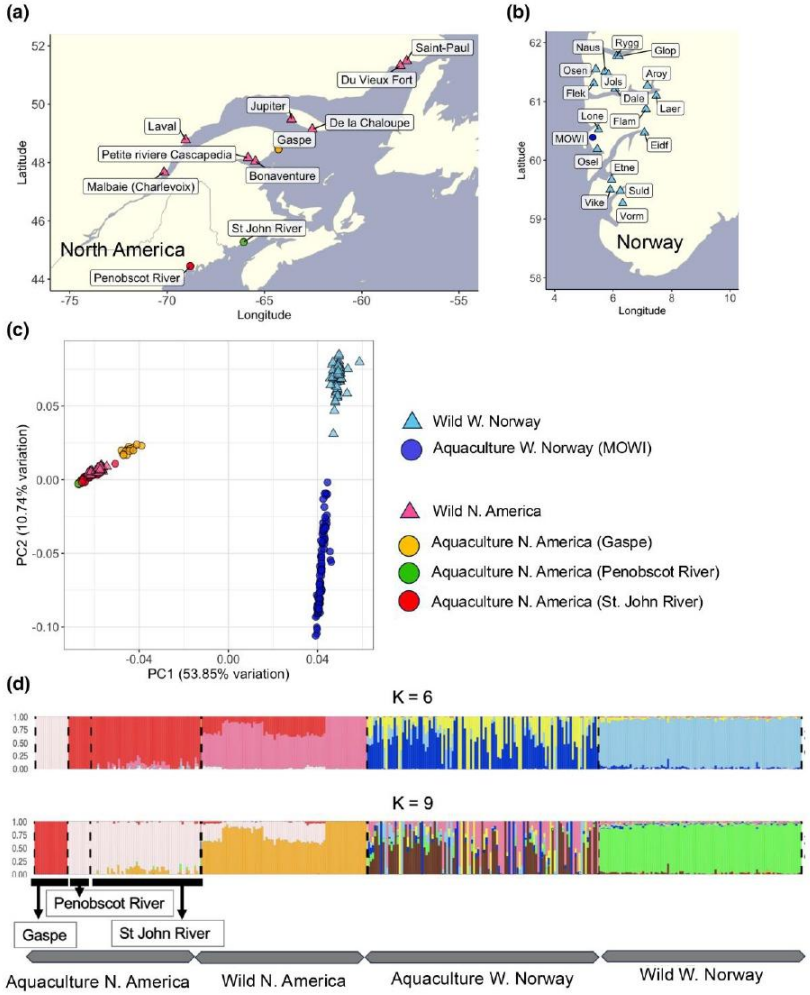


Fig. 1. Sample distribution and genetic structure of Atlantic salmon lines from western Norway and North America. a) Geographic distribution of Atlantic salmon population in North America. Aquaculture samples were used from Gaspe (Gaspe, $n = 16$), St John River ($n = 53$), and Penobscot River ($n = 11$). Wild samples came from 8 rivers: Malbaie Charlevoix ($n = 10$), Bonaventure ($n = 10$), Petite riviere Cascapedia ($n = 10$), Laval ($n = 10$), De la Chaloupe ($n = 10$), Jupiter ($n = 10$), Du Vieux Fort ($n = 10$), and Saint-Paul ($n = 10$). b) Geographic distribution of Atlantic salmon population from Western Norway. Aquaculture samples come from MOWI Norwegian breeding line ($n = 112$). Wild counterparts were collected among 17 rivers: Vorma (Vorm, $n = 2$), Suldalslågen (Suld, $n = 10$), Vikedalselva i Vindafjord (Vike, $n = 10$), Etneelva (Etne, $n = 2$), Eidfjordvassdraget (Eidf, $n = 2$), Oselva i Os (Osel, $n = 2$), Loneelva i Osterøy (Lone, $n = 2$), Flåmselva (Flam, $n = 10$), Lærdalselva (Laer, $n = 2$), Årøy (Aroy, $n = 10$), Daleelva (Dale, $n = 2$), Flekkeelva (Flek, $n = 2$), Nausta (Naus, $n = 10$), Jølstra (Jols, $n = 10$), Oselwassdraget (Osen, $n = 10$), Ryggelva (Rygg, $n = 10$), and Gloppenelva (Glop, $n = 2$). c) Scatterplot of PCA. PC1 and PC2 were calculated for all Atlantic salmon samples. d) Population genetic structure analysis. All samples were clustered into 6 and 9 genetic components ($K = 6$ and $K = 9$) according to the lowest cross-validation error.

Downloaded from https://academic.oup.com/gbe/article/17/4/evaf063/8115854 by guest on 08 January 2026

Bradbury et al. 2022), the Norwegian aquaculture population demonstrated higher genetic heterogeneity than its wild counterpart, possibly reflecting the admixed origin from multiple wild populations. Conversely, North American aquaculture and wild populations were genetically more similar, except for the Gaspe individuals, which were largely distinct, with limited representation in wild populations. This reflects more recent coancestry among North American individuals resulting from the later development of aquaculture in Eastern North America. We also observed a small amount of admixture from the wild East Atlantic group in both wild and aquaculture North American populations, reflecting known past hybridization events that have occurred both through natural secondary contact and as a result of escaped aquaculture individuals with admixed European and North American ancestry (Bradbury et al. 2022).

We delineated our study populations into 4 main groups for the following analysis: West Norwegian Aquaculture Group, Wild West Norwegian Group, North American Aquaculture Group, and Wild North American Group. This grouping was based on the tight PCA clustering of all North American populations and the smaller variance among wild Norwegian populations on PC2 than within the MOWI aquaculture strain. The groups, therefore, reflect both the continent-scale geographical origin, Western Norway versus North America, and the selective environment, those that are wild versus those that have undergone domestication.

Identification of Putatively Selected Regions, Genes, and SNPs

The genome scans to identify potential genomic regions under artificial selection in domestic strains were carried out on 1,139,312 SNPs for Tajima's D and F_{ST} and 1,053,324 SNPs for XP-EHH, depending on the filtering requirements (see Methods). Tajima's D revealed 1,053 windows in the Norwegian aquaculture population and 1,032 windows in the North American aquaculture populations that are putatively under selection. Regarding F_{ST}, the analysis identified 3,177 windows in the West Norwegian pair and 3,088 in the North American pair as potentially under divergent selection. Finally, XP-EHH identified 1914 SNPs in Western Norway and 1982 in North America that are under putatively recent directional selection (Fig. 2).

We then asked if there were any regions or SNPs suggested as under selection by multiple methods. For the pairs in Western Norway and North America, we identified 8 and 2 windows, respectively, where regions were suggested to be non-neutral by both the F_{ST} and Tajima's D metrics. Similarly, there are 104 and 339 shared SNPs suggested to be non-neutral by both F_{ST} and XP-EHH in Western Norway and North America, respectively. Meanwhile, no SNPs overlapped between Tajima's D results and XP-EHH results on either continent. The limited overlap among loci detected as outliers between different methods could be because each method is sensitive to detecting selection over different timescales (Oleksyk et al. 2010). Lack of overlapped loci among methods has been reported previously for domesticated Atlantic salmon (López et al. 2019).

To assess the statistical significance of overlapping genomic sweep regions identified by multiple metrics, we performed random size-matched window sampling simulations ($n = 1000$). For Tajima's D and F_{ST} comparisons, the observed overlap fell below the bottom 1% of the simulated distribution. In contrast, for XP-EHH and F_{ST}, the observed overlap exceeded the maximum value in the simulations (supplementary fig. S4, Supplementary Material online). These findings suggest that Tajima's D and F_{ST} capture selection signatures at different evolutionary timescales, leading to lower-than expected overlap, whereas XP-EHH and F_{ST} tend to highlight regions where selective sweeps and resulting population differentiation cooccur.

To infer which functional groups of genes have been subjected to selection, we identified genes that overlap with windows or SNPs suggested to be non-neutral by multiple methods, using the genes from the Ensembl v111 annotation of the Atlantic salmon reference genome. For these genes, we conducted Gene Ontology (GO) enrichment analysis (Table 1). The "Focal adhesion (false discovery rate, FDR = 0.015)" category has been detected in Western Norway from the intersect F_{ST}/Tajima's D. No GO categories were detected in other groups of genes detected by multiple methods. Therefore, we expanded our investigation to include results from XP-EHH alone, which can detect higher resolution and more recent directional selection. For the genes that contain SNPs identified by XP-EHH, the categories "semaphorin receptor binding (FDR = 0.01)" and "organelle organization (FDR = 0.013)" were enriched in Western Norway, while the category "intracellular protein-containing complex (FDR = 0.0077)" was enriched in North America. A complete list of genes overlapping with XP-EHH outliers is available in Supplementary material (supplementary table S3, Supplementary Material online).

Parallel Selective Sweeps on the Shared SNPs in the MHC Locus in 2 Atlantic Salmon Lineages

We assessed parallel genomic patterns by comparing candidate positively selected XP-EHH SNPs of aquaculture groups across continents, as this method was likely to be most specific to recent divergent selection. We found 82 SNPs with extended haplotypes in both aquaculture lineages (Fig. 3a). Of these, 58 SNPs fell within the *tapbp* gene (ENSSSAG00000040723, 14:64331687-64338464) and the rest, 24 SNPs fell within the adjacent *daxx_chr14* gene (ENSSSAG00000040701, 14:64304883-64330023) on chromosome 14. Both genes are located into the major

Downloaded from https://academic.oup.com/gbe/article/17/4/evaf063/8115854 by guest on 08 January 2026

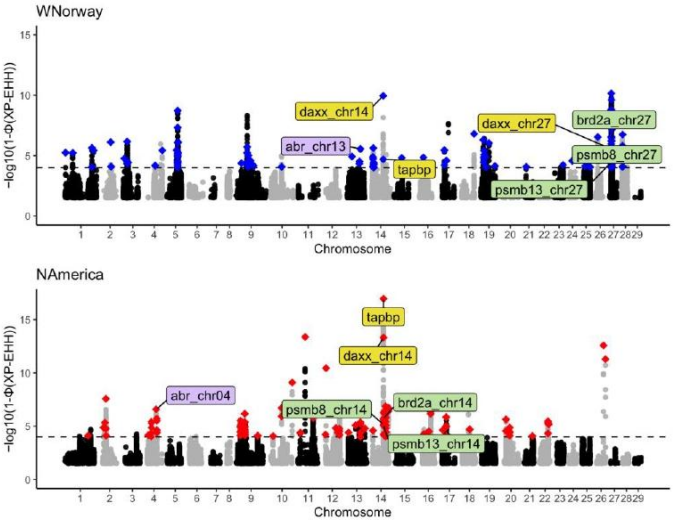


Fig. 2. Cross-population test on extended haplotype homozygosity (XP-EHH). Each dot is SNPs, while the *x* axis displays the position across the genome. The *y* axis represents the −log10(*P*-value the XP-EHH (cross population extended haplotype homozygosity) values. Diamonds indicate SNPs with significantly high XP-EHH values that overlap genes in aquaculture populations from Western Norway and North America, respectively. We adopted −log10(*P*-value) = 4 as the cutoff (dashed line) for significant XP-EHH values (Materials and Methods). Gray and black dots are other tested SNPs. Only SNPs with a -log10(*P*-value) >1 were plotted to ensure the computational efficiency. Labeled genes are genes with exact same SNPs under parallel selection (yellow) or selection on the homeologous regions, with matching colors in both populations showing the WGD-derived gene pairs. The suffix denotes the chromosomal location of the copy. The entire SNP list with significant XP-EHH is available in supplementary table S1, Supplementary Material online.

Table 1 GO enrichment results on genes under putative selection

FDR	nGenes	Fold enrichment	Pathways	Population	Tests
0.015	3	13.8	Focal adhesion	W. Norway	Intersect F_{ST}/Tajima D
0.01	4	27.8	Semaphorin receptor binding	W. Norway	XPEHH
0.013	14	3.6	Organelle organization	W. Norway	XPEHH
0.0077	6	11.2	Intracellular protein-containing complex	N. America	XPEHH

For genes that overlap with windows or SNPs suggested to be non-neutral by multiple methods, or XPEHH alone, we conducted GO enrichment analysis using ShinyGo v0.77 (Ge, Jung and Yao 2020). Categories are shown when the enrichment FDR was < 0.05, Fold Enrichment > 2, and the GO category contained ≥3 our input genes (nGenes). FDR is adjusted from the hypergeometric test. Fold Enrichment is defined as the percentage of genes in the input list belonging to a pathway, divided by the corresponding percentage in the background genes (Ge, Jung and Yao 2020).

histocompatibility complex (MHC) class IA region. Notably, 5 of the shared SNPs are missense SNPs. To check if this result reflected genotyping errors owing to the high diversity and duplicated nature of the MHC region, we manually inspected the SNPs using the BAM files in IGV (Robinson et al. 2011). We assessed that these SNPs are likely real (supplementary fig. S6, Supplementary Material online) and are reliably genotyped.

Variation in the MHC region is known to be maintained by long-term balancing selection as observed as trans-species polymorphisms (Tsukamoto et al. 2012; Teixeira et al. 2015). Given that variation in the MHC region has been maintained since before the divergence of the lineages, we would expect high Tajima's D values in this region. To investigate it, we conducted haplotype analysis and calculated Tajima's D on this MHC region on chromosome 14 and random, size-matching 500 control regions in the 4 groups. Haplotype clustering using the 125 MHC SNPs on chromosome 14 (chr14:64326415-64339381) suggested that there are 2 independent selective sweeps in the West Norwegian aquaculture group and North American Aquaculture group and that high diversity in

Downloaded from https://academic.oup.com/gbe/article/17/4/evaf063/8115854 by guest on 08 January 2026

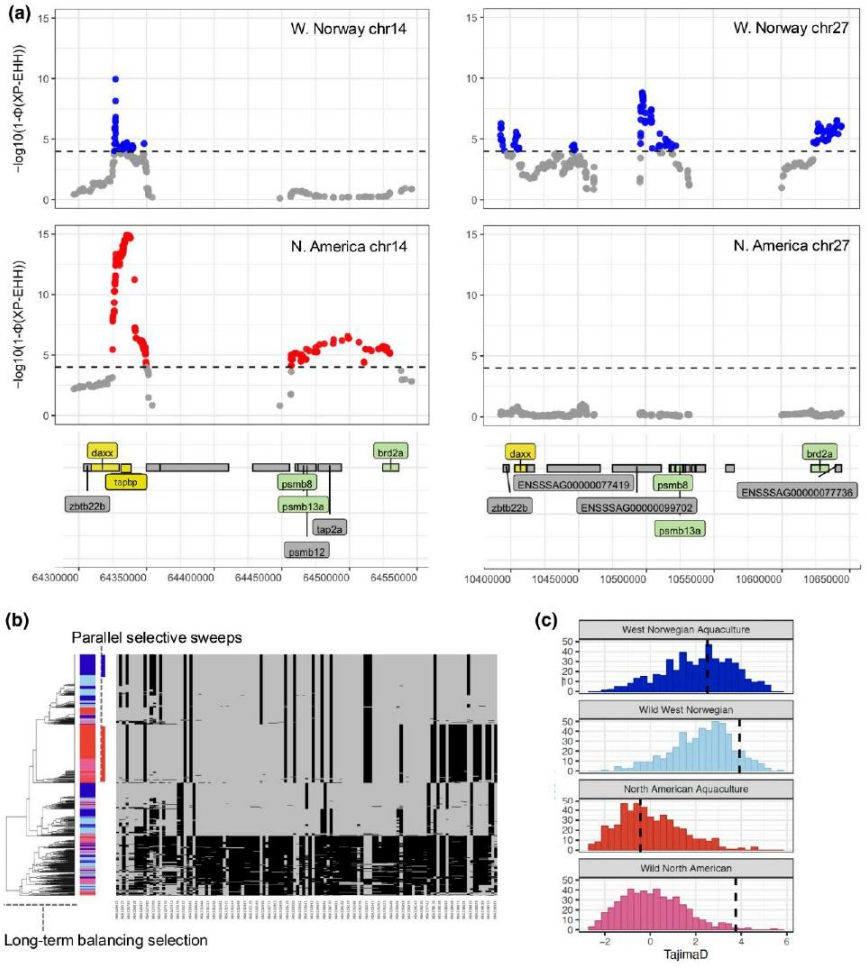


Fig. 3. The MHC region under parallel selective sweeps in both continents. a) Zoomed-in view of XP-EHH results on the MHC loci on chromosomes 14 and 27, exhibiting parallel selective sweeps. The *x* axis shows the genomic position, and the *y* axis represents the −log10(*P*-value) of the XP-EHH (cross population extended haplotype homozygosity) values. Candidate SNPs with significantly high XP-EHH values (above the threshold −log10(*P*) = 4, indicated by the dashed line) are shown, with matching SNPs across populations indicating parallel selection signals. Genes under parallel selection or in homeologous regions with shared selective signals are highlighted accordingly, and genes with signals specific to one population are also labeled. b) Haplotype pattern in the MHC region (126 SNPs from chr14: 64326415-64339381), with each column representing a SNP and each row a phased haplotype from the four population groups. The group is labeled accordingly. c) Distribution of Tajima's *D* values for each population, shown as histograms. The vertical dashed line indicates the observed Tajima's *D* value at the MHC locus. Neutral expectations are derived from 500 randomly sampled size-matched regions per population.

Downloaded from https://academic.oup.com/gbe/article/17/4/evaf063/8115854 by guest on 08 January 2026

Table 2 Genes under parallel selection based on the XP-EHH test

Gene ensembl ID	Gene name	XP-EHH signatures	Region	Chr	Start	End
ENSSSAG00000040701	daxx_chr14	Both_Populations	MHC	14	64,304,883	64,330,023
ENSSSAG00000040723	Tapbp	Both_Populations	MHC	14	64,331,687	64,338,464
ENSSSAG00000093386	tap2a	NAmerica	MHC	14	64,476,753	64,493,938
ENSSSAG00000085704	psmb13a_chr14	Parallel_NAmerica	MHC	14	64,463,937	64,467,854
ENSSSAG00000041842	psmb8b_chr14	Parallel_NAmerica	MHC	14	64,459,783	64,462,906
ENSSSAG00000042614	brd2a_chr14	Parallel_WNorway	MHC	14	64,524,519	64,536,467
ENSSSAG00000077402	daxx_chr27	Parallel_WNorway	MHC	27	10,402,918	10,411,454
ENSSSAG00000077187	tcf19	WNorway	MHC	27	10,323,665	10,328,897
ENSSSAG00000077419	Uba	WNorway	MHC	27	10,426,981	10,465,906
ENSSSAG00000078726	Vhsv	WNorway	MHC	27	10,801,247	10,804,941
ENSSSAG00000078638	vps52	WNorway	MHC	27	10,789,984	10,840,875
ENSSSAG00000077397	zbtb22b	WNorway	MHC	27	10,394,411	10,398,690
ENSSSAG00000077772	col11a2	WNorway	MHC	27	10,645,732	10,694,689
ENSSSAG00000077561	psmb13a_chr27	Parallel_WNorway	MHC	27	10,521,614	10,527,869
ENSSSAG00000085574	psmb8a_chr27	Parallel_WNorway	MHC	27	10,518,241	10,521,084
ENSSSAG00000077713	brd2a_chr27	Parallel_NAmerica	MHC	27	10,621,388	10,634,539
ENSSSAG00000048882	abr_chr13	Parallel_WNorway	…	13	77,743,389	77,966,280
ENSSSAG00000001855	abr_chr4	Parallel_NAmerica	…	4	56,687,253	56,919,875

We described the gene name in this table if the gene or duplicated gene pair is overlapped with 1 or more outlier SNP(s) in XP-EHH test in the Western Norwegian aquaculture group and North American aquaculture group. Start and End refer to the starting and ending locations of genes on the chromosome, respectively, based on the Ssa v.3.1 annotation on Ensembl. Duplicated genes, gray shade (see also Figs. 2 and 3) are defined according to Gillard et al. 2021.

the haplotypes has been maintained in Atlantic salmon since before the lineage divergence between the lineages of North America and Europe (Fig. 3b). In the wild North American group, the Tajima's D value for this region was at the 1st percentile compared to the control regions, while the Tajima's D value for this region was in the top 56% in the North American aquaculture group. In Western Norway, the Tajima's D value for this region was at the 43rd percentile in the aquaculture group and at the 10th percentile in the wild group (Fig. 3c). These results suggest that the wild groups show signs of long-term balancing selection in the MHC region on chr14, while the aquaculture groups appear to have undergone selective sweeps since domestication began. Similar trends were observed in the 2 continents; however, likely due to demographic influences, the distribution of Tajima's D values in the control regions is generally higher in Western Norway compared to those in North America (Fig. 3c).

Parallel Signatures of Selection in Alternate Ohnolog Copies Across Lineages

To further investigate if and how the salmonid-specific WGD contributes to the domestication of Atlantic salmon genetically, we examined whether one of the gene pairs from salmonid-specific WGD has adaptive signatures in the Western Norway aquaculture group while the other gene pair has adaptive signatures in the North American aquaculture group. By investigating XP-EHH outlier SNPs in genic regions, we detected 4 loci with ohnologous genes (i.e. genes duplicated from the WGD) under aquaculture-specific selective sweeps, 1 containing potentially multiple sweep signals. In addition to the *daxx* gene on chromosome 14, which was identified as a gene under selection in both lineages, its ohnolog on chromosome 27 (ENSSSAG00000077402) was identified as under selection in the Western Norwegian aquaculture group. In addition, we identified 3 ohnolog gene pairs under parallel domestication-related sweeps across the 2 distinct geographical regions: the bromodomain-containing 2a (*brd2a*) and proteasome 20S subunit beta 13a (*psmb13*) genes within 100 kb of *daxx* on chromosome 14 in the North American aquaculture group and chromosome 27 in the Western Norwegian aquaculture group, and ABR activator of RhoGEF and GTPase (*abr*) on chromosome 4 in the North American aquaculture group and chromosome 13 in the Western Norwegian aquaculture group (Figs. 2 and 3a and Table 2). The *brd2* gene has previously been reported as under positive selection in Atlantic salmon in aquaculture in Canada compared to their wild counterparts (López et al. 2019), which aligns with its identification as part of the selective sweep in this study.

In addition to the shared genes above, the MHC class I region (Lukacs et al. 2007, 2010) also contains salmonid-lineage-specific gene copies with overlapping XP-EHH outliers. In the Western Norwegian aquaculture group, these include *uba*, *vhsv*, *col11A2*, *tcf19*, *zbtb22b*, *psmb8a*, and *vsp52* on chromosome 27. In contrast, *the psmb8b* and *tap2a* genes on the homeolog chromosome 14 show selection signatures in the North American

Downloaded from https://academic.oup.com/gbe/article/17/4/evaf063/8115854 by guest on 08 January 2026

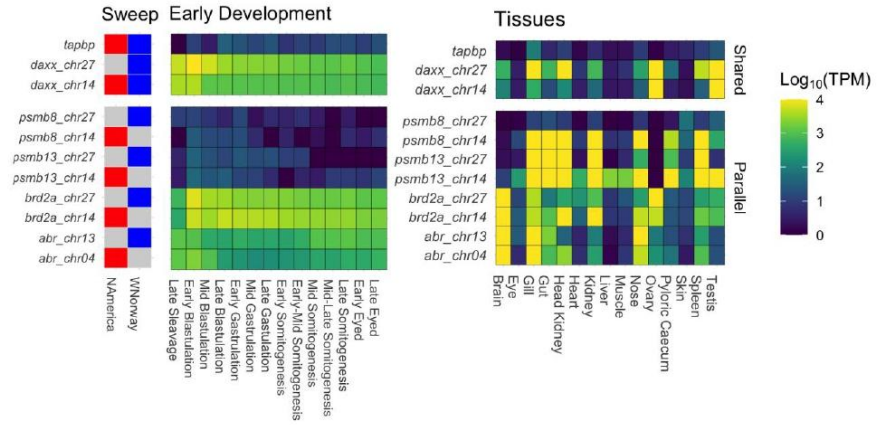


Fig. 4. Expression of genes under parallel selection or parallel selection on the homeologous regions in early development stages and mature tissues. Gene expression is represented as log-transformed TPM. The sweep panel indicates the gene that has undergone selective sweeps in North America or Western Norway. The left heatmap shows the gene expression levels during early development, from late gastrulation to the eyed egg stage. The right heatmap shows the gene expression levels across different tissues in the mature Atlantic salmon. We used publicly available RNA sequencing data from early development stages (AQUA-FAANG project No. PRJEB51855) and from various tissues of healthy fish (NCBI Project No. SRP011583) including brain, eye, gill, gut, head kidney, heart, kidney, liver, muscle, nose, ovary, caecum, skin, spleen, and testis of Atlantic salmon in Europe.

aquaculture group (Figs. 2 and 3a and Table 2). These lineage-specific selection patterns on homeolog regions suggest parallel selection targeting the MHC class I region.

To estimate the degree of selection signals in ohnolog genes expected by random chance, we performed a simulation analysis by randomly placing putative genes with outlier XP-EHH values across the genome 10,000 times. For cases where "one gene exhibited selection in one population while its duplicate showed selection in the other population," we observed 4 such pairs in empirical data, with a simulated probability of 0.03 of observing 4 or more by chance. Similarly, for cases where "the same gene exhibited selection in both populations," 2 such genes were observed, with a simulated probability of 0.027. These results indicate that while the observed selection patterns in duplicated genes are not extremely rare, they are also not entirely expected by chance.

Expression Pattern of the Genes Under Putative Selection

We leveraged publicly available RNA-seq datasets spanning various developmental stages and tissues. This allowed us to infer the potential functions of these putatively adaptive genes and regulatory difference of homeolog gene copies.

Both *daxx* homeolog gene copies, showed high expression in the ovary, testis, and during early blastulation stages (Fig. 4), suggesting a role in both reproduction and early development. Additionally, *daxx_chr27* is also highly expressed in the gill, spleen, and head kidney, indicating its role in immunity. This pattern suggests that both copies may contribute to multiple physiological processes.

The *tapbp* gene, showed low expression in most tissues in the healthy fish. This gene was, though, moderately expressed in the gill (Fig. 4), which serves multiple important functions, including gas exchange, osmoregulation, and acting as a gateway for potential pathogens and its expression pattern suggests a potential role in the immune response (Amill et al. 2024). It is reported that *tapbp* expression is upregulated in a rainbow trout monocyte/macrophage cell line in response to viral infection (Sever et al. 2014).

Both homeologous copies of *brd2* show high expression in the brain and gill, while *brd2_chr27* is also expressed in ovary and *brd2_chr14* is expressed in head kidney, kidney, and nose (Fig. 4). As with *daxx*, this pattern indicates that each copy may have common and divergent roles. Both copies also display high expression levels during the early stages of development, particularly in early blastulation (Fig. 4). This is consistent with the established roles of *brd2* in reproductive and developmental processes, including spermatogenesis and folliculogenesis (Rhee et al. 1998), and egg polarity in zebrafish (*Danio rerio*) (DiBenedetto et al. 2008). *Brd2* is also responsible for nervous system development, morphogenesis and differentiation of the digestive tract in zebrafish (Murphy et al. 2017). In the GO analysis, both *daxx* and *brd2* appear in the organelle organization category that is enriched in genes under selection in Norwegian aquaculture (Table 1).

Downloaded from https://academic.oup.com/gbe/article/17/4/evaf063/8115854 by guest on 08 January 2026

Another duplicate pair *Psmb8*, codes for a beta subunit of the type 8 proteasome, which facilitates endopeptidase activity and proteasomal protein catabolism (Arima et al. 2011). *Psmb8a_chr14* is highly expressed in many tissues, including immune-related tissues such as the gill, head kidney, and spleen, while *psmb8a_chr27* did not show such high expression, suggesting possible regulatory difference (Fig. 4). Both copies of *abr*, which activate cerebellar development in multiple vertebrates (Chuang et al. 1995; Mulherkar et al. 2014; Guo et al. 2020), were highly expressed in the brain (Fig. 4), also suggesting a neurological function in Atlantic salmon.

Discussion

Domestication often involves unintentional selection for traits such as tameness, reduced stress, and changes in neurological and behavioral responses, enabling species to better coexist with humans (Fleming et al. 1996; Islam et al. 2020). In contrast, artificial selection deliberately targets desirable traits linked to dietary, economic, or social factors. Over the past 50 years (~15 generations), independent Atlantic salmon populations have undergone both domestication and artificial selection for key production traits, resulting in specialized phenotypes significantly distinct from their wild counterparts (Milla et al. 2021). Parallel evolution across different populations is expected due to similar captive environments and selection pressures (Bourret et al. 2013; Elmer et al. 2014). Moreover, aquaculture species that experienced WGD, exemplified with salmonids, may have benefited from this genomic event, potentially hosting the genetic resources to adapt to novel, artificial environments, and intensive selection of aquaculture (Lien et al. 2016; Chen et al. 2019; Xu et al. 2019; Du et al. 2020; Redmond et al. 2023; Rondeau et al. 2023).

In this study, we investigated how domestication and artificial selection have shaped the genomic landscapes of farmed populations compared to the wild. Genome-wide scans using whole-genome sequencing data from Atlantic salmon populations in North America and Western Norway revealed evidence of genetic divergence between farmed and wild populations. We also identified parallel selection sweeps driven by positive selection in the 2 farmed strains, addressing a knowledge gap in previous research. Among these, our findings highlight parallel ohnologs—a rare case where WGD-derived homeologous regions show shared selection signals during independent domestication events.

Parallel Selective Sweeps in MHC Genes in Aquaculture Across Continents: Likely Retention of Variation Under Long-term Balancing Selection

Our comparative study identified 82 SNPs within the *daxx* and adjacent *tapbp* genes, including 5 nonsynonymous SNPs, shared selective sweeps in aquaculture population pairs in Western Norway and North America. Both genes are members of the MHC class IA region in Atlantic salmon (Grimholt 2018), suggesting crucial roles in the immune system. The *tapbp* gene encodes a transmembrane glycoprotein, which stabilizes the peptide loading complex of MHC class I molecules, aiding in T cell immune recognition (Ortmann et al. 1997; Lehner et al. 1998; Garbi et al. 2003). The *daxx* gene has a broad range of functions in vertebrates such as regulation of telomere length and stability (Heaphy et al. 2011) and chromatin remodeling (Drané et al. 2010; Goldberg et al. 2010; Rapkin et al. 2013; Lovejoy et al. 2020). Subsequent research has shown that daxx is also involved in apoptosis (Khelifi et al. 2005; Yao et al. 2014; Détrée and Gonçalves 2019) and as a tumor suppressor (Mahmud and Liao 2019). However, its interaction with the MHC class I in fish, especially in salmon remain elusive.

Our findings suggest that long-term balancing selection has maintained high allelic diversity in the MHC regions across populations in North America and Western Norway that diverged ~1 million years ago (Rougemont and Bernatchez 2018), as observed in other vertebrates (McConnell et al. 2016; Veríssimo et al. 2023). This diversity allows the shared, ancient SNPs in the MHC region to undergo parallel selection in aquaculture conditions. In contrast, non-MHC regions likely lack shared SNPs already due to genetic drift, which has led to fixation in each lineage and reduced the potential for parallel selection. In addition, genomic structural differences between the 2 lineages may have reduced the number of shared variants outside the MHC regions.

In natural environments, we can assume that the allelic diversity at the MHC locus of Atlantic salmon has been maintained through balancing selection. However, in aquaculture settings, specific alleles have been selected due to common selective pressures for increased immune response to infectious diseases in high-density net pen populations (Krkosek 2010). Alternatively, this selection may target other phenotypic effects of these genes exemplified by the fact that *daxx* also plays a key role in apoptosis (Michaelson et al. 1999; Khelifi et al. 2005). While nonsynonymous SNPs have been identified, pinpointing the causal SNPs needs more targeted approaches, such as MHC-specific sequencing, to fully resolve this rapidly evolving region (Sundaram et al. 2020) and functional analysis, such as gene editing. In the long term, coevolution between hosts and pathogens could lead to currently advantageous alleles becoming disadvantageous (Spurgin and Richardson 2010).

In addition, lineage-specific selective sweeps in other MHC genes further support the role of artificial selection in shaping immunity, even when the specific SNPs and genes with adaptive signatures differ between the 2 domesticated lines.

Downloaded from https://academic.oup.com/gbe/article/17/4/evaf063/8115854 by guest on 08 January 2026

Parallel Selection on Homeologous Regions

We observed parallel selection on 2 paired regions of salmon-specific WGD-derived ohnologs (chr14-27 and chr4-13), and one major region is the MHC region with high Tajima's D value in the aquaculture populations. This suggests balancing selection has been contributed in maintaining adaptive immune variation across duplicates in addition to prolonged tetraploid recombination, as observed in the late rediploidization LORe regions (Robertson et al. 2017). However, further in-depth research is needed to fully resolve how WGD-derived variation has contributed to adaptation, particularly in the context of domestication.

In addition, we observed 4 WGD-derived gene pairs under parallel domestication-related selection across 2 distinct geographical lineages: *psmb8*, *psmb13a*, *brd2a* on chr14 and chr27 (genes on MHC regions), and *abr* on chr4 and chr13 (gene involved in neurological and brain functions; Chuang et al. 1995; Mulherkar et al. 2014; Guo et al. 2020). Such selection signatures may be associated with ongoing adaptation to aquaculture conditions. In aquaculture environments, fish are raised under higher density conditions than in the wild, with controlled lighting, temperature, and salinity levels. In such artificial environments, advantageous alleles in captivity may differ from beneficial alleles in nature. Disease outbreaks are a major challenge in aquaculture, regardless of species (Cain 2022). Disease resistance, a key trait in aquaculture, may be enhanced by such immune genes on homeologous regions under selection. While this suggests that WGD-derived variation may contribute to specific adaptations, such occurrences are rare and its impact appears to be limited, likely due to the ancient nature of the WGD event and subsequent gene diversification or pseudogenization. Nonetheless, these genes may play a supporting role in the adaptation of Atlantic salmon to aquaculture environments. In this context, the contribution of WGD primary contributions seem to be (i) allowing for increased immune variation maintained among duplicated immune gene complexes and (ii) reducing pleiotropy through subfunctionalization (Marchant et al. 2019). While genetic changes resulting from WGD may contribute to specific adaptations, domestication has occurred across diverse phylogenetic lineages with varying genetic backgrounds, indicating that multiple genomic pathways can facilitate adaptation to human-mediated environments.

The role of WGD-derived variation in adaptation is not unique to Atlantic salmon. Parallel selection on WGD-originated gene pairs has been reported in other taxa, such as vibrant skin color in common carp and leaf-heading traits in *Brassica* species (Cheng et al. 2016; Wang et al. 2024), demonstrating that duplicated genes can contribute to adaptive evolution although the underlying genetic mechanisms may differ across taxa. Our findings provide limited direct evidence that WGD derived variation facilitated adaptation to aquaculture conditions, indicating that many other factors, including additional facets of species biology and genetics, selective breeding and environmental pressures, are also crucial in shaping domestication outcomes. Further studies integrating functional validation and broader taxonomic comparisons will be necessary to disentangle these contributions.

Sema Genes Under Selection in the Western Norwegian Aquaculture Group

Building on understanding neural-related genetic selection in aquaculture, we underscore the sema genes (Alto and Terman 2017). GO enrichment analysis revealed that genetic SNPs under selection related to semaphorins were overrepresented in the Western Norwegian aquaculture group (*sema 4ab*, *sema5ba*, *sema6d*, and *sema6e*). Semaphorin is involved in axon guidance and synapse formation (Luo et al. 1993; Jongbloets and Pasterkamp 2014). Semaphorins have also been detected under selection in previous studies of aquaculture Atlantic salmon populations (Gutierrez et al. 2016; Naval-Sanchez et al. 2020), supporting our findings. Such genes under lineage-specific selection should also contribute to the domestication of Atlantic salmon. It has been reported that structural variation showing genetic divergence between farmed and wild Atlantic salmon is linked to synaptic genes responsible for behavioral variation in Atlantic salmon (Bertolotti et al. 2020). Such work suggests that individuals with greater stress tolerance have been selectively favored in aquaculture environments. Meanwhile, it is reported that fast-growing fish tend to be more aggressive (Nicieza and Metcalfe 1999). So, it is also possible that the fast-growing fish have been selected, potentially leading to the inadvertent selection of genes associated with aggressive behaviors as a byproduct.

Conclusion

Our findings contribute to a better understanding of the genomic mechanisms underlying aquaculture adaptation, across continents. By identifying immune and developmental genes under artificial selection, this study highlights key processes involved in the domestication of Atlantic salmon. These results also emphasize the role of genetic diversity and redundancy in shaping evolutionary trajectories and resilience to challenges in aquaculture-specific environments.

Materials and Methods

Studied Populations

The present study involved 370 Atlantic salmon individuals from 2 wild and aquaculture populations across distinct

Downloaded from https://academic.oup.com/gbe/article/17/4/evaf063/8115854 by guest on 08 January 2026

phylogeographical lineages, aiming to compare genomic landscapes of domestic strains with local wild counterparts. Whole-genome sequencing data was retrieved from publicly available datasets.

We used 80 North American aquaculture samples derived from the Gaspe, St. John, and Penobscot Rivers (NCBI PRJEB34225) and 112 Norwegian MOWI strain samples from the River Bolstad, Årøy, and marine captures near Osterfjord and Sotra (NCBI PRJEB47441; Glover et al. 2009). Wild counterparts included 80 North American individuals from 8 Quebec rivers and 98 Norwegian individuals from 17 rivers in Western Norway (NCBI PRJEB38061; Bertolotti et al. 2020), see Fig. 1; supplementary tables S1 and S2, Supplementary Material online.

In North America, aquaculture samples were taken from Gaspe ($n = 16$), St. John ($n = 53$), and Penobscot Rivers ($n = 11$). Wild samples were from 8 rivers: Malbaie Charlevoix, Bonaventure, Petite rivière Cascapedia, Laval, De la Chaloupe, Jupiter, Du Vieux Fort, and Saint-Paul ($n = 10$ each). In Western Norway, aquaculture samples were from the MOWI Norwegian breeding line ($n = 112$). Wild counterparts were from 17 rivers: Vorma, Suldalslågen, Vikedalselva, Etneelva, Eidfjordvassdraget, Oselva i Os, Loneelva i Osterøy, Flåmselva, Lærdalselva, Årøy, Daleelva, Flekkeelva, Nausta, Jølstra, Oselvvassdraget, Ryggelva, and Gloppenelva (ranging from $n = 2$ to $n = 10$, supplementary tables S1 and S2, Supplementary Material online).

The North American lineages originate ostensively from single rivers and so should not have an admixed founding population; however, there is documented supplementation from Canadian rivers early in the domestication history into Maine. These lineages also show evidence of European admixture despite use of European fish in Canadian aquaculture never having been approved (Bradbury et al. 2022). The origins of domesticated North American salmon strains are the Penobscot River strain (Gao et al. 2023), the St. John's River strain (Liu et al. 2017) and the Grand Cascapedia River (Nash and Waknitz 2003). These strains originated from the Gulf of Maine, the Bay of Fundy, and the Gaspé Bay, Quebec, respectively. Sampled populations were primarily from Quebec including the Petite Cascapedia River, which neighbors the source population of the Gaspé strain (Fig. 1; supplementary tables S1 and S2, Supplementary Material online). Historically, the Gaspé and Miramichi Rivers were used for population supplementation in Maine during the 1920s–1960s. This provides a basis for comparing Canadian wild populations with domesticated strains, as the Penobscot strain originated after Canadian stock introductions (FWS Training Center: https://trainingcenter.fws.gov/resources/knowledge-resources/salmon/asalmon4.html). The admixture history of North American domesticated lineages is evident in the PCA plot (Fig. 1).

The European lineage we analyzed, MOWI, is derived from a mixture of west coast lineages, primarily from the Vosso drainage basin and the River Årøy, in the 1960s. The wild sampled rivers include the source population River Årøy in Western Norway. The Vosso population has collapsed since salmon were taken from it for aquaculture; it is highly disturbed by introgression from farmed escapees, which is high in part because of population collapse from multiple other anthropogenic disturbances. The loss of the original Vosso source population due to farm-wild introgression is well-documented. Sægrov et al. (1997) reported that escaped farmed Atlantic salmon replaced the original stock in the River Vosso (Sægrov et al. 1997). Similarly, Glover et al. (2012) provided a spatio-temporal analysis of Atlantic salmon population genetic structure throughout Norway (Glover et al. 2012). Sampling efforts include Årøy, which was used as part of the wild dataset alongside other west coast populations. Vosso, however, was excluded due to its near-total replacement by aquaculture escapees (Våge 1995). As a result, contemporary samples from Vosso cannot, unfortunately, be used as representative of the original source population. A sample of 16 other west coast populations in the surrounding region was used in place of this source population.

The founding population size of aquaculture groups is unknown, though it is assumed that there will have been a population bottleneck at the founding of the domesticated lineages. However, this will have been countered by the admixture during the founding of the European lineage and the supplementation of domesticated lineages in Maine with Canadian and likely also European salmon. Resulting in domesticated lineages that have a history of population bottlenecks, admixture, and strong artificial selection. Disentangling these processes from the data will be complex without historical samples. The influence of admixture can be seen in the wider spread of domesticated lineages in PCA space (European on PC2 and North American across PC1 and PC2, see Fig. 1) compared to the wild lineages used. The Domesticated lineages cover greater PCA space than the combined spread of the wild populations in both continents, suggesting admixture is more important than population bottlenecks. Despite the Gaspé lineage have the closest source population sampled; this is actually the furthest in PCA space from the cluster of wild populations. The North American wild populations cluster tightly and overlap with the Maine and Bay of Fundy domesticated lineages, probably reflecting the use of Canadian salmon to supplement early aquaculture further south.

Data Preparation

Mapping, SNP calling and quality control were performed independently for each of the 4 datasets (Aquaculture Norwegian, Wild Norwegian, Aquaculture North American, and Wild North American) using the same pipeline and parameters. Illumina reads were mapped to the

Downloaded from https://academic.oup.com/gbe/article/17/4/evaf063/8115854 by guest on 08 January 2026

Atlantic salmon genome (Ssal_v3.1; GCA_905237065.2; Stenløkk 2023) using bwa-mem2 (Li and Durbin 2009). Duplicate reads were marked with gatk4-spark:4.3.0.0 MarkDuplicates. Genetic variation was identified using gatk4-spark:4.3.0.0 HaplotypeCaller with default parameters and the individual genotypes were merged with gatk4-spark:4.3.0.0 GenotypeGVCFs. To avoid the detection of false positives, SNPs were filtered by the following parameters with gatk4-spark:4.3.0.0 (McKenna et al. 2010) : « QD < 2.0 » which filters out SNPs with a low-quality score by depth of coverage; « QUAL < 50.0 » to sort SNPs with a low-quality score; « SOR > 4.0 » that removes SNPs where the number of reads supporting the reference allele is significantly skewed toward 1 strand (forward or reverse) compared to the other; « FS > 60.0 » to filter out SNPs whose reads supporting the reference allele versus the alternative allele are significantly different between the 2 strands; and finally « ReadPosRankSum < -8.0 » to remove SNPs with a low score for the difference between the mean position of the reference allele and the alternative allele among the reads supporting the SNP (https://gatk.broadinstitute.org/hc/en-us). From this large pruning, we kept only biallelic markers (SNPs) with depth >4 and <30, confident call rate >90% (GQ) and missingness lower than 10% using VCFtools (Danecek et al. 2011). The whole analytical pipeline is summarized in supplementary fig. S1, Supplementary Material online.

Population Structure Analysis

We used 2 clustering methods to investigate genetic structure among populations and have a broad representation of the genetic differentiation between North American and West Norwegian lineages. Additional filtering steps specific to these analyses were applied to remove loci that deviated from Hardy–Weinberg equilibrium, loci with a minor allele frequency (MAF) below 5%, and linked SNPs using PLINK v1.9 (Purcell et al. 2007). Linkage disequilibrium pruning was performed using a sliding window approach (i.e. 50 SNP windows, 10 SNP step size, and an R^2 threshold of 0.5). We ran ADMIXTURE (Alexander et al. 2009; Alexander and Lange 2011) with a number of ancestral populations set from 1 to 20 (K). The optimal K was selected based on the lowest cross-validation error. We then conducted a PCA using PLINK v1.9 (Purcell et al. 2007) on the same filtered dataset. The results of both analyses (i.e. eigenvectors and genetic variation from PCA; genetic components from ADMIXTURE) were visualized in R 4.2.2 using the *ggplot2* (Wickham 2011) and *pophelper* (Francis 2017) R packages.

Detection of Selection Sweeps

For the detection of putative genomic regions under artificial selection in domestic strains, we applied a genome scan using 3 statistical methods based on (i) population differentiation (F_{ST}; Weir and Cockerham 1984); (ii) site frequency (Tajima's D; Tajima 1989), and (iii) cross-population test of extended haplotype homozygosity (XP-EHH; Sabeti et al. 2002). These tests were first carried out to compare aquaculture and wild populations from the same geographical region independently in order to detect lineage-specific selection signatures. Then, we investigated whether some genomic regions putatively under selection are shared by populations of different origins, indicative of parallel evolution. In the context of domestication, we expect 2 main scenarios in terms of selective sweeps. A hard selective sweep occurs when a beneficial allele arises and rapidly reaches fixation within a population. There is little genetic variation surrounding the selected allele because the advantageous SNP quickly "sweeps" through the population, leaving little time for new mutations or recombination to occur. As a result, the genetic diversity around the selected site is reduced, and the favored allele is present on a single haplotype in the population. A soft selective sweep occurs when multiple beneficial alleles, either derived from different mutations or standing genetic variation, rise in frequency or are favored within a population. During a soft sweep, there can be multiple haplotypes carrying different advantageous alleles and the genetic diversity around the selected site is not as reduced as in a hard sweep, and multiple haplotypes may show evidence of positive selection.

First, F_{ST} identifies selection using differences in allele frequencies between populations, here between domesticated and wild populations from each geographic area independently. These statistics were calculated according to the pairwise Weir and Cockerham's (1984) estimator using VCFtools (Danecek et al. 2011) in 20,000 bp nonoverlapping sliding windows using wild individuals as the reference population. We extracted the top 1% values as outliers.

Second, Tajima's D assumes that neutral genetic variation is maintained by a balance between mutation, genetic drift, and gene flow. The test is computed as the ratio between the number of segregating sites and the average pairwise nucleotide differences. Deviations from this neutral evolution can indicate putative selection or demographic events. Markers under selection can be expressed as positive and negative values, while zero values indicate a putative absence of selection. A significant positive Tajima's D value may indicate an excess of intermediate-frequency alleles, which could be caused by balancing selection, population subdivision, or recent population expansion. A significant negative Tajima's D value may indicate an excess of low-frequency alleles, which could be caused by positive selection, population bottlenecks, or purifying selection. In our study, Tajima's D was conducted on the 2 aquaculture populations independently by

Downloaded from https://academic.oup.com/gbe/article/17/4/evaf063/8115854 by guest on 08 January 2026

nonoverlapping sliding windows of 20,000 bp. As the domestication involved bottlenecks and focus on directional selection, we are more interested in negative values. The lowest values at the 1% level were taken as the significance level. F_{ST} and Tajima's D were calculated from vcf files resulting from the soft filtering (i.e depth, low quality SNPs, and missingness; see *Data preparation* section).

Finally, we investigated patterns of extended haplotypes between pairs of populations using the *rehh* R package (Gautier et al. 2017). Haplotype analyses are based on extended haplotype homozygosity (EHH) that represents the model of a selective sweep in which an adaptive de novo mutation appears on a haplotype and rapidly evolves toward fixation, thus reducing diversity around the locus (Sabeti et al. 2002). A hard sweep induces a signal of high haplotype homozygosity to be observed extending from an adaptive locus as strong rapid selection causes strong linkage that is not broken by recurrent mutation or recombination. In other words, under conditions of positive selection, a beneficial new allele can rapidly increase in frequency and create long stretches of homozygosity, or identical copies of the haplotype, around the selected site. This process results in longer and linked haplotypes detectable by EHH methods. In our case, we applied a cross-population (XP-EHH) test that compares the distribution of haplotype homozygosity between aquaculture and wild populations. When there is divergent selection between the 2 environments (wild and aquaculture), the allele selected in the focal population (aquaculture) will be on a longer haplotype. This signal will not be present when selection is concordant between the 2 environments. Negative scores of XP-EHH suggest that selection occurred in the reference population, whereas positive scores show selection in the domestic population. Unlike Tajima and F_{ST}, these tests cannot directly handle SNP genotypes because they require haplotypes. Added to the soft filtering, SNPs with MAF below 5% were removed. The remaining SNPs were phased to infer the haplotypes of each population using BEAGLE v4.1 (Browning and Browning 2007) with a window size of 10,000 and 600 bp overlap. One-sided *P*-values were obtained as $-\log10(1-\Phi[\text{XP-EHH}])$, where Φ(XP-EHH) represents the Gaussian distribution function for the statistic. FDR test was performed to estimate the number of false positives and determine the significant threshold. We used $-\log10(P\text{-value}) = 4$ (alpha = 0,0001) as a threshold to define significant values of XP-EHH. As XP-EHH searches for the extended haplotype, which is typical of a hard sweep, this method will be more sensitive in detecting hard sweeps than soft selective sweeps.

SNP Annotation

Significant windows or SNPs from the 3 comparative analyses were intersected using BEDtools (Quinlan 2014) to examine markers detected by multiple methods and thus increase the repeatability of the results. For each geographical region, 3 intersections were tested: (i) F_{ST}/Tajima's D, (ii) F_{ST}/XP-EHH, and (iii) Tajima's D/XP-EHH. Candidate SNPs detected by 2 or more methods were then intersected with annotated genes in the Atlantic salmon reference genome (Ssal_v3.1; GCA_905237065.2 ; Stenløkk 2023) using BEDtools window to identify SNPs under selection that are contained in the same windows as the genes or that are directly mapped to genes. The overlapping genes were input into ShinyGO 0.77 (Ge et al. 2020; http://bioinformatics.sdstate.edu/go/) to extract biological functions and enrichment. All the intersects were tested independently to observe population-specific signatures. We then tested parallel evolution using only the significant SNPs from XP-EHH, as this was the most suitable method. Categories were extracted where the enrichment FDR was <0.05, Fold Enrichment >2, and the GO category contained ≥3 genes (Table 1).

Gene Expression Analysis

We used available RNA sequencing data online (i) from early development stages (AQUA-FAANG project No. PRJEB51855) and (ii) from tissues (NCBI Project No. SRP011583) including brain, eye, gill, gut, head kidney, heart, kidney, liver, muscle, nose, ovary, caecum, skin, spleen, and testis. Data were quantified using the latest version of the Ensembl transcriptome (Ssa v.3.1). The fastq files were first filtered using fastp v0.23.2 (Chen 2023) with default settings in order to remove potential sequencing adaptors in the reads. Then, the reads were mapped using the lightweight mapper salmon v1.5.2 (Patro et al. 2017), which tracks, by default, the position and orientation of all mapped fragments. This information is used in conjunction with the abundances from online inference to compute per-fragment conditional probabilities (Patro et al. 2017). The produced TPM (transcripts per million) was visualized using the package *ggplot2* in R.

Test for Balancing Selection on the MHC Region on Chromosome 14

Investigation of balancing selection was carried out using Tajima's D (Tajima 1989). Tajima's D value was computed in the MHC region on chromosome 14 between 64,320,000 and 64,340,000 bp. Then, Tajima's D was computed on 500 windows of 20,000 pb with more than 15 SNPs, randomly selected across the genome. We used variant effect predictor (McLaren et al. 2016) to predict SNPs consequences on coding genes.

Supplementary Material

Supplementary material is available at *Genome Biology and Evolution* online.

Downloaded from https://academic.oup.com/gbe/article/17/4/evaf063/8115854 by guest on 08 January 2026

Acknowledgements

The authors thank Joris Bertrand from the Laboratoire Génome et Développement des Plantes (LGDP) in Perpignan (France) for his advice on genetic structure analysis. We acknowledge the use of the Orion computing cluster at the Norwegian University of Life Sciences (NMBU). We thank Louise Chavarie for her comments on the geography of Canada.

Authors Contributions

Pauline Buso (Conceptualization, Data curation, Investigation, Writing—Original Draft Preparation, Review and Editing Manuscript), Célian Diblasi (Investigation, Software, Review and Editing Manuscript), Domniki Manousi (Investigation, Software, Review and Editing Manuscript), Jun Soung Kwak (Investigation, Software, Review and Editing Manuscript), Kristina Stenløkk (Investigation, Software, Review and Editing Manuscript), Nicola Barson (Conceptualization, Investigation, Writing—Original Draft Preparation, Review and Editing Manuscript), and Marie Saitou (Conceptualization, Investigation, Writing—Original Draft Preparation, Review and Editing Manuscript)

Funding

This work was initiated by PB as part of the Master degree in Functional Biology and Ecology funded by the EUR TULIP-GS N°ANR-18-EUR-0019 a grant managed by the French National Research Agency (ANR) in the framework of the program "Investing for the Future". The study was supported by the Research Council of Norway (grant nos. 325874, 275310, and 221734). D.M. was supported by the Norwegian University of Life Sciences (NMBU), BIOVIT, PhD funding.

Data Availability

Script used for the analyses in this work are available at https://github.com/paulinebuso/ATLANTIDES_pipelines

Literature Cited

Alexander DH, Lange K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics. 2011:12(1):246. https://doi.org/10.1186/1471-2105-12-246.

Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009:19(9):1655–1664. https://doi.org/10.1101/gr.094052.109.

Alto LT, Terman JR. Semaphorins and their signaling mechanisms. In: Terman JR, editor. Semaphorin signaling. New York (NY): Springer New York (Methods in Molecular Biology); 2017. p. 1–25.

Amill F, Gauthier J, Rautio M, Derome N. Characterization of gill bacterial microbiota in wild Arctic char (*Salvelinus alpinus*) across lakes, rivers, and bays in the Canadian Arctic ecosystems. Microbiol Spectr. 2024:12(3):e02943-23. https://doi.org/10.1128/spectrum.02943-23.

Arima K, Kinoshita A, Mishima H, Kanazawa N, Kaneko T, Mizushima T, Ichinose K, Nakamura H, Tsujino A, Kawakami A, et al. Proteasome assembly defect due to a proteasome subunit beta type 8 (PSMB8) mutation causes the autoinflammatory disorder, Nakajo-Nishimura syndrome. Proc Natl Acad Sci U S A. 2011:108(36):14914–14919. https://doi.org/10.1073/pnas.1106015108.

Baduel P, Quadrana L, Hunter B, Bomblies K, Colot V. Relaxed purifying selection in autopolyploids drives transposable element over-accumulation which provides variants for local adaptation. Nat Commun. 2019:10(1):5818. https://doi.org/10.1038/s41467-019-13730-0.

Bailey SF, Blanquart F, Bataillon T, Kassen R. What drives parallel evolution? How population size and mutational variation contribute to repeated evolution. BioEssays. 2017:39(1):e201600176. https://doi.org/10.1002/bies.201600176.

Bertolotti AC, Layer RM, Gundappa MK, Gallagher MD, Pehlivanoglu E, Nome T, Robledo D, Kent MP, Røsæg LL, Holen MM, et al. The structural variation landscape in 492 Atlantic salmon genomes. Nat Commun. 2020:11(1):5176. https://doi.org/10.1038/s41467-020-18972-x.

Bicskei B, Bron JE, Glover KA, Taggart JB. A comparison of gene transcription profiles of domesticated and wild Atlantic salmon (*Salmo salar* L.) at early life stages, reared under controlled conditions. BMC Genomics. 2014:15(1):884. https://doi.org/10.1186/1471-2164-15-884.

Bourret V, Kent MP, Primmer CR, Vasemägi A, Karlsson S, Hindar K, McGinnity P, Verspoor E, Bernatchez L, Lien S. SNP-array reveals genome-wide patterns of geographical and potential adaptive divergence across the natural range of Atlantic salmon (*Salmo salar*). Mol Ecol. 2013:22(3):532–551. https://doi.org/10.1111/mec.12003.

Bradbury IR, Lehnert SJ, Kess T, Van Wyngaarden M, Duffy S, Messmer AM, Wringe B, Karoliussen S, Dempson JB, Fleming IA, et al. Genomic evidence of recent European introgression into North American farmed and wild Atlantic salmon. Evol Appl. 2022:15(9):1436–1448. https://doi.org/10.1111/eva.13454.

Brenna-Hansen S, Li J, Kent MP, Boulding EG, Dominik S, Davidson WS, Lien S. Chromosomal differences between European and North American Atlantic salmon discovered by linkage mapping and supported by fluorescence in situ hybridization analysis. BMC Genomics. 2012:13(1):432. https://doi.org/10.1186/1471-2164-13-432.

Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet.2007:81(5):1084–1097. https://doi.org/10.1086/521987.

Bull JK, Stanford BCM, Bokvist JK, Josephson MP, Rogers SM. Environment and genotype predict the genomic nature of domestication of salmonids as revealed by gene expression. Proc R Soc Lond B Biol Sci. 2022:289(1988):20222124. https://doi.org/10.1098/rspb.2022.2124.

Cain K. The many challenges of disease management in aquaculture. J World Aquac Soc. 2022:53(6):1080–1083. https://doi.org/10.1111/jwas.12936.

Chen S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. iMeta. 2023:2(2):e107. https://doi.org/10.1002/imt2.107.

Chen Z, Omori Y, Koren S, Shirokiya T, Kuroda T, Miyamoto A, Wada H, Fujiyama A, Toyoda A, Zhang S, et al. De novo assembly of the goldfish (*Carassius auratus*) genome and the evolution of genes after whole-genome duplication. Sci Adv. 2019:5(6):eaav0547. https://doi.org/10.1126/sciadv.aav0547.

Downloaded from https://academic.oup.com/gbe/article/17/4/evaf063/8115854 by guest on 08 January 2026

Cheng F, Wu J, Cai C, Fu L, Liang J, Borm T, Zhuang M, Zhang Y, Zhang F, Bonnema G, et al. Genome resequencing and comparative variome analysis in a Brassica rapa and Brassica oleracea collection. Sci Data. 2016:3(1):160119. https://doi.org/10.1038/sdata.2016.119.
Chuang TH, Xu X, Kaartinen V, Heisterkamp N, Groffen J, Bokoch GM. Abr and Bcr are multifunctional regulators of the Rho GTP-binding protein family. Proc Natl Acad Sci U S A. 1995:92(22):10282–10286. https://doi.org/10.1073/pnas.92.22.10282.
Cross TF, King J. Genetic effects of hatchery rearing in Atlantic salmon. Aquaculture. 1983:33(1-4):33–40. https://doi.org/10.1016/0044-8486(83)90384-8.
Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. The variant call format and VCFtools. Bioinformatics. 2011:27(15):2156–2158. https://doi.org/10.1093/bioinformatics/btr330.
Debes PV, Normandeau ERIC, Fraser DJ, Bernatchez L, Hutchings JA. Differences in transcription levels among wild, domesticated, and hybrid Atlantic salmon (*Salmo salar*) from two environments. Mol Ecol. 2012:21(11):2574–2587. https://doi.org/10.1111/j.1365-294X.2012.05567.x.
Détrée C, Gonçalves AT. Transcriptome mining of apoptotic mechanisms in response to density and functional diets in Oncorhynchus mykiss and role in homeostatic regulation. Comp Biochem Physiol Part D Genomics Proteomics. 2019:31:100595. https://doi.org/10.1016/j.cbd.2019.100595.
Diamond J. Evolution, consequences and future of plant and animal domestication. Nature. 2002:418(6898):700–707. https://doi.org/10.1038/nature01019.
DiBenedetto AJ, Guinto JB, Ebert TD, Bee KJ, Schmidt MM, Jackman TR. Zebrafish brd2a and brd2bare paralogous members of the bromodomain-ET (BET) family of transcriptional coregulators that show structural and expression divergence. BMC Dev Biol. 2008:8(1):39. https://doi.org/10.1186/1471-213X-8-39.
Drané P, Ouararhni K, Depaux A, Shuaib M, Hamiche A. The death-associated protein DAXX is a novel histone chaperone involved in the replication-independent deposition of H3.3. Genes Dev. 2010:24(12):1253–1265. https://doi.org/10.1101/gad.566910.
Du K, Stöck M, Kneitz S, Klopp C, Woltering JM, Adolfi MC, Feron R, Prokopov D, Makunin A, Kichigin I, et al. The sterlet sturgeon genome sequence and the mechanisms of segmental rediploidization. Nat Ecol Evol. 2020:4(6):841–852. https://doi.org/10.1038/s41559-020-1166-x.
Ebadi M, Bafort Q, Mizrachi E, Audenaert P, Simoens P, Van Montagu M, Bonte D, Van de Peer Y. The duplication of genomes and genetic networks and its potential for evolutionary adaptation and survival during environmental turmoil. Proc Natl Acad Sci U S A. 2023:120(41):e2307289120. https://doi.org/10.1073/pnas.2307289120.
Elmer KR, Fan S, Kusche H, Luise Spreitzer M, Kautt AF, Franchini P, Meyer A. Parallel evolution of Nicaraguan crater lake cichlid fishes via non-parallel routes. Nat Commun. 2014:5(1):5168. https://doi.org/10.1038/ncomms6168.
Fernandez AR, Sáez A, Quintero C, Gleiser G, Aizen MA. Intentional and unintentional selection during plant domestication: herbivore damage, plant defensive traits and nutritional quality of fruit and seed crops. New Phytol. 2021:231(4):1586–1598. https://doi.org/10.1111/nph.17452.
Fleming IA, Jonsson B, Gross MR, Lamberg A. An experimental study of the reproductive behaviour and success of farmed and wild Atlantic Salmon (*Salmo salar*). J Appl Ecol. 1996:33(4):893. https://doi.org/10.2307/2404960.
Ford JS, Myers RA. A global assessment of salmon aquaculture impacts on wild salmonids. PLoS Biol. 2008:6(2):e33. https://doi.org/10.1371/journal.pbio.0060033.
Francis RM. Pophelper : an R package and web app to analyse and visualize population structure. Mol Ecol Resour. 2017:17(1):27–32. https://doi.org/10.1111/1755-0998.12509.
Frantz LAF, Bradley DG, Larson G, Orlando L. Animal domestication in the era of ancient genomics. Nat Rev Genet. 2020:21(8):449–460. https://doi.org/10.1038/s41576-020-0225-0.
Gao G, Waldbieser GC, Youngblood RC, Zhao D, Pietrak MR, Allen MS, Stannard JA, Buchanan JT, Long RL, Milligan M, et al. The generation of the first chromosome-level de novo genome assembly and the development and validation of a 50 K SNP array for the St. John River aquaculture strain of North American Atlantic salmon. G3 (Bethesda). 2023:13(9):jkad138. https://doi.org/10.1093/g3journal/jkad138.
Garbi N, Tiwari N, Momburg F, Hämmerling GJ. A major role for tapasin as a stabilizer of the TAP peptide transporter and consequences for MHC class I expression. Eur J Immunol. 2003:33(1):264–273. https://doi.org/10.1002/immu.200390029.
Gautier M, Klassmann A, Vitalis R. Rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. Mol Ecol Resour. 2017:17(1):78–90. https://doi.org/10.1111/1755-0998.12634.
Ge SX, Jung D, Yao R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. Bioinformatics. 2020:36(8):2628–2629. https://doi.org/10.1093/bioinformatics/btz931.
Gillard GB, Grønvold L, Røsæg LL, Holen MM, Monsen Ø, Koop BF, Rondeau EB, Gundappa MK, Mendoza J, Macqueen DJ, et al. Comparative regulomics supports pervasive selection on gene dosage following whole genome duplication. Genome Biol. 2021:22(1):103. https://doi.org/10.1186/s13059-021-02323-0.
Gjederm T. Selection for growth rate and domestication in Atlantic salmon. Z Tierzüchtung Züchtungsbiologie. 1979:96(1-4):56–59. https://doi.org/10.1111/j.1439-0388.1979.tb00199.x.
Gjedrem T, Gjøen HM, Gjerde B. Genetic origin of Norwegian farmed Atlantic salmon. Aquaculture. 1991:98(1-3):41–50. https://doi.org/10.1016/0044-8486(91)90369-I.
Glover KA, Otterå H, Olsen RE, Slinde E, Taranger GL, Skaala Ø. A comparison of farmed, wild and hybrid Atlantic salmon (*Salmo salar* L.) reared under farming conditions. Aquaculture. 2009:286(3-4):203–210. https://doi.org/10.1016/j.aquaculture.2008.09.023.
Glover KA, Quintela M, Wennevik V, Besnier F, Sørvik AGE, Skaala Ø. Three decades of farmed escapees in the wild: a spatio-temporal analysis of Atlantic salmon population genetic structure throughout Norway. PLoS One. 2012:7(8):e43129. https://doi.org/10.1371/journal.pone.0043129.
Glover KA, Solberg MF, McGinnity P, Hindar K, Verspoor E, Coulson MW, Hansen MM, Araki H, Skaala Ø, Svåsand T. Half a century of genetic interaction between farmed and wild Atlantic salmon: status of knowledge and unanswered questions. Fish Fish. 2017:18(5):890–927. https://doi.org/10.1111/faf.12214.
Goldberg AD, Banaszynski LA, Noh K-M, Lewis PW, Elsaesser SJ, Stadler S, Dewell S, Law M, Guo X, Li X, et al. Distinct factors control histone variant H3.3 localization at specific genomic regions. Cell. 2010:140(5):678–691. https://doi.org/10.1016/j.cell.2010.01.003.
Grimholt U. Whole genome duplications have provided teleosts with many roads to peptide loaded MHC class I molecules. BMC Evol Biol. 2018:18(1):25. https://doi.org/10.1186/s12862-018-1138-9.
Gundappa MK, To T-H, Grønvold L, Martin SAM, Lien S, Geist J, Hazlerigg D, Sandve SR, Macqueen DJ. Genome-wide reconstruction of rediploidization following autopolyploidization across one hundred million years of salmonid evolution. Mol Biol Evol. 2022:39(1):msab310. https://doi.org/10.1093/molbev/msab310.
Guo D, Yang X, Shi L. Rho GTPase regulators and effectors in autism spectrum disorders: animal models and insights for therapeutics. Cells. 2020:9(4):835. https://doi.org/10.3390/cells9040835.

Downloaded from https://academic.oup.com/gbe/article/17/4/evaf063/8115854 by guest on 08 January 2026

Gutierrez AP, Yáñez JM, Davidson WS. Evidence of recent signatures of selection during domestication in an Atlantic salmon population. Mar Genomics. 2016:26:41–50. https://doi.org/10.1016/j.margen.2015.12.007.

Harvey AC, Skilbrei OT, Besnier F, Solberg MF, Sørvik A-GE, Glover KA. Implications for introgression: has selection for fast growth altered the size threshold for precocious male maturation in domesticated Atlantic salmon? BMC Evol Biol. 2018:18(1):188. https://doi.org/10.1186/s12862-018-1294-y.

Heaphy CM, de Wilde RF, Jiao Y, Klein AP, Edil BH, Shi C, Bettegowda C, Rodriguez FJ, Eberhart CG, Hebbar S, et al. Altered telomeres in tumors with *ATRX* and *DAXX* mutations. Science. 2011:333(6041):425–425. https://doi.org/10.1126/science.1207313.

Houston RD, Bean TP, Macqueen DJ, Gundappa MK, Jin YH, Jenkins TL, Selly SLC, Martin SAM, Stevens JR, Santos EM, et al. Harnessing genomics to fast-track genetic improvement in aquaculture. Nat Rev Genet. 2020:21(7):389–409. https://doi.org/10.1038/s41576-020-0227-y.

Houston RD, Macqueen DJ. Atlantic salmon (*Salmo salar* L.) genetics in the 21st century: taking leaps forward in aquaculture and biological understanding. Anim Genet. 2019:50(1):3–14. https://doi.org/10.1111/age.12748.

Huang X, Huang S, Han B, Li J. The integrated genomics of crop domestication and breeding. Cell. 2022:185(15):2828–2839. https://doi.org/10.1016/j.cell.2022.04.036.

Innan H, Kim Y. Pattern of polymorphism after strong artificial selection in a domestication event. Proc Natl Acad Sci U S A.2004:101(29):10667–10672. https://doi.org/10.1073/pnas.0401720101.

Islam SS, Wringe BF, Bradbury IR, Fleming IA. Behavioural variation among divergent European and North American farmed and wild Atlantic salmon (*Salmo salar*) populations. Appl Anim Behav Sci. 2020:230:105029. https://doi.org/10.1016/j.applanim.2020.105029.

Izawa T. Reloading DNA history in rice domestication. Plant Cell Physiol. 2022:63(11):1529–1539. https://doi.org/10.1093/pcp/pcac073.

Jin Y, Olsen RE, Harvey TN, Østensen M, Li K, Santi N, Vadstein O, Bones AM, Vik JO, Sandve SR, et al. Comparative transcriptomics reveals domestication-associated features of Atlantic salmon lipid metabolism. Mol Ecol. 2020:29(10):1860–1872. https://doi.org/10.1111/mec.15446.

Jongbloets BC, Pasterkamp RJ. Semaphorin signalling during development. Development. 2014:141(17):3292–3297. https://doi.org/10.1242/dev.105544.

Khelifi AF, D'Alcontres MS, Salomoni P. Daxx is required for stress-induced cell death and JNK activation. Cell Death Differ. 2005:12(7):724–733. https://doi.org/10.1038/sj.cdd.4401559.

Klemetsen A, Amundsen P-A, Dempson JB, Jonsson B, Jonsson N, O'Connell MF, Mortensen E. Atlantic salmon *Salmo salar* L., brown trout *Salmo trutta* L. and Arctic charr *Salvelinus alpinus* (L.): a review of aspects of their life histories. Ecol Freshw Fish. 2003:12(1):1–59. https://doi.org/10.1034/j.1600-0633.2003.00010.x.

Krkosek M. Host density thresholds and disease control for fisheries and aquaculture. Aquac Environ Interact. 2010:1(1):21–32. https://doi.org/10.3354/aei0004.

Larson G, Piperno DR, Allaby RG, Purugganan MD, Andersson L, Arroyo-Kalin M, Barton L, Climer Vigueira C, Denham T, Dobney K, et al. Current perspectives and the future of domestication studies. Proc Natl Acad Sci U S A. 2014:111(17):6139–6146. https://doi.org/10.1073/pnas.1323964111.

Lehner PJ, Surman MJ, Cresswell P. Soluble tapasin restores MHC class I expression and function in the tapasin-negative cell line.220. Immunity. 1998:8(2):221–231. https://doi.org/10.1016/S1074-7613(00)80474-4.

Lehnert SJ, Bentzen P, Kess T, Lien S, Horne JB, Clément M, Bradbury IR. Chromosome polymorphisms track trans-Atlantic divergence and secondary contact in Atlantic salmon. Molec Ecol. 2019:28(8):2074–2087. https://doi.org/10.1111/mec.15065.

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

Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, Hvidsten TR, Leong JS, Minkley DR, Zimin A, et al. The Atlantic salmon genome provides insights into rediploidization. Nature. 2016:533(7602):200–205. https://doi.org/10.1038/nature17164.

Liu L, Ang KP, Elliott JAK, Kent MP, Lien S, MacDonald D, Boulding EG. A genome scan for selection signatures comparing farmed Atlantic salmon with two wild populations: testing colocalization among outlier markers, candidate genes, and quantitative trait loci for production traits. Evol Appl. 2017:10(3):276–296. https://doi.org/10.1111/eva.12450.

López ME, Benestan L, Moore J-S, Perrier C, Gilbey J, Di Genova A, Maass A, Diaz D, Lhorente J-P, Correa K, et al. Comparing genomic signatures of domestication in two Atlantic Salmon (*Salmo salar* L.) populations with different geographical origins. Evol Appl. 2019:12(1):137–156. https://doi.org/10.1111/eva.12689.

Lovejoy CA, Takai K, Huh MS, Picketts DJ, de Lange T. ATRX affects the repair of telomeric DSBs by promoting cohesion and a DAXX-dependent activity. PLoS Biol. 2020:18(1):e3000594. https://doi.org/10.1371/journal.pbio.3000594.

Lukacs MF, Harstad H, Bakke HG, Beetz-Sargent M, McKinnel L, Lubieniecki KP, Koop BF, Grimholt U. Comprehensive analysis of MHC class I genes from the U-, S-, and Z-lineages in Atlantic salmon. BMC Genomics. 2010:11(1):154. https://doi.org/10.1186/1471-2164-11-154.

Lukacs MF, Harstad H, Grimholt U, Beetz-Sargent M, Cooper GA, Reid L, Bakke HG, Phillips RB, Miller KM, Davidson WS, et al. Genomic organization of duplicated major histocompatibility complex class I regions in Atlantic salmon (*Salmo salar*). BMC Genomics. 2007:8(1):251. https://doi.org/10.1186/1471-2164-8-251.

Luo Y, Raible D, Raper JA. Collapsin: a protein in brain that induces the collapse and paralysis of neuronal growth cones. Cell. 1993:75(2):217–227. https://doi.org/10.1016/0092-8674(93)80064-L.

Macqueen DJ, Johnston IA. A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. Proc R Soc Lond B Biol Sci. 2014:281(1778):20132881. https://doi.org/10.1098/rspb.2013.2881.

Mahmud I, Liao D. DAXX in cancer: phenomena, processes, mechanisms and regulation. Nucleic Acids Res. 2019:47(15):7734–7752. https://doi.org/10.1093/nar/gkz634.

Marchant A, Cisneros AF, Dubé AK, Gagnon-Arsenault I, Ascencio D, Jain H, Aubé S, Eberlein C, Evans-Yamamoto D, Yachie N, et al. The role of structural pleiotropy and regulatory evolution in the retention of heteromers of paralogs. eLife. 2019:8:e46754. https://doi.org/10.7554/eLife.46754.

McConnell SC, Hernandez KM, Wcisel DJ, Kettleborough RN, Stemple DL, Yoder JA, Andrade J, de Jong JLO. Alternative haplotypes of antigen processing genes in zebrafish diverged early in vertebrate evolution. Proc Natl Acad Sci U S A. 2016:113(34):E5014–E5023. https://doi.org/10.1073/pnas.1607602113.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010:20(9):1297–1303. https://doi.org/10.1101/gr.107524.110.

Downloaded from https://academic.oup.com/gbe/article/17/4/evaf063/8115854 by guest on 08 January 2026

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. The ensembl variant effect predictor. Genome Biol. 2016:17(1):122. https://doi.org/10.1186/s13059-016-0974-4.
Michaelson JS, Bader D, Kuo F, Kozak C, Leder P. Loss of Daxx, a promiscuously interacting protein, results in extensive apoptosis in early mouse development. Genes Dev. 1999:13(15):1918–1923. https://doi.org/10.1101/gad.13.15.1918.
Mignon-Grasteau S, Boissy A, Bouix J, Faure J-M, Fisher AD, Hinch GN, Jensen P, Le Neindre P, Mormède P, Prunet P, et al. Genetics of adaptation and domestication in livestock. Livest Prod Sci. 2005:93(1):3–14. https://doi.org/10.1016/j.livprodsci.2004.11.001.
Mihajlovic L, Iyengar BR, Baier F, Barbier I, Iwaszkiewicz J, Zoete V, Wagner A, Schaerli Y. A direct experimental test of Ohno's hypothesis. Elife. 2025:13. https://doi.org/10.7554/eLife.97216.3.
Milla S, Pasquet A, El Mohajer L, Fontaine P. How domestication alters fish phenotypes. Rev Aquac. 2021:13(1):388–405. https://doi.org/10.1111/raq.12480.
Mulherkar S, Uddin MD, Couvillon AD, Sillitoe RV, Tolias KF. The small GTPases RhoA and rac1 regulate cerebellar development by controlling cell morphogenesis, migration and foliation. Dev Biol. 2014:394(1):39–53. https://doi.org/10.1016/j.ydbio.2014.08.004.
Murphy T, Melville H, Fradkin E, Bistany G, Branigan G, Olsen K, Comstock CR, Hanby H, Garbade E, DiBenedetto AJ. Knockdown of epigenetic transcriptional co-regulator Brd2a disrupts apoptosis and proper formation of hindbrain and midbrain-hindbrain boundary (MHB) region in zebrafish. Mech Dev. 2017:146:10–30. https://doi.org/10.1016/j.mod.2017.05.003.
Nash CE, Waknitz FW. Interactions of Atlantic salmon in the Pacific Northwest. Fish Res. 2003:62(3):237–254. https://doi.org/10.1016/S0165-7836(03)00063-8.
Naval-Sanchez M, McWilliam S, Evans B, Yáñez JM, Houston RD, Kijas JW. Changed patterns of genomic variation following recent domestication: selection sweeps in farmed Atlantic salmon. Front Genet. 2020:11:264. https://doi.org/10.3389/fgene.2020.00264.
Naylor RL, Hardy RW, Buschmann AH, Bush SR, Cao L, Klinger DH, Little DC, Lubchenco J, Shumway SE, Troell M. A 20-year retrospective review of global aquaculture. Nature. 2021:591(7851):551–563. https://doi.org/10.1038/s41586-021-03308-6.
Nicieza AG, Metcalfe NB. Costs of rapid growth: the risk of aggression is higher for fast-growing salmon. Funct Ecol. 1999:13(6):793–800. https://doi.org/10.1046/j.1365-2435.1999.00371.x.
Oleksyk TK, Smith MW, O'Brien SJ. Genome-wide scans for footprints of natural selection. Philos Trans R Soc Lond B Biol Sci. 2010:365(1537):185–205. https://doi.org/10.1098/rstb.2009.0219.
Ortmann B, Copeman J, Lehner PJ, Sadasivan B, Herberg JA, Grandea AG, Riddell SR, Tampé R, Spies T, Trowsdale J, et al. A critical role for tapasin in the assembly and function of multimeric MHC class I-TAP complexes. Science. 1997:277(5330):1306–1309. https://doi.org/10.1126/science.277.5330.1306.
Pajic P, Pavlidis P, Dean K, Neznanova L, Romano R-A, Garneau D, Daugherity E, Globig A, Ruhl S, Gokcumen O. Independent amylase gene copy number bursts correlate with dietary preferences in mammals. eLife. 2019:8:e44628. https://doi.org/10.7554/eLife.44628.
Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017:14(4):417–419. https://doi.org/10.1038/nmeth.4197.
Podgorniak T, Dhanasiri A, Chen X, Ren X, Kuan P-F, Fernandes J. Early fish domestication affects methylation of key genes involved in the rapid onset of the farmed phenotype. Epigenetics. 2022:17(10):1281–1298. https://doi.org/10.1080/15592294.2021.2017554.
Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007:81(3):559–575. https://doi.org/10.1086/519795.
Quinlan AR. BEDTools: the Swiss-army tool for genome feature analysis. Curr Protoc Bioinformatics. 2014:47(1):11.12.1–11.12.34. https://doi.org/10.1002/0471250953.bi1112s47.
Rapkin LM, Ahmed K, Li R, Bazett-Jones DP. A role for the histone chaperone DAXX in the structural organization of heterochromatin. Epigenetics Chromatin. 2013:6(S1):P121. https://doi.org/10.1186/1756-8935-6-S1-P121.
Redmond AK, Casey D, Gundappa MK, Macqueen DJ, McLysaght A. Independent rediploidization masks shared whole genome duplication in the sturgeon-paddlefish ancestor. Nat Commun. 2023:14(1):2879. https://doi.org/10.1038/s41467-023-38714-z.
Rhee K, Brunori M, Besset V, Trousdale R, Wolgemuth DJ. Expression and potential role of *Fsrg1*, a murine bromodomain-containing homologue of the *Drosophila* gene *female sterile homeotic*. J Cell Sci. 1998:111(23):3541–3550. https://doi.org/10.1242/jcs.111.23.3541.
Robertson FM, Gundappa MK, Grammes F, Hvidsten TR, Redmond AK, Lien S, Martin SAM, Holland PWH, Sandve SR, Macqueen DJ. Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification. Genome Biol. 2017:18(1):111. https://doi.org/10.1186/s13059-017-1241-z.
Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. Nat Biotechnol. 2011:29(1):24–26. https://doi.org/10.1038/nbt.1754.
Rondeau EB, Christensen KA, Johnson HA, Sakhrani D, Biagi CA, Wetklo M, Despins CA, Leggatt RA, Minkley DR, Withler RE, et al. Insights from a chum salmon (*Oncorhynchus keta*) genome assembly regarding whole-genome duplication and nucleotide variation influencing gene function. G3 (Bethesda). 2023:13(8):jkad127. https://doi.org/10.1093/g3journal/jkad127.
Rougemont Q, Bernatchez L. The demographic history of Atlantic salmon (*Salmo salar*) across its distribution range reconstructed from approximate Bayesian computations*: ATLANTIC SALMON HISTORY AND LINKED SELECTION. Evolution. 2018:72(6):1261–1277. https://doi.org/10.1111/evo.13486.
Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature. 2002:419(6909):832–837. https://doi.org/10.1038/nature01140.
Salman-Minkov A, Sabath N, Mayrose I. Whole-genome duplication as a key factor in crop domestication. Nat Plants. 2016:2(8):16115. https://doi.org/10.1038/nplants.2016.115.
Sever L, Vo NTK, Bols NC, Dixon B. Expression of tapasin in rainbow trout tissues and cell lines and up regulation in a monocyte/macrophage cell line (RTS11) by a viral mimic and viral infection. Dev Comp Immunol. 2014:44(1):86–93. https://doi.org/10.1016/j.dci.2013.11.019.
Sægrov H, Hindar K, Kålås S, Lura H. Escaped farmed Atlantic salmon replace the original salmon stock in the River Vosso, western Norway. ICES J Mar Sci. 1997:54(6):1166–1172. https://doi.org/10.1016/S1054-3139(97)80023-9.
Solberg MF, Robertsen G, Sundt-Hansen LE, Hindar K, Glover KA. Domestication leads to increased predation susceptibility. Sci Rep. 2020:10(1):1929. https://doi.org/10.1038/s41598-020-58661-9.

Downloaded from https://academic.oup.com/gbe/article/17/4/evaf063/8115854 by guest on 08 January 2026

Spurgin LG, Richardson DS. How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. Proc R Soc Lond B Biol Sci. 2010:277(1684):979–988. https://doi.org/10.1098/rspb.2009.2084.
Stenløkk KSR. Genomic structural variations as drivers of adaptation in salmonid fishes. 2023 [accessed 2025 Apr 9]. https://nmbu.brage.unit.no/nmbu-xmlui/handle/11250/3093077.
Sundaram AYM, Garseth ÅH, Maccari G, Grimholt U. An illumina approach to MHC typing of Atlantic salmon. Immunogenetics. 2020:72(1-2):89–100. https://doi.org/10.1007/s00251-019-01143-8.
Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989:123(3):585–595. https://doi.org/10.1093/genetics/123.3.585.
Teixeira JC, de Filippo C, Weihmann A, Meneu JR, Racimo F, Dannemann M, Nickel B, Fischer A, Halbwax M, Andre C, et al. Long-term balancing selection in LAD1 maintains a missense trans-Species polymorphism in humans, chimpanzees, and bonobos. Mol Biol Evol. 2015:32(5):1186–1196. https://doi.org/10.1093/molbev/msv007.
Teletchea F. Fish domestication in aquaculture: 10 unanswered questions. Anim Front. 2021:11(3):87–91. https://doi.org/10.1093/af/vfab012.
Tsukamoto K, Miura F, Fujito NT, Yoshizaki G, Nonaka M. Long-lived dichotomous lineages of the proteasome subunit Beta type 8 (PSMB8) gene surviving more than 500 million years as alleles or paralogs. Mol Biol Evol. 2012:29(10):3071–3079. https://doi.org/10.1093/molbev/mss113.
Våge R. Avlsarbeidet til NLA og A/S MOWI [Master's thesis]. [Molde (Norway)]: Molde University College; 1995.
Veríssimo A, Castro LFC, Muñoz-Mérida A, Almeida T, Gaigher A, Neves F, Flajnik MF, Ohta Y. An ancestral major histocompatibility complex organization in cartilaginous fish: reconstructing MHC origin and evolution. Mol Biol Evol. 2023:40(12):msad262. https://doi.org/10.1093/molbev/msad262.
Vigne J-D. The origins of animal domestication and husbandry: a major change in the history of humanity and the biosphere. C R Biol. 2011:334(3):171–181. https://doi.org/10.1016/j.crvi.2010.12.009.
Wang G-D, Xie H-B, Peng M-S, Irwin D, Zhang Y-P. Domestication genomics: evidence from animals. Annu Rev Anim Biosci. 2014:2(1):65–84. https://doi.org/10.1146/annurev-animal-022513-114129.
Wang M, Li X, Wang C, Zou M, Yang J, Li X-D, Guo B. Asymmetric and parallel subgenome selection co-shape common carp domestication. BMC Biol. 2024:22(1):4. https://doi.org/10.1186/s12915-023-01806-9.
Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. Evolution. 1984:38(6):1358–1370. https://doi.org/10.1111/j.1558-5646.1984.tb05657.x
Wickham H. Ggplot2: ggplot2. Wiley Interdiscip Rev Comput Stat. 2011:3(2):180–185. https://doi.org/10.1002/wics.147.
Wiener P, Wilkinson S. Deciphering the genetic basis of animal domestication. Proc R Soc Lond B Biol Sci. 2011:278(1722):3161–3170. https://doi.org/10.1098/rspb.2011.1376.
Witt KE, Huerta-Sánchez E. Convergent evolution in human and domesticate adaptation to high-altitude environments. Philos Trans R Soc Lond B Biol Sci. 2019:374(1777):20180235. https://doi.org/10.1098/rstb.2018.0235.
Wood TE, Bruke JM, Riesberg LH. Parallel genotypic adaptation: when evolution repeats itself. Genet. 2005:180(1):295–306. https://doi.org/10.1534/genetics.108.091173.
Xu P, Xu J, Liu G, Chen L, Zhou Z, Peng W, Jiang Y, Zhao Z, Jia Z, Sun Y, et al. The allotetraploid origin and asymmetrical genome evolution of the common carp Cyprinus carpio. Nat Commun. 2019:10(1):4625. https://doi.org/10.1038/s41467-019-12644-1.
Yao Z, Zhang Q, Li X, Zhao D, Liu Y, Zhao K, Liu Y, Wang C, Jiang M, Li N, et al. Death domain-associated protein 6 (Daxx) selectively represses IL-6 transcription through histone deacetylase 1 (HDAC1)-mediated histone deacetylation in macrophages. J Biol Chem. 2014:289(13):9372–9379. https://doi.org/10.1074/jbc.M113.533992.

Associate editor: Bonnie Fraser

Downloaded from https://academic.oup.com/gbe/article/17/4/evaf063/8115854 by guest on 08 January 2026

PAPER III

Recurrent emergence and ancient persistence of structural variants across continents in Atlantic salmon

Diblasi Célian[1], Barson Nicola[1], Saitou Marie[1]

[1] Department of Animal and Aquacultural Sciences, Section for Genome Biology, Faculty of Biosciences, Norwegian University of Life Sciences, 1433 Ås, Norway

Contact :

Célian Diblasi: celian.diblasi@nmbu.no

Nicola Barson: nicola.barson@nmbu.no

Marie Saitou: marie.saitou@nmbu.no

Author contribution:

The study was conceived by all authors. Data was generated by CD and MS. Bioinformatic and computational analyses were conducted by CD and MS. Interpretation of the results were done by all authors. Drafting paper was done by all authors. Manuscript editing was done by all authors.

Abstract :

Structural variants (SVs) are a major source of genomic diversity, yet their population-level dynamics remain poorly understood. We investigated the evolutionary origins of SVs across four Atlantic salmon (*Salmo salar*) lineages that differ in geography (Europe vs North America) and domestication status (wild vs farmed). Using sensitive discovery and stringent genotyping, we generated a high-confidence SV map and identified large rearrangements through local PCA. We found that SVs were enriched in repetitive regions, particularly segmental duplications and LTR retrotransposons, indicating mutational hotspots shaped by genome architecture. In addition, a majority of SVs were shared across the four lineages, suggesting long-term persistence of ancestral polymorphisms despite deep divergence. By contrast, a small subset of large SVs exhibited complex haplotype structures and multimodal sequence distance distributions consistent with recurrent origins. One immune gene–rich copy-number

variable region on chromosome 3 showed signs of domestication-related selection. Together, these results suggest that ancient maintenance and recurrent mutation jointly structure the SV landscape across continents.

Introduction:

Structural variants (SVs), including insertions, deletions, inversions, duplications, and translocations, are major sources of genomic diversity. Unlike single nucleotide polymorphisms (SNPs), which involve changes at individual base pairs, SVs can affect kilobase- to megabase-sized regions, altering gene dosage, regulatory landscapes, and chromosome structure. As such, SVs can have a greater impact on gene function, expression, and regulation compared to SNPs (Hämälä et al., 2021; Scott et al., 2021; Stuart et al., 2025). Recent advances in genome sequencing have facilitated the detection and characterization of SVs, yet their evolutionary significance remains poorly understood. While some SVs may be selectively neutral, others have been implicated in adaptive processes, reproductive isolation, and population divergence (Faria et al., 2019; Berdan et al., 2024; Hämälä et al., 2024). Beyond their mechanistic role in genome architecture, SVs play critical roles in evolutionary processes. They have been associated with local adaptation, phenotypic diversification, and reproductive isolation.

Despite their evolutionary importance, the population-level dynamics of SVs remain poorly understood. This gap in knowledge is partly due to technical limitations: SVs are often large, repetitive, complex, and challenging to genotype reliably with short-read sequencing, leading to their underrepresentation in population-scale studies. As a result, fundamental questions about the evolutionary trajectories of SVs have remained unresolved. In particular, whether SVs are observed across populations, and if so, through which processes, recurrence or selection has rarely been tested directly in natural systems (Lindtke et al., 2017; Wellenreuther et al., 2019; Aqil et al., 2023; Ferguson et al., 2024). Although short-read sequencing poses inherent challenges for

SV discovery, recent advances in bioinformatic methods have improved the ability to detect and genotype SVs with higher confidence (The Danish Pan-Genome Consortium et al., 2018; Kosugi, 2019). By applying such approaches, it is now possible to address long-standing questions about the origins and evolutionary dynamics of SVs even in species with large and complex genomes (Mérot et al., 2020; Berdan et al., 2021).

When shared SVs are detected across populations, several non-mutually exclusive mechanisms may underline them. These include the retention of ancestral polymorphism, particularly when divergence is recent; long-term maintenance by balancing selection, which may be especially relevant for SVs due to potential heterozygote advantage (Berdan et al., 2021). Recurrent mutations may also arise in structurally unstable genomic regions, such as those enriched in transposable elements and segmental duplications (Boettger et al., 2016). These regions are known hotspots for SV formation: repeated sequences promote non-allelic homologous recombination (NAHR), while transposable element activity can introduce double-strand breaks and erroneous repair (Kidd et al., 2010; Carvalho and Lupski, 2016). Hybridization or introgression can also transfer SVs across taxa (Almarri et al., 2020; Quan et al., 2021). Disentangling the relative contributions of these processes is critical for understanding how SVs shape evolutionary trajectories, yet few natural systems provide the necessary conditions for rigorous testing.

Atlantic salmon (*Salmo salar*) offers a particularly valuable model for addressing these questions. The species comprises two deeply diverged lineages, European and North American, that split roughly one million years ago and have since evolved largely in isolation (Payne et al., 1971; King et al., 2007). Both lineages have also undergone independent domestication in recent decades, providing replicated contrasts between wild and farmed populations exposed to distinct selective pressures. Furthermore, major karyotypic variation, including chromosome fusions in the North American lineage, combined with residual sequence homology from the

salmonid-specific whole genome duplication, creates a structurally dynamic genomic background prone to recurrent SV formation (Lubieniecki et al., 2010; Lien et al., 2016).

In this study, we leverage this unique system to investigate the evolutionary origins and trajectories of SVs in Atlantic salmon. By integrating high-confidence SV genotyping with genome-wide haplotype structure and functional annotations, we aim to disentangle the contributions of selection, recurrence, and shared ancestry to the distribution of SVs across populations with contrasting evolutionary histories.

Results:

Construction of a high-confidence SV map across Atlantic salmon lineages

To construct a high-confidence structural variant (SV) map, we analyzed four Atlantic salmon lineages comprising 368 individuals. We first applied Manta, a graph-based SV caller that integrates paired-end and split-read signals to achieve high sensitivity, identifying a total of 302,843 putative SVs. While highly sensitive, Manta's detection strategy may also capture variants with limited genotyping resolution. To improve accuracy in genotyping, which is particularly important for comparative analyses across lineages, we subsequently used BayesTyper, a k-mer-based probabilistic genotyper that jointly considers read support and variant presence probabilities.

This two-step approach, combining sensitive discovery with stringent probabilistic refinement, yielded a final set of 6,683 high-confidence SVs (hereafter, BayesTyper_SVs; **Supplementary Figure 1**). The substantial reduction in variant number reflects the conservative nature of BayesTyper's genotyping criteria, which are designed to minimize false positives, particularly in repetitive or low-coverage regions. Importantly, comparative analyses of retained and discarded SVs revealed that key features such as size and type were broadly preserved (**Supplementary Figure 2**), indicating that the refinement process enhanced genotyping reliability without introducing strong biases.

To better understand the characteristics of the high-confidence SV set, we next examined their size distribution, genomic context, and population-level patterns. The size distribution of BayesTyper_SVs exhibited two distinct peaks, with 27% of variants falling within the 50-100 bp range. A secondary enrichment was observed around 1,432-1,436 bp, consistent with previously reported signatures of the Tc1 mariner transposable element (Bertolotti et al., 2020). Among the 92 SVs in this range, 83 showed >95% identity to the Tc1 mariner reference (EF685967.1), corroborating earlier findings and supporting the technical robustness of our detection pipeline (**Figure 1A**). In terms of genomic distribution, SVs were significantly enriched in intronic regions compared to exons or coding sequences (**Figure 1B**). This pattern aligns with the expectation that purifying selection constrains the maintenance of structural variation in functionally important genomic regions, a trend observed in other vertebrate genomes as well (Zhou et al., 2019; NHGRI Centers for Common Disease Genomics et al., 2020; Hämälä et al., 2021).

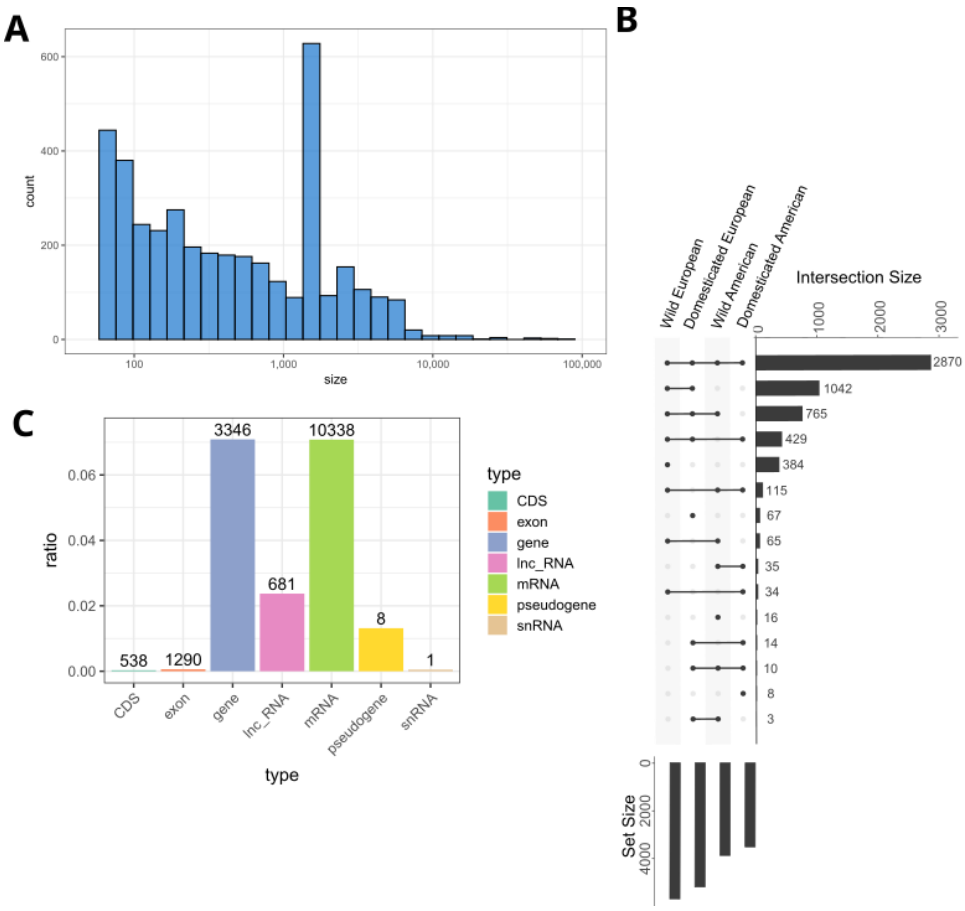


Figure 1: Overview of BayesTyper_SVs features

A: Size of SVs in bp on a log10 scale. **B:** Upset plot of the SVs in the four lineages. Lineages name indicates their domestication status (Domesticated vs wild) and geographic origin (American vs European). **C:** Proportion of SVs in genomic features, calculated for each annotation by taking the number of annotations overlapped for each type (indicated above the bars) divided by the total number of annotations.

SVs were enriched in repetitive regions, particularly LTR retrotransposons

To explore genomic features that may facilitate the formation of SVs, we assessed the extent to which BayesTyper_SVs overlapped with repetitive elements. Specifically, we focused on the overlap of SVs with segmental duplications and transposable elements,

both known to promote genome instability through mechanisms such as non-allelic homologous recombination (NAHR). Our analysis revealed that SVs overlapped with repetitive sequences at a significantly higher rate than expected by chance (**Figure 2**). SVs overlapping segmental duplications were particularly enriched, ranking in the top 0.5% of a null distribution generated from random genomic locations. This enrichment is consistent with the well-known role of segmental duplications in facilitating NAHR. These regions may also be generally more susceptible to the processes that form SVs. Taken together, this suggests that segmental duplications often coincide with genomic areas where SVs arise more frequently (Escaramís et al., 2015; Carvalho and Lupski, 2016). We further found that SVs were strongly enriched within or near LTR retrotransposons, both in terms of direct overlap and within 5 kb flanking regions. In contrast, DNA transposons showed a significant depletion of overlapping SVs. These contrasting patterns likely reflect fundamental differences in transposition mechanisms: LTR retrotransposons propagate via a "copy-and-paste" mechanism and often introduce larger structural changes, while DNA transposons use a "cut-and-paste" strategy, which is less likely to promote extensive rearrangements (Gray, 2000; Burssed et al., 2022).

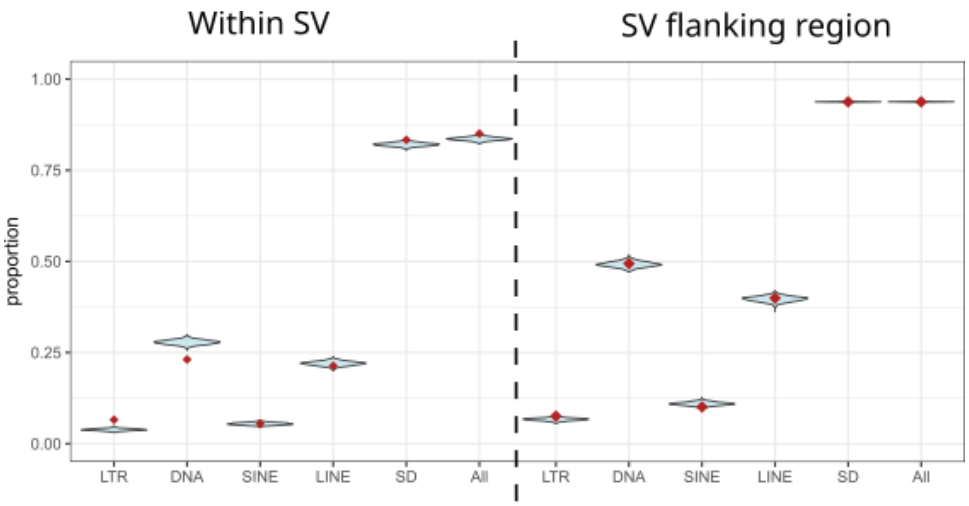


Figure 2: Proportion of SVs overlapped or surrounded by repeats compared to random expectations.

The blue distribution represents the proportion of BayesTyper_SVs overlapping (left) or surrounding (right, 5kb around the SV) the given repeat type for randomly located SV in 1000 hypotheticals dataset. The red dot indicates the observed proportion of SVs overlapping (or surrounding) the given repeat type. Notice that one SV can overlap or surround several repeat types.

Nearly half of SVs were shared across all four lineages

Principal component analysis (PCA) based on genome-wide SVs largely recapitulated known population structure from SNP data , showing clear continental differentiation and a domestication-related separation within Europe (**Supplementary Figure 3A**; Buso et al.), suggesting that many SVs are neutral. One individual of our dataset, however, displayed a highly divergent haplotype structure and was consistently isolated in PCA space across multiple chromosomes. Based on this consistent outlier pattern, we excluded the individual from downstream analyses to avoid technical distortion in population-level analysis.

Analysis of SV sharing revealed that 49% of BayesTyper_SVs were found in all four lineages (European and North American, wild and domesticated), a surprisingly high proportion given the deep divergence and limited recent gene flow between these groups (**Figure 1C**). In contrast, only 17.8% were shared exclusively between European lineages, and just 0.6% between North American lineages. Given the deep divergence and historical isolation between European and North American lineages, these results were not fully in line with expectations: SVs shared across all four lineages were more common than those shared within lineages of the same continent. This suggests a more complex history than previously assumed, potentially involving either long-term maintenance of ancient polymorphisms or recurrent SV formation at structurally unstable loci. We revisit these scenarios and their supporting evidence in subsequent sections. In addition, postglacial introgression from European lineages into North America, as well as historical admixture or contamination of domesticated North American stocks with European-origin fish, could also have contributed to the observed sharing. While these admixture scenarios are not explicitly tested here, they provide alternative, non-mutually exclusive explanations for the extensive SV overlap observed across lineages

Identification of large-scale structural variants using PCA-based clustering

Large structural variants (SVs) are known to influence genome evolution at macroscopic scales, contributing to phenotypic divergence, speciation, reproductive isolation, and karyotypic shifts by affecting chromosomal structure, and regulatory landscape. However, their detection remains challenging, particularly for variants exceeding the resolution of short-read sequencing (Mahmoud et al., 2019; Gong et al., 2021). To complement our previous SV discovery, we applied a window-based PCA clustering approach (**see Materials and Methods**, Mérot et al., 2021), which is designed to detect genomic regions with unusual haplotype structure potentially indicative of large SVs.

Using this method, we identified 25 genomic regions, ranging from 16.9 kb to 3.1 Mb, hereafter referred to as "Lostruct_SVs" (**Supplementary Figure 4, Supplementary Table 1**). These regions were further investigated using population genetic statistics to assess signatures consistent with large SVs. In most cases, heterozygous individuals showed elevated heterozygosity and nucleotide diversity (π), and comparisons between homozygote classes revealed high values of dXY and FST, consistent with strong differentiation and structural divergence (**Figures 3 and 4**). In addition, we observed that all identified SVs were present in at least one copy in all lineages. This observation is not expected from the method that is expected to have high sensitivity to non-shared SVs.

To further examine whether the regions detected by lostruct corresponded to discrete structural variants, we inspected the local haplotype structure within each candidate region. We performed principal component analyses (PCA) based on SNPs located within the putative SV boundaries to visualize potential genotype clusters among individuals. In most regions, the PCA revealed three clearly separated clusters, corresponding to the expected genotypic classes of a biallelic SV (AA, AB, and BB). However, a subset of regions displayed a more complex pattern, with six partially overlapping clusters (**Figure 4D**). This unexpected pattern suggested the possibility of more than two alleles segregating at these loci, as observed in other species (González et al., 2014; Knief et al., 2016). To evaluate this hypothesis, we compared heterozygosity across the six groups. If three distinct alleles were present, we would expect the three heterozygous combinations to show elevated heterozygosity relative to homozygotes. This pattern was not consistently observed **(Supplementary Figure 5**), providing limited support for a true multiallelic model. Nonetheless, these regions illustrate that haplotype structure in large SVs can be unexpectedly complex and may not always conform to simple biallelic expectations.

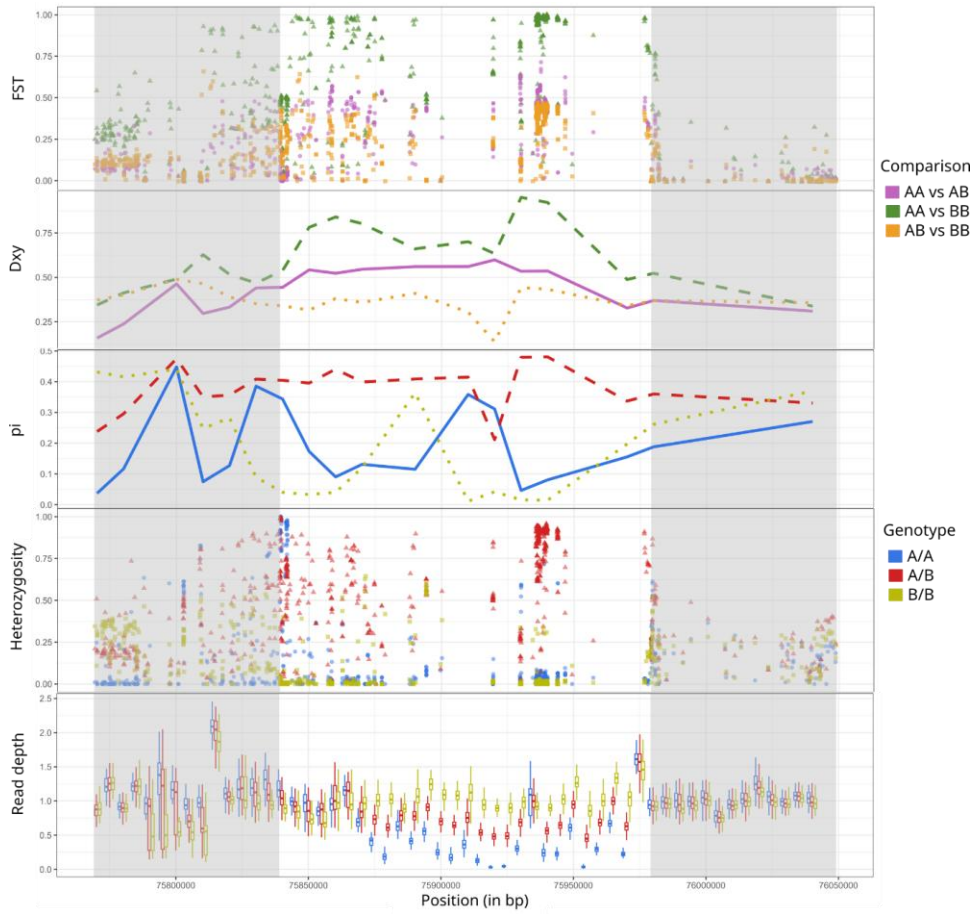


Figure 3: Population genetic summary statistics across one Lostruct_SV region.

The SV represented is located on ssa03:75839027 – 75979106. Each panel shows variation in genetic statistics along genomic position (x-axis) for the three inferred genotypes of the SV (A/A, A/B, B/B). Grey shading indicates the boundaries of the putative SV **A:** FST for each genotype comparison. **B:** dxy for each genotype comparison. **C:** Pi for each genotype: **D:** Heterozygosity for each genotype. E: Read depth for each genotype.

To test whether the large SVs detected through haplotype structure reflected underlying copy number variation (CNV), we examined normalized sequencing read depth across all 25 Lostruct_SV regions, stratified by SV genotype. Several regions, including ssa03:75.8–75.9 Mb (ssa03_75mb8_75mb9), ssa03:81.2–81.3 Mb, and

ssa27:11.8–12.3 Mb, showed clear genotype-specific depth shifts, consistent with deletions or duplications contributing to the SV signals detected via local PCA. Among these, ssa03_75mb8_75mb9 exhibited one of the most pronounced CNV-associated patterns (**Figure 4E**). BLAST searches against the Danio rerio genome revealed that ssa03_75mb8_75mb9 was particularly rich in immune-related genes: 8 out of 10 annotated genes showed high similarity to immunoglobulin sequences in other teleost fish species (≥85% identity, E-value < 1e−4) . Given the known rapid evolution of immune gene families and the high density of segmental duplications in this region, the structural variation in ssa03_75mb8_75mb9 likely reflects both genomic instability and immune-related functional specialization. No significant Gene Ontology (GO) enrichment was detected (FDR < 0.05) on the 21 Lostruct_SV regions overlapping annotated genes and many regions contained poorly annotated genes.

Contrasting frequency patterns of large SVs within and between continents

To assess how large structural variants (SVs) are retained across lineages, we examined allele frequencies of the 25 Lostruct_SVs identified through haplotype-based clustering. These SVs were prioritized because they show clear population differentiation signals and, due to their large size and proximity to functional regions, are more likely to be influenced by selection than smaller SVs.

Within each continent, we observed a strong positive correlation in SV frequencies between wild and domesticated lineages (Europe: slope = 0.65, R^2 = 0.4, p = 7.3 x 10^{-4}; America: slope = 0.83, R^2 = 0.79, p = 3 x 10^{-}9; **Figure 5A,B**). This pattern is consistent with expectations given the relatively recent origin of domestication (~15 generations). A single outlier, ssa03_75mb8_75mb9, the immune-related region on chromosome 3, showed a pronounced frequency increase in domesticated European salmon (0.69) relative to wild counterparts (0.19), deviating beyond the 95% confidence limits of the regression line. This deviation could result from either a founder effect or positive selection associated with domestication.

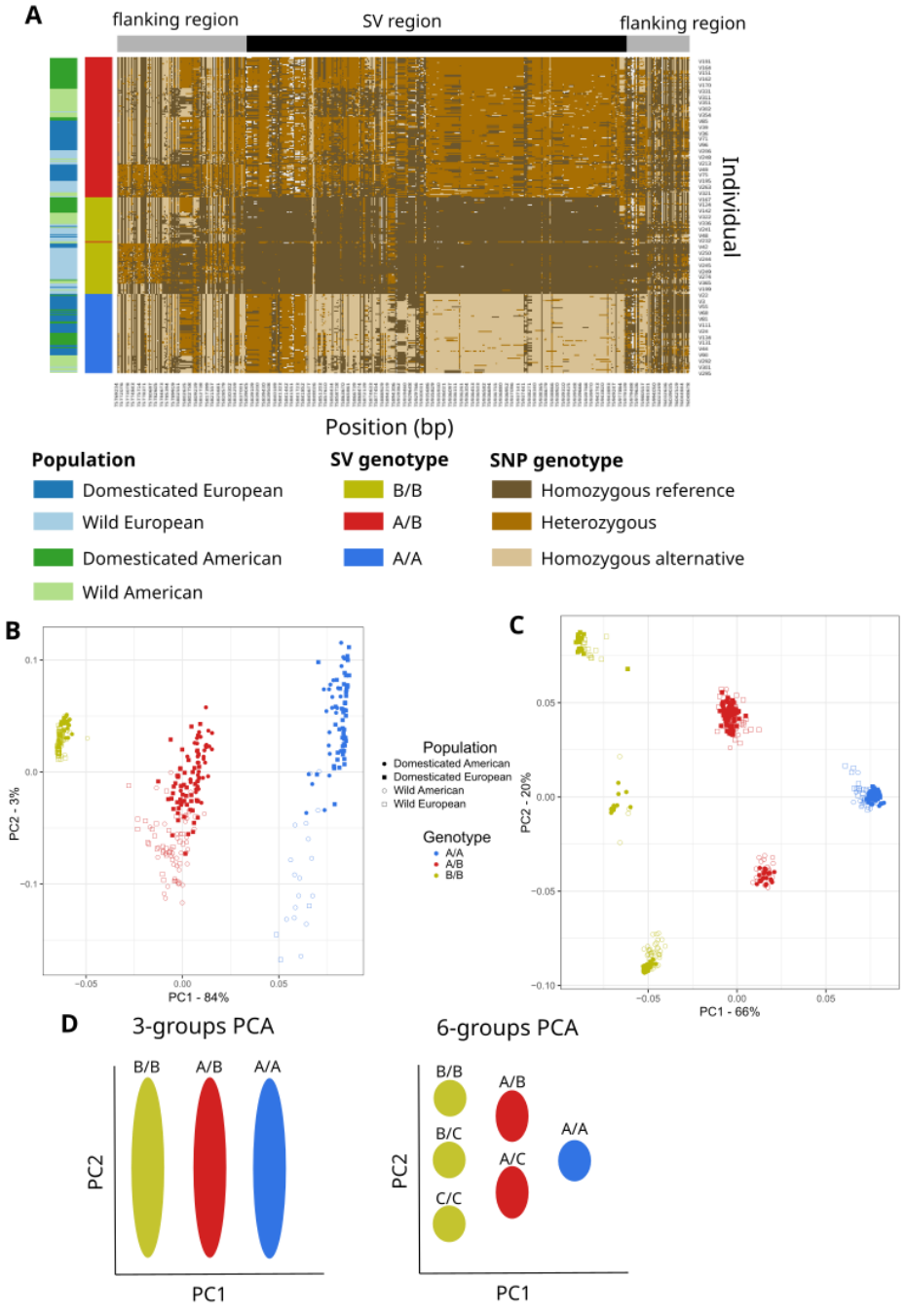

A
flanking region
SV region
flanking region
Individual
Position (bp)
Population
Domesticated European
Wild European
Domesticated American
Wild American
SV genotype
B/B
A/B
A/A
SNP genotype
Homozygous reference
Heterozygous
Homozygous alternative
B
PC2 - 3%
PC1 - 84%
Population
Domesticated American
Domesticated European
Wild American
Wild European
Genotype
A/A
A/B
B/B
C
PC2 - 20%
PC1 - 66%
D
3-groups PCA
B/B
A/B
A/A
PC2
PC1
6-groups PCA
B/B
A/B
B/C
A/C
A/A
C/C
PC2
PC1

Figure 4. Genotypic clustering and haplotype structure of large structural variants detected by local PCA.

(**A**) Haplotype visualization for the lostruct_SV ssa03:75839027 – 75979106. The heatmap represents individual SNP genotypes relative to the reference genome (light = homozygous reference, dark = homozygous alternative, intermediate = heterozygous). Colored bars on the left indicate lineage (green = Domesticated American, light green = wild American, blue = domesticated European, light blue = wild European). Bars on the right indicate the SV genotype inferred by invClust (yellow = A/A, red = A/B, blue = B/B). (**B, C**) Principal component analyses (PCA) within two representative SV regions (ssa03:75.8–75.9 Mb (**B**) and ssa05:44.0–44.1 Mb (**C**)). Each point represents an individual. Color denotes SV genotype (yellow = A/A, red = A/B, blue = B/B); shape denotes lineage (filled circle = domesticated American, open circle = wild American, filled square = domesticated European, open square = wild European). (**D**) Schematic representation of PCA cluster configurations among the 25 Lostruct_SVs. Left panel shows the typical three-group pattern expected for a biallelic SV (A/A, A/B, B/B). Right panel shows an unusually six-group pattern, potentially reflecting more complex or recurrent variation. Ellipses illustrate the relative position and separation of genotype clusters.

When comparing European and North American lineages, we expected a weakly positive correlation because many SVs likely predate continental divergence. Contrary to this expectation, allele frequencies were negatively correlated in both domesticated (slope = −0.77, R^2 = 0.43, p = 3.6 × 10^{-4}; **Figure 5C**) and wild (slope = −0.64, R^2 = 0.36, p = 1.4 × 10^{-4}; **Figure 5D**) groups. This contrast with the within-continent pattern suggests that ancient SVs may have experienced divergent or balancing selection pressures acting in opposite directions across continents.

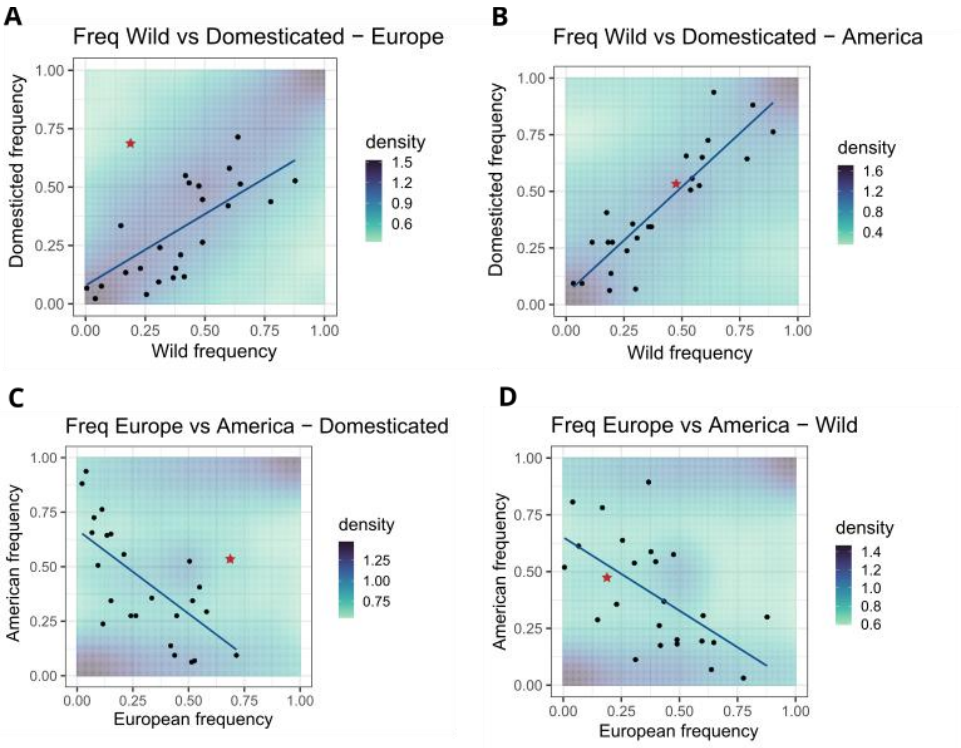


Figure 5. Frequency correlations of large structural variants (Lostruct_SVs) among populations.

(**A, B**) Allele frequencies of 25 Lostruct_SVs in wild (x-axis) versus domesticated (y-axis) populations for Europe (**A**) and North America (**B**). Each point represents a single SV; background shading indicates the frequency density of bayestyper_SVs. The red star marks ssa03_75mb8_75mb9, which shows a strong frequency shift in domesticated Europe. (**C, D**) Comparison of SV frequencies between European (x) and American (y) populations for domesticated (**C**) and wild (**D**) groups. Blue lines show regression fits; shading indicates the frequency density of bayestyper_SVs. The red star again corresponds to ssa03_75mb8_75mb9.

Previous SNP-based selection scans, including XPEHH analyses conducted in the same American and European lineages pairs investigated here, reported several candidate regions under artificial selection during domestication (Buso et al., 2025). Among the SVs identified in the present study, 3 Lostruct_SVs in the American population and 3 SVs (1 Lostruct_SV, 2 Bayestyper_SVs) in the European lineage overlapped with these

previously reported candidate regions (**Supplementary Figure 6, Supplementary Table 2**) (FDR = 10-4).

Simulated evolutionary scenarios reveal distinct retention patterns of ancient and recurrent SVs

To better understand the evolutionary processes shaping SV diversity, we simulated forward-time models of SV evolution under different scenarios using SLiM v4.2.2 (Haller and Messer, 2023) (**see Materials and Methods**). This was motivated by the unexpectedly high proportion of SVs shared across the four sampled lineages, which could result either from recurrent mutations or from long-term maintenance of ancient polymorphisms by selection.

Ancient SVs (SVs arising before lineage divergence) were consistently more likely to be retained across all lineages than recurrent SVs (**Figure 6A**), regardless of the recurrence rate (**Supplementary Figure 7**). Under balancing selection scenarios, recurrent SVs were more likely to persist than under directional selection, whereas the opposite trend was observed for ancient SVs. Ancient SVs with directional selection tended to become rapidly fixed and were therefore excluded from further analysis, as fixed variants would not be detectable in lineage comparisons (**Figure 6A**). Overall, the simulations indicate that SVs shared across all four lineages are more likely to represent ancient polymorphisms than recurrent events.

To further characterize the evolutionary origins of the detected structural variants, we examined the distribution of pairwise haplotype distances (1 − IBS) among individuals homozygous for alternative SV genotypes. For each SV, we compared three distance categories: between individuals homozygous for haplotype 1 (G1–G1, AA vs AA), between individuals homozygous for haplotype 2 (G2–G2, BB vs BB), and between the two homozygote groups (G1–G2, AA vs BB). Forward simulations indicated that SVs of ancient origin typically yield unimodal distance distributions, with within-group distances forming a single peak and between-group distances showing a single, broader mode corresponding to deep divergence between

haplotypes. In contrast, recurrently arising SVs are expected to produce multimodal distributions, as independent mutational events generate several distinct haplotypes within the same genotypic class.

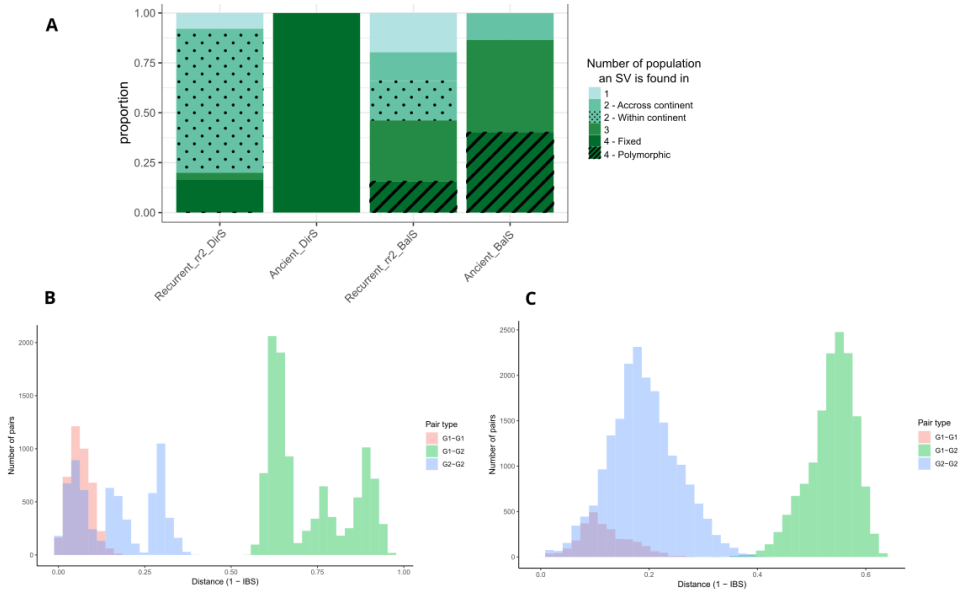


Figure 6: Evolutionary history of simulated SV and inferred recurrent SVs

A: Labels indicate the history of SVs (recurrent or ancient), as well as the selection type (directional selection = DirS; balancing selection = BalS) and the recurrent rate (rr2 = 2 SVs per simulation on average). Simulations were run with a selection coefficient of 1.1. **B-C**: Genetic distance between SNPs in haplotype, G1-G1: AA vs AA; G1-G2:AA vs BB; G2-G2: BB vs BB, for two lostruct_SVs. Under recurrent SV, the distribution is expected to be multimodal for G1-G2 (**B**, ssa05:44020000-44084228), which is not the case for other SVs (**C**, ssa12:63150000-63800000)

When applying this simulation-informed framework to the empirical data, most SVs showed the unimodal "ancient" pattern. However, four loci exhibited multimodal distance distributions consistent with recurrent origin. A representative example is shown in **Figure 6B**. In this case, distances among G1 individuals (G1–G1, AA vs AA, red) form a single narrow peak, indicating a uniform haplotype background. In

contrast, distances among G2 individuals (G2–G2, BB vs BB, blue) and between G1 and G2 (G1–G2, AA vs BB, green) form three distinct peaks, reflecting the coexistence of multiple divergent haplotypes classified under the same genotypic state. This pattern implies that the G2 allele has arisen independently more than once, producing several haplotype lineages that now segregate in the population. All four loci displaying this pattern (ssa16_33750000_34020915, ssa05_44020000_44084228, ssa16_60620000_60880000, ssa21_38333640_38459102, **Supplementary Figure 8**) also exhibited complex six-cluster structures in PCA space (exemplified in **Figure 4D**). We found that these four regions were surrounded by more TEs than random regions but not significantly more than other Lostructs_Svs (**Supplementary Figure 9**). For BayesTyper_SVs, similar analysis did not reveal clear evidence of recurrent SVs, in particular due to the short size of these SVs, not suited for this haplotype comparison method.

Discussion:

Structural variants (SVs) contribute substantially to genome evolution, adaptation, and speciation. However, their evolutionary dynamics remain less understood than those of single-nucleotide polymorphisms, particularly in animals with large and repetitive genomes.

Here, we combined short-read SV genotyping, haplotype-based clustering, and forward simulations to investigate the evolutionary history of SVs across four Atlantic salmon lineages differing in geographic origin and domestication status. This integrative framework allowed us to contrast ancient polymorphisms maintained across lineages with recurrent mutations occurring at structurally variable genomic regions.

Nearly half of SVs were shared among all four lineages, suggesting the long-term persistence of ancestral polymorphisms, while a small number of large SVs exhibited multimodal haplotype-distance distributions and complex six-cluster PCA patterns indicative of recurrent origin. One notable SV on chromosome 3, encompassing

immunoglobulin-related genes, showed a marked frequency increase in domesticated European salmon, consistent with recent artificial selection. These findings reveal that the SV landscape of Atlantic salmon reflects both deep evolutionary maintenance and ongoing structural turnover.

Structural variants in Atlantic salmon were strongly enriched within or near repetitive regions, particularly segmental duplications and LTR retrotransposons, indicating that the genomic context plays a major role in their formation. Such enrichment is consistent with the well-established role of repetitive sequences in promoting non- allelic homologous recombination (NAHR) and replication-based rearrangements (Kidd et al., 2010; Carvalho and Lupski, 2016). Segmental duplications provide extensive sequence homology that facilitates ectopic pairing, leading to recurrent deletions, duplications, or inversions at the same loci (Emanuel and Shaikh, 2001). The strong enrichment we observed among LTR retrotransposons further suggests that retroelement proliferation contributes to ongoing genome instability. Because LTR elements propagate by a copy-and-paste mechanism, they can mediate unequal crossover and generate large insertion–deletion polymorphisms, whereas DNA transposons, which rely on a cut-and-paste mechanism, are less likely to promote large rearrangements (Gray, 2000; Burssed et al., 2022).

These observations indicate that SV formation in the Atlantic salmon genome is not random but is biased toward structurally variable regions. This bias provides a mechanistic explanation for why certain loci repeatedly generate SVs, as seen for the recurrent variants detected in later analyses. The co-localization of SVs with repetitive and duplicated regions thus highlights the role of intrinsic genome architecture in shaping the mutational landscape of large structural variation.

Shared SVs across lineages and the persistence of ancestral polymorphisms

Nearly half of the high-confidence Bayestyper_SVs were shared among all four Atlantic salmon lineages, an unexpectedly high proportion given the deep divergence and long-term geographic isolation between European and North American lineages.

This extensive sharing cannot be readily explained by recurrent mutation alone, as our simulations showed that recurrent SVs are rarely retained across multiple lineages. Instead, it indicates that a large fraction of SVs represents ancient polymorphisms that have persisted since before the continental split.

The preferential localization of these shared SVs in intronic and intergenic regions further supports this interpretation, as such regions are less affected by purifying selection and more likely to tolerate long-term structural variation. Similar patterns have been reported in other taxa, where chromosomal inversions or copy-number variants have been maintained for millions of years under balancing or fluctuating selection (Lindtke et al., 2017; Göktay et al., 2021; Saitou et al., 2021; Aqil et al., 2023). The persistence of ancestral SVs may therefore reflect a combination of neutrality and balancing forces that prevent fixation or loss. Moreover, within each continent, SV frequencies were positively correlated between wild and domesticated lineages, consistent with recent shared ancestry and largely neutral dynamics.

Large-scale SVs and complex haplotype structures

While most SVs exhibited patterns consistent with long-term maintenance, a subset of large-scale variants revealed unexpectedly complex haplotype structures. Using local PCA analyses, we identified 25 genomic regions (Lostruct_SVs) showing pronounced clustering of haplotypes, consistent with the presence of large rearrangements. In most regions, individuals grouped into three discrete clusters corresponding to the genotypic classes expected for a biallelic SV (AA, AB, and BB). However, several loci deviated from this pattern and displayed six partially overlapping clusters, implying that multiple distinct haplotypes segregate at the same genomic position. Former examples of such multilayered clustering revealing complex variants with evidence for tri-allelic loci (González et al., 2014; Knief et al., 2016; Lehnert et al., 2025). However, the observed SVs in our study did not match these expectations, suggesting a more complex history. Our forward simulations predicted that recurrent SVs can produce this type of multilocus haplotype configuration. Consistent with this expectation, these same loci exhibited multimodal

haplotype-distance distributions in the empirical data, suggesting the presence of multiple divergent haplotypes within the same genotypic category. Such recurrent structural changes contribute to "cryptic" haplotype diversity that may be overlooked when SVs are treated as simple, single-origin, biallelic variants.

Functional relevance and domestication-related selection

Among the large structural variants identified, the region ssa03_75mb8_75mb9 on chromosome 3 showed a pronounced allele-frequency increase in domesticated European salmon (0.19 in wild versus 0.69 in domesticated individuals), while no comparable shift occurred in North American populations. Although a founder effect associated with the establishment of domesticated stocks could contribute to the observed frequency shift, the pattern remains consistent with context-dependent selection acting on structurally unstable immune regions. The biological context of the region points to a plausible target of selection. The locus overlaps with a copy-number variable segment enriched for immunoglobulin-like genes, situated within a cluster of segmental duplications. Such architecture is well known to facilitate both recurrent rearrangements and rapid immune-gene evolution (Traherne et al., 2010; Lin and Gokcumen, 2019; Rodriguez et al., 2023). Read-depth profiles further indicate that this locus encompasses both copy-number losses among genotypic classes. Under aquaculture conditions, characterized by high population density and elevated pathogen exposure, selection on immune-related copy-number variation is plausible (Krkosek, 2010; Li et al., 2017; Buso et al., 2025).

Similar immune-related SVs have been implicated in adaptation to parasite pressure and disease resistance in other species, supporting the notion that structural variation can serve as an important substrate for rapid immune evolution (Penso-Dolfin et al., 2020; Göktay et al., 2021; Karageorgiou et al., 2024). The example of ssa03_75mb8_75mb9 thus illustrates how functional gene content and artificial selection can shape the evolutionary trajectory of specific SVs in domesticated salmon.

Frequency divergence across continents

The contrasting patterns of allele-frequency correlations within and between continents provide important insights into the evolutionary forces shaping large SVs in Atlantic salmon. Within each continent, frequencies were positively correlated between wild and domesticated lineages, consistent with recent shared ancestry and largely neutral evolution during the short timescale of domestication. In sharp contrast, frequencies across continents were negatively correlated for both wild and domesticated lineages. Such a pattern is unlikely to result from recurrent mutation, which would not systematically produce inverse frequency shifts across multiple loci, and instead suggests that ancient SVs have been maintained under divergent selective pressures.

Divergent selection between continents may arise from long-term differences in environmental conditions, such as temperature regimes, photoperiod, or pathogen communities, that differentially affect the fitness of alternative SV alleles (Lehnert et al., 2020). Under this scenario, the same ancestral variants could be favored in one continent but disfavored in the other, producing the observed negative correlation. Alternatively, asynchronous selection through time, where the relative advantage of each allele fluctuates, could also generate similar frequency reversals among geographically isolated populations (Lehnert et al., 2025). Such temporally or spatially varying selection regimes are well known to maintain polymorphism in other taxa and they provide a parsimonious explanation for the widespread retention of SVs across divergent lineages despite continental-scale differentiation.

These findings imply that the SV landscape of Atlantic salmon reflects a long-term balance between geographically variable selection, and historical demographic events. Ancient SVs have persisted across continents, but their allele frequencies have been reshaped by region-specific selective pressures, resulting in parallel yet directionally opposed genetic landscapes.

Simulations and inference of recurrent SVs

Forward-time simulations provided a theoretical context for interpreting how different evolutionary processes can produce the observed patterns of SV sharing. Under a wide range of parameters, recurrent mutations were predicted to be short-lived or geographically restricted, whereas ancient variants were expected to persist across lineages under neutrality or balancing selection. These simulations thus predict that widespread SVs shared across lineages most likely represent long-standing polymorphisms, while SVs restricted to one or a few lineages may have originated through recurrent mutation.

Applying this simulation-informed framework to empirical data allowed us to identify a small number of loci that follows the recurrent scenario. Four PCA_SVs exhibited multimodal haplotype-distance distributions (1–IBS) and complex six-cluster PCA configurations, patterns predicted by simulations for recurrent mutations. In these loci, one haplotype class was genetically homogeneous, whereas the alternative class comprised multiple divergent haplotypes, suggesting that similar structural rearrangements have arisen independently at the same genomic sites. Such patterns are difficult to explain by ancestral polymorphism alone and are instead consistent with recurrent mutation in regions of structural instability. The convergence of structural and population-genetic evidence thus indicates that recurrent SV formation is an ongoing process in the Atlantic salmon genome. Although these events are relatively rare compared to the abundance of ancestral SVs, their occurrence highlights the dynamic nature of genome evolution, where ancient variation provides continuity while recurrent mutation continuously introduces new genomic configurations.

Methodological considerations and future directions

Our approach combines population-scale short-read genotyping, haplotype-based clustering, and forward simulations to investigate the evolutionary history of structural variants. While this integrative framework provides a coherent view of SV

dynamics, several technical aspects warrant consideration for future refinement. Short-read sequencing inevitably limits the resolution of complex rearrangements, particularly at repetitive breakpoints, where mapping ambiguity can lead to merged or fragmented variant calls. Because our pipeline was conservative against false positives and ambiguous variants, some true variants were likely missed; however, the variants retained after stringent filtering are expected to represent a highly reliable subset of the SV landscape.

The identification of recurrent-like SVs was based on haplotype-distance patterns derived from simulations and qualitative inspection of multimodality. Although not strictly quantitative, this simulation-informed criterion provides a consistent and transparent basis for distinguishing recurrent from ancient origins. Similar qualitative approaches have been successfully applied in studies of inversion polymorphisms and complex haplotype architectures in natural populations (Boettger et al., 2016; Xu et al., 2017; Porubsky et al., 2020; Schaal et al., 2022). Future work integrating long-read sequencing, optical mapping, and direct breakpoint reconstruction will allow more precise quantification of recurrence rates and mutational mechanisms.

As an additional consideration, hybridization was not explicitly modeled in our analytical framework. Although the lineages examined here are generally thought to have experienced long-term geographic isolation, postglacial introgression events have been documented in parts of their range, and traces of European ancestry may persist in some domesticated North American populations (Bradbury et al., 2022). Such histories, if present, could contribute to a small fraction of shared SVs. Future analyses incorporating explicit admixture models will help clarify the extent to which hybridization has influenced the observed patterns.

Finally, integrating structural variation with transcriptomic and phenotypic data will be essential to evaluate the functional consequences of SV polymorphisms. Finally, integrating structural variation with transcriptomic and phenotypic data will be important for evaluating the functional consequences of SV polymorphisms. Across diverse taxa, SVs have been shown to contribute to gene expression variation,

regulatory divergence, and adaptive phenotypes beyond what is captured by SNP-based analyses alone (GTEx Consortium et al., 2017; Alonge et al., 2020; Scott et al., 2021; Falker-Gieske et al., 2022; Zhang et al., 2024). In Atlantic salmon, applying comparable integrative approaches allows these general effects to be examined in a system shaped by recent and rapid domestication, deep continental divergence, and a structurally dynamic genome influenced by whole-genome duplication and lineage-specific chromosomal rearrangements.

Material and methods:

Sampling design and whole-genome sequencing data

We analyzed whole-genome sequencing data from four Atlantic salmon lineages representing two continents (Europe and North America) and two domestication status (wild and domesticated). In total, 369 individuals were initially considered. Farmed North American samples (n = 80) originated from three source rivers in North America (NCBI Project PRJEB34225). Farmed European samples (n = 112) belonged to the MOWI strain, derived from the River Bolstad in the Vosso drainage, the River Årøy, and marine captures near Osterfjord and Sotra in Western Norway (Glover et al., 2009, NCBI Project No. PRJEB47441). As wild counterparts, we used 79 wild North American individuals collected from eight rivers in Québec and 98 wild European individuals from 17 rivers in Western Norway (Bertolotti et al., 2020, NCBI Project No. PRJEB38061).

Paired end Illumina reads from all individuals were mapped to the Atlantic salmon reference genome Ssal_v3.1 (GCA_905237065.2) using bwa-mem2 (Li and Durbin, 2009). One wild North American sample that could not be reliably genotyped for structural variants was removed from downstream analyses.

SNP detection

SNP calling followed the pipeline described in Buso et al., 2025 for the mapping and variant calling. Mapping, variant calling and quality control were performed

independently for each of the four datasets (Farmed Norwegian, Wild Norwegian, Farmed North American, Wild North American) using the same pipeline and parameters. Illumina reads were mapped to the Atlantic salmon genome (Ssal_v3.1; GCA_905237065.2) using bwa-mem2 (Li and Durbin, 2009). Duplicate reads were marked with gatk4-spark:4.3.0.0 MarkDuplicates. Genetic variation was identified using gatk4-spark:4.3.0.0 HaplotypeCaller with default parameters and the individual genotypes were merged with gatk4-spark:4.3.0.0 GenotypeGVCFs. To avoid the detection of false positives, variants were filtered by the following parameters with gatk4-spark:4.3.0.0 (McKenna et al., 2010) : "QD < 2.0" which filters out variants with a low quality score by depth of coverage; "QUAL < 50.0" to sort variants with a low-quality score; SOR > 4.0 that removes variants where the number of reads supporting the reference allele is significantly skewed towards one strand (forward or reverse) compared to the other; FS > 60.0: to filter out variants whose reads supporting the reference allele versus the alternative allele are significantly different between the two strands; and finally ReadPosRankSum < -8.0: to remove variants with a low score for the difference between the mean position of the reference allele and the alternative allele among the reads supporting the variant (https://gatk.broadinstitute.org/hc/en-us).

Variants detected in all lineages were then merged together in a single file using bcftools v.1.9 (Danecek et al., 2021), in order to homogenize variants across all populations. Variants with missingness higher than 10% were removed using vcftools v0.1.16 (Danecek et al., 2011).

Read-based structural variant discovery and genotyping

Initial discovery of structural variants (SVs) relied on read-based calling. For each individual, we used Manta v1.6.2 (Chen et al., 2016) to detect candidate SVs using paired-end and split-read information. Within each lineage, individual-level calls were merged using SURVIVOR v1.0.7 (Jeffares et al., 2017) with parameter "100 1 0 0 0 30". in order to cluster SVs with similar breakpoints and to construct a lineage-wide

candidate SV set. Variants were first filtered by Manta quality score, retaining only SVs with QUAL > 100 (**Supplementary Figure 1**).

Because the Manta pipeline alone does not clearly distinguish between truly absent variants and uncalled genotypes, and because precise genotyping was essential for comparative analyses, we refined SV genotypes using BayesTyper v1.5 (The Danish Pan-Genome Consortium et al., 2018) BayesTyper uses a k-mer based probabilistic model that combines read support and expected genotype configurations. We provided BayesTyper with the SVs discovered by Manta and followed the batch processing recommendations from the official documentation (**Supplementary Figure 1**). Due to computational limitations, individuals were divided into batches of five, yielding 74 batches that were genotyped separately and then merged using bcftools (Danecek et al., 2021) as recommended in the bayestyper guidelines.

We removed SVs with more than 75 % missing genotypes and a BayesTyper quality score below 10 using vcftools v0.1.16 (Danecek et al., 2011). During the BayesTyper step, information on SV type was not retained, so SV types were reassigned by matching the genomic coordinates of BayesTyper-genotyped variants to the original Manta calls (**Supplementary Figure 1**). We identified one individual that appeared as a consistent outlier in genome-wide principal component analyses was excluded from population genetic analyses using this data set. The final high-confidence SV set, referred to as BayesTyper_SVs, contained 6,683 variants and formed the basis for genome-wide analyses of SV size distribution, genomic context, and sharing across lineages.

Annotation of genomic context and repetitive elements

Gene annotations for the Ssal_v3.1 reference genome were used to assign SVs to genomic features. We used bedtools v2.30 (Quinlan and Hall, 2010) to intersect BayesTyper_SVs with gene models and to classify variants according to their location in exons, introns, or intergenic regions.

To investigate the association between SVs and repetitive elements, we first built a de novo transposable element (TE) library for Atlantic salmon using RepeatModeler v2.0.4 (Flynn et al., 2020). Elements classified as "unknown" by RepeatModeler were reclassified using DeepTE tool (Yan et al., 2020). To reduce redundancy among TE families and to avoid double counting similar families under different names, we clustered the TE consensus sequences using CD-HIT v4.8.1 following recommended parameters from published TE curation workflows (Goubert et al., 2022).

Genome-wide TE annotation was then performed using the curated TE library. Segmental duplication tracks were retrieved from the UCSC Genome Browser for Ssal_v3.1. Using bedtools v2.30 (Quinlan and Hall, 2010), we calculated the proportion of BayesTyper_SVs that overlapped or were located within 5 kb of each repeat category, including segmental duplications, LTR retrotransposons, DNA transposons, and other TE classes. To assess enrichment relative to random expectation, we generated null distributions by simulating sets of random genomic intervals matched to the observed SVs in size and chromosome and computing the same overlap statistics.

To investigate the Tc1 mariner peak observed in the SV size distribution (Bertolotti et al., 2020), we extracted sequences for SVs in the 1,432-1,436 bp size range and aligned them against the Tc1 mariner reference sequence EF685967.1 using BLASTn

(Korf et al., 2003). SVs showing at least 95 % sequence identity were classified as Tc1 mariner associated.

Detection of large structural variants using local PCA

Large structural variants are difficult to detect via short-read based callers which often can not span the whole SV. To complement the read-based pipeline, we used a local SNP-based PCA approach to detect large SVs implemented in the R package lostruct (Li and Ralph, 2019). Genome-wide SNP data were partitioned into windows of 1,000 SNPs along each chromosome, and principal component analyses were performed in each window using the run_lostruct.R script. We then used the built-in

"summarize_run.Rmd" script to identify windows with extreme PCA configurations relative to the genomic background.

Most windows displayed PCA patterns typical of genome-wide background structure and were therefore not retained for further analysis. Instead, we focused on windows exhibiting clear deviations from this background pattern, which are indicative of structurally divergent haplotypes expected under the presence of large SVs. These atypical windows were subsequently clustered using k-means based on their major principal components. For each candidate region, we merged adjacent windows with similar PCA structure, resulting in 51 putative SV regions.

For these regions, we calculated FST between the inferred genotype groups using vcftools and visually inspected FST along the chromosome. Regions showing extended blocks of elevated FST were retained, and their approximate breakpoints were refined based on the FST profiles. Genotypes for each region were then assigned using invClust (https://github.com/isglobal-brge/invClust/), which infers karyotypes of inversion-like polymorphisms from SNP data. Regions in which only two genotype classes could be confidently resolved were excluded. After this filtering, 25 regions remained and are referred to as Lostruct_SVs (**Supplementary figure 5, Supplementary table 1**).

Population genetic statistics and potential copy number variants

For each Lostruct_SV, we calculated population genetic statistics within and between inferred genotype classes. Heterozygosity and nucleotide diversity (π) were computed within each genotype using the R packages vcfR and adegenet (Jombart, 2008; Knaus and Grünwald, 2017). Between-genotype differentiation was summarized using dXY and FST, calculated with pixy v1.2.10.beta2 and vcftools respectively (Danecek et al., 2011; Korunes and Samuk, 2021).

To test whether Lostruct_SVs corresponded to underlying copy number variation, we estimated per-base read depth using mosdepth and summarized normalized coverage across SV regions for each genotype class. Regions showing clear genotype-

specific coverage shifts were interpreted as candidates for deletions or duplications contributing to the haplotype structure detected by local PCA.

These statistics were computed for the SV region including flanking regions corresponding to 50% of the size of the SVs (upstream and downstream). The PCAs displayed in Figure 4B and C were only computed on the SV region itself, to only reflect the structure of the SV.

Gene content and functional annotation of large SVs

For Lostruct_SVs overlapping annotated genes, we extracted gene models using bedtools and summarized gene content. To investigate the gene content of the immune-related region ssa03_75mb8_75mb9, we conducted BLASTn (Korf et al., 2003) searches of its annotated genes against the Danio rerio genome GRCz11 and recorded matches with at least 85 % identity and E-values < 1e-4, which were predominantly immunoglobulin genes.

We used ShinyGO (V 0.80, http://bioinformatics.sdstate.edu/go/) to perform Gene Ontology (GO) enrichment analyses for Lostruct_SVs containing annotated genes. No GO terms remained significant after multiple testing correction at FDR < 0.05.

Simulations of structural variant evolution

To explore the processes underlying the sharing of SVs among lineages, we performed forward-time simulations using SLiM v4.2.2 (Haller and Messer, 2023) following a framework similar to Berdan et al., 2021, with different types of selection (directional selection or heterozygote advantage).and different types of evolutionary history (variant being ancestral or appearing in parallel). We modeled a four-population system representing European and North American, wild and domesticated lineages, and simulated structural variants under alternative scenarios.

In “ancient SV” scenarios, SVs arose before lineage divergence and could be maintained or lost in each descendant lineage. In “recurrent SV” scenarios, SVs could

arise independently on different haplotype backgrounds after lineage divergence, allowing for multiple mutational origins. For both ancient and recurrent SVs, we considered different selection regimes, including directional selection and balancing selection. For each combination of parameters, we performed 10,000 replicate simulations and retained only those in which the SV remained polymorphic in at least one lineage, because fixed variants would not be detectable in our lineage comparisons.

SNP mutations were simulated with a rate of 1.06×10^{-8} per site per generation (Stenløkk et al., 2022). Each scenario was run for 2,000 generations. For analysis of haplotype structure, we sampled 50 individual genomes per population from the final generation. Full details of demographic settings, recombination rates, selection coefficients, and recurrence rates are provided as scripts on GitHub.

Haplotype distance and IBS analyses

To distinguish between ancient and recurrent SV origins, we examined pairwise haplotype distances within and between SV genotype classes. For each Lostruct_SV and BayesTyper_SV, we extracted SNPs within the putative SV boundaries and computed 1 − IBS using PLINK v2 and 1.9 (Chang et al., 2015) treating 1 − IBS as a measure of haplotype divergence.

For each SV, we defined three distance categories, namely distances between individuals homozygous for genotype 1 (G1-G1, AA vs AA), between individuals homozygous for genotype 2 (G2-G2, BB vs BB), and between the two homozygote groups (G1-G2, AA vs BB). The distributions of these three-distance classes were compared to expectations from the SLiM simulations. Ancient SVs are expected to produce unimodal distance distributions with a single broad mode for between-genotype distances, whereas recurrent SVs are expected to generate multimodal patterns reflecting multiple independent haplotype backgrounds.

We used these patterns to classify candidate recurrent SVs among Lostruct_SVs and to confirm that BayesTyper_SVs, which are typically shorter, did not provide sufficient haplotype information to robustly detect recurrent origins.

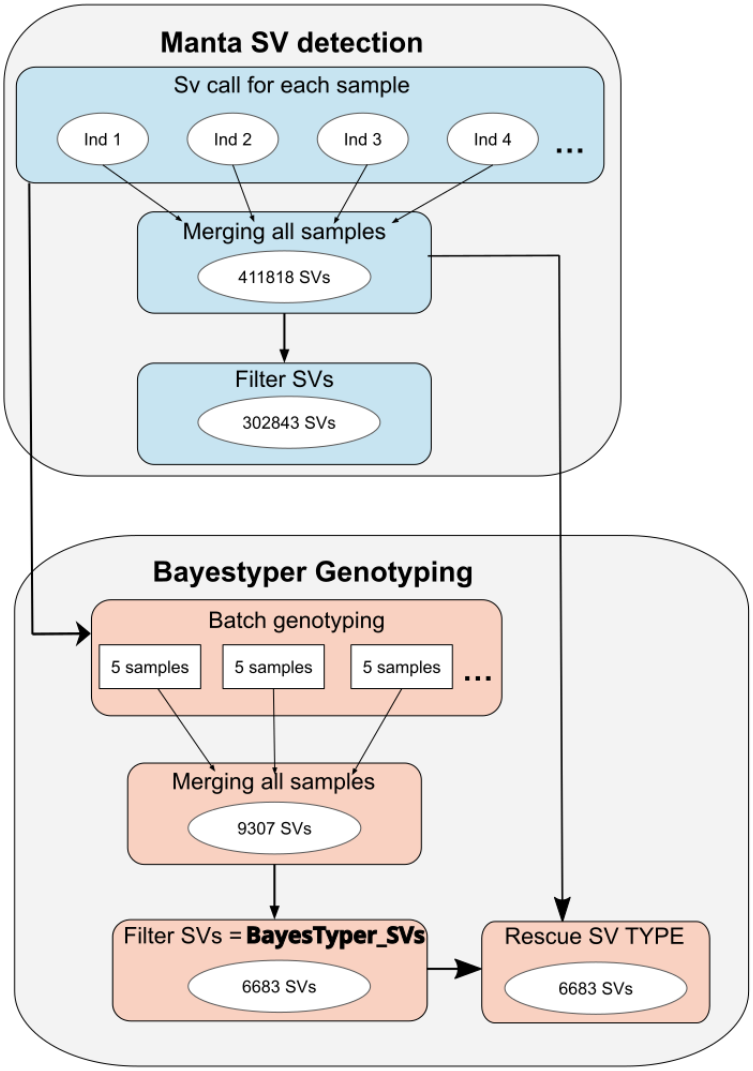
Data and Code availability

All scripts used for read mapping, variant calling, structural variant genotyping, annotation, generation of transposable elements library, population genetic analyses, simulations, and figure generation are available at https://github.com/celiand/SV_evolutionary_history, together with configuration files and example commands enabling full reproducibility of the analyses described in this study.

Acknowledgments:

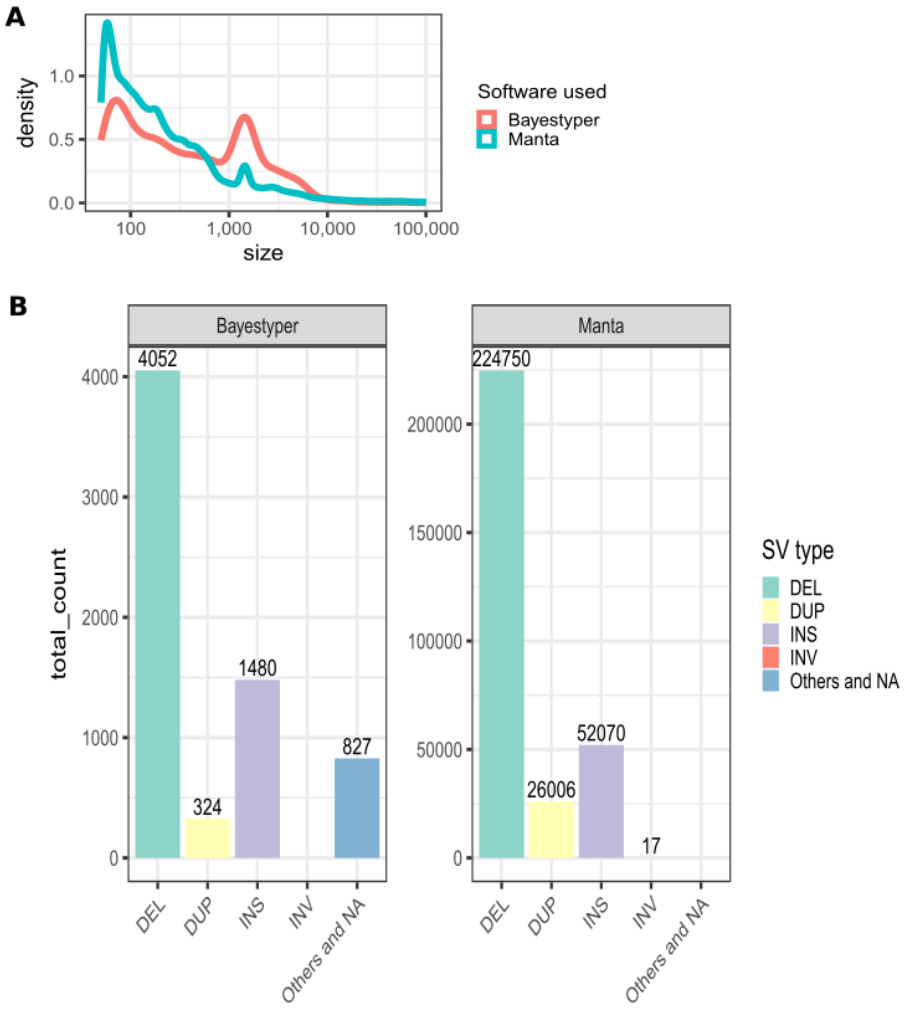
We acknowledge the use of the Orion computing cluster at the Norwegian University of Life Sciences (NMBU). We thank Erik Sandertun Røed from Centre for Ecological and Evolutionary Synthesis – University of Oslo, for his help and advice with SLiM simulations. The study was supported by the Research Council of Norway (grant nos. 325874, 275310, and 221734).

Supplementary materials:



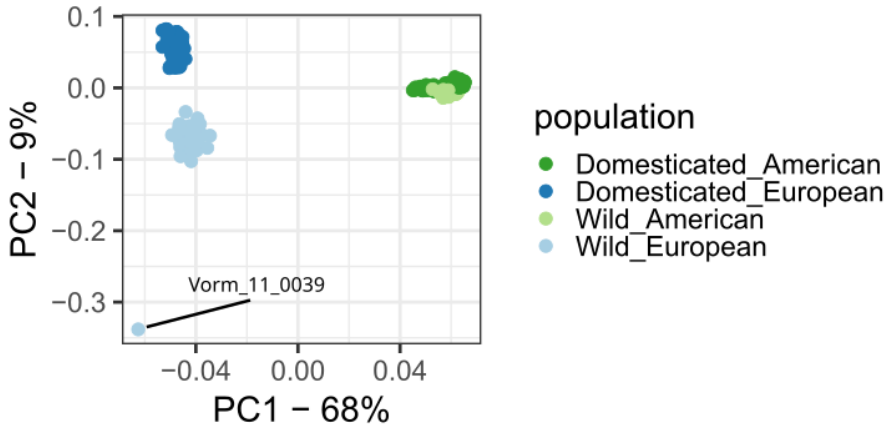
Supplementary Figure 1: BayesTyper_SVs detection pipeline

Top: Manta-based detection, bottom: BayesTyper genotyping. The final set BayesTyper_SVs used in this study is indicated in bold.



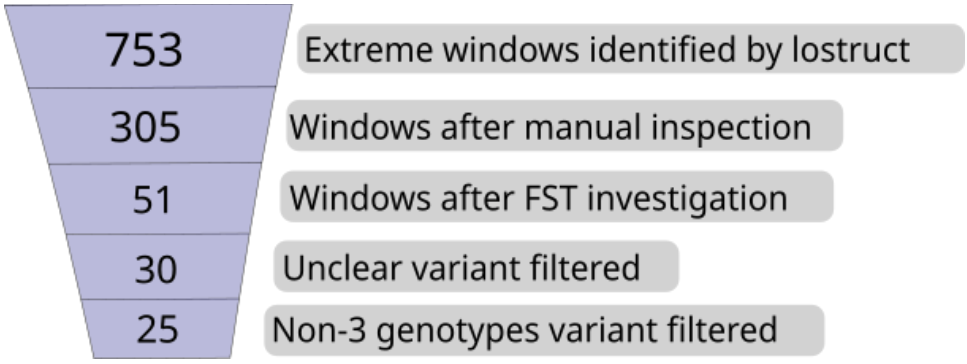
Supplementary Figure 2: Comparison of SVs between manta and bayestyper

A: Density of SV size (in log10). **B:** Number of SV of different types. SV types for bayestyper are not directly present but are rescued as described in Supplementary Figure 1 and Material and methods.

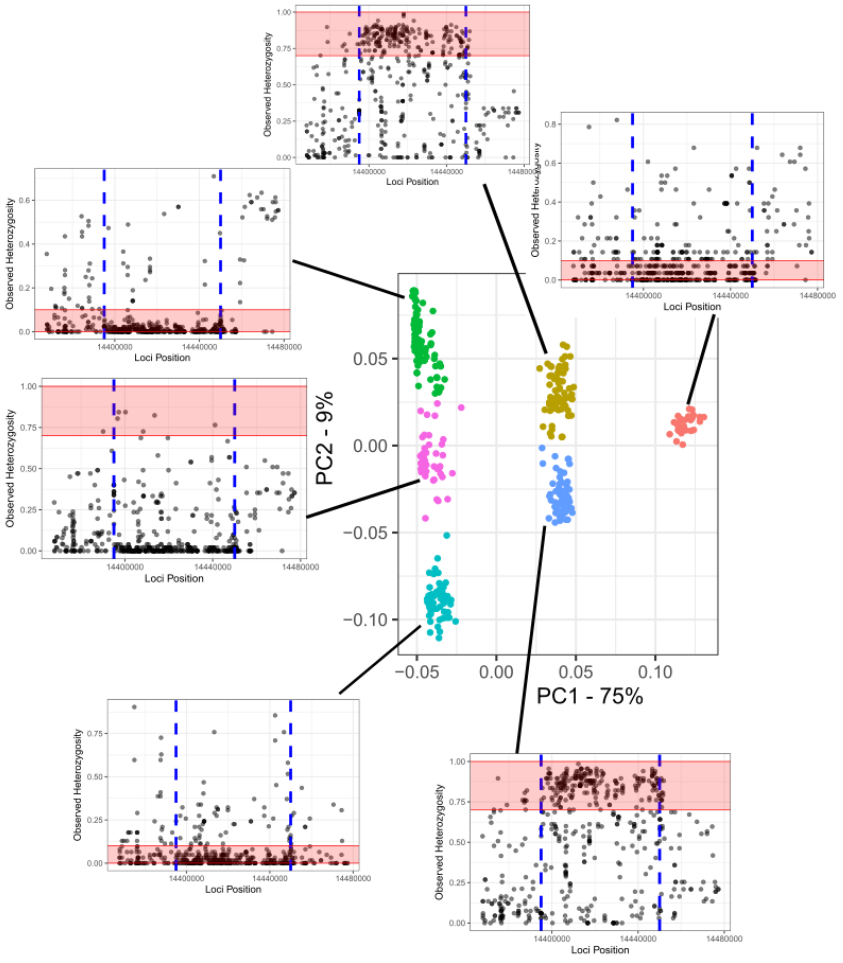


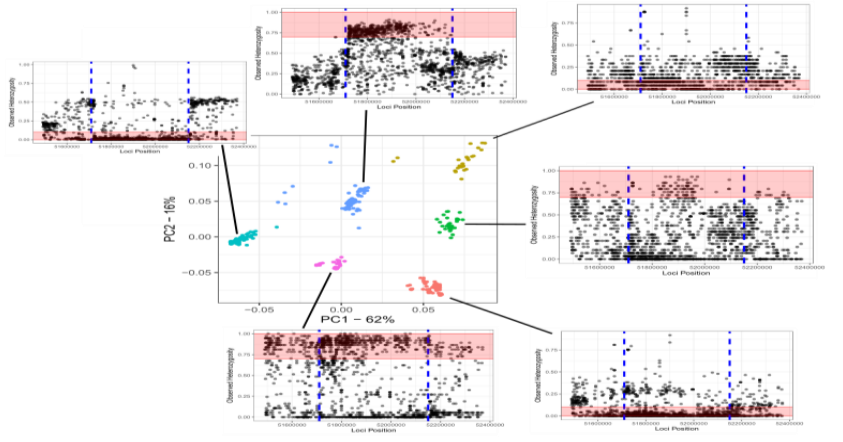
Supplementary Figure 3: Whole genome PCA using BayesTyper_SVs

PCA based on BayesTyper_SVs distinguish continents on PC1, and domestication status on PC2. The outlier individual (Vorm_11_0039 – Wild European) is indicated. While investigating this outlier by doing chromosome by chromosome PCA, we observed that the same individual was an outlier in 16 out of 29 chromosomes.



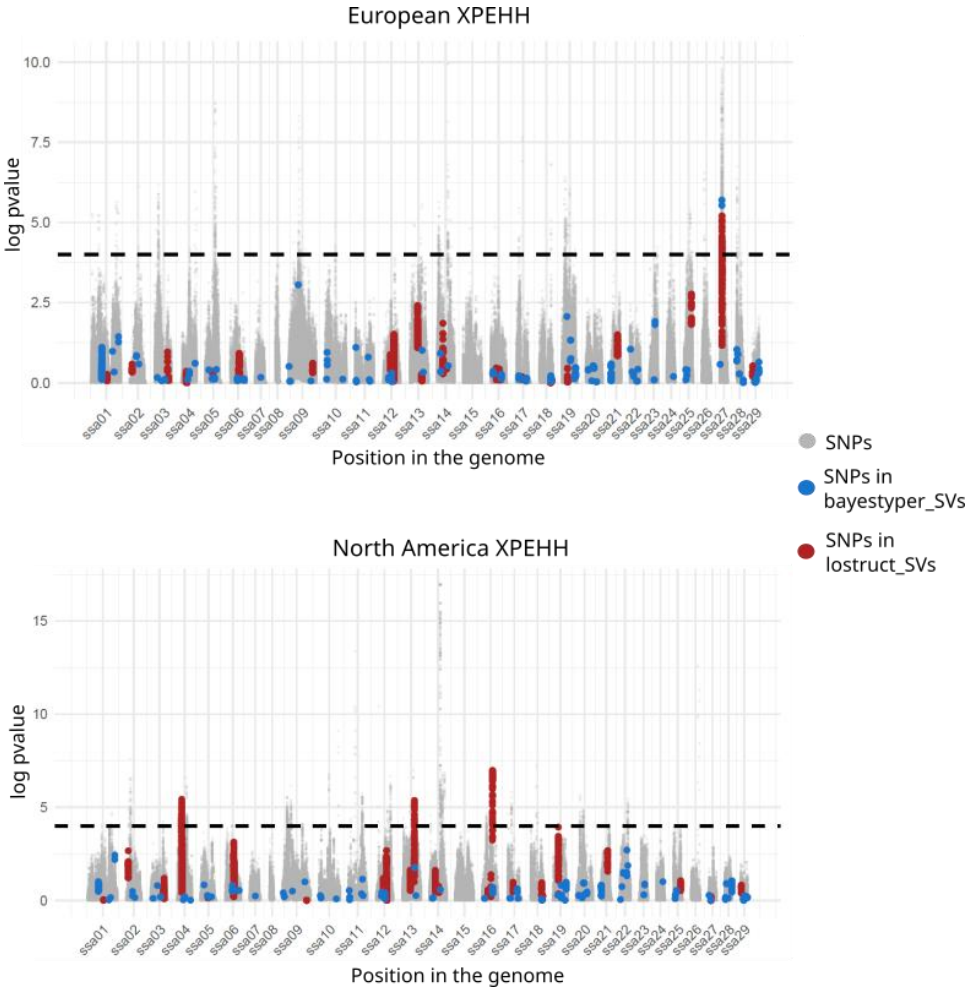
Supplementary Figure 4: lostruct_SVs detection method pipeline


Observed Heterozygosity
Loci Position
14400000
14440000
14480000
PC2 - 9%
PC1 - 75%
0.05
0.00
−0.05
−0.10
0.10


PC2 - 16%
PC1 – 62%
0.10
0.05
0.00
−0.05

Supplementary figure 5: Heterozygosity within PCA cluster

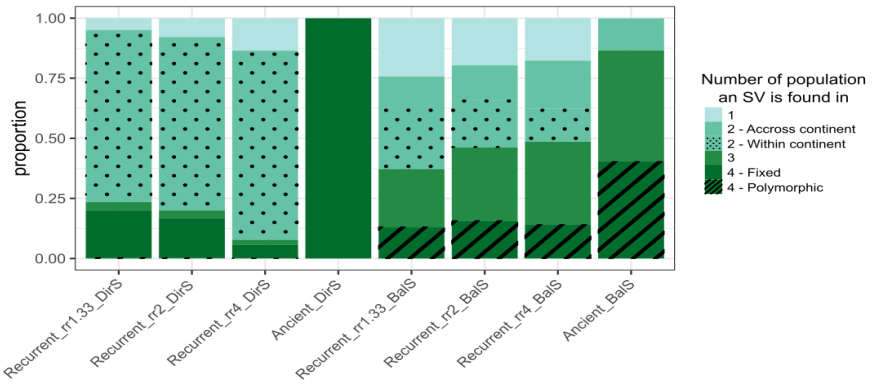
Top : ssa02: 14395000-14450000 ; bottom: ssa13: 51710000-52150000. Red-shaded area indicates heterozygosity expectations in the case of an SV with 3 alleles (**See Figure 4D).** Vertical dotted lines indicate the SV boundary.



Supplementary Figure 6: Genome-wide distribution of XP-EHH signals and overlap with structural variants

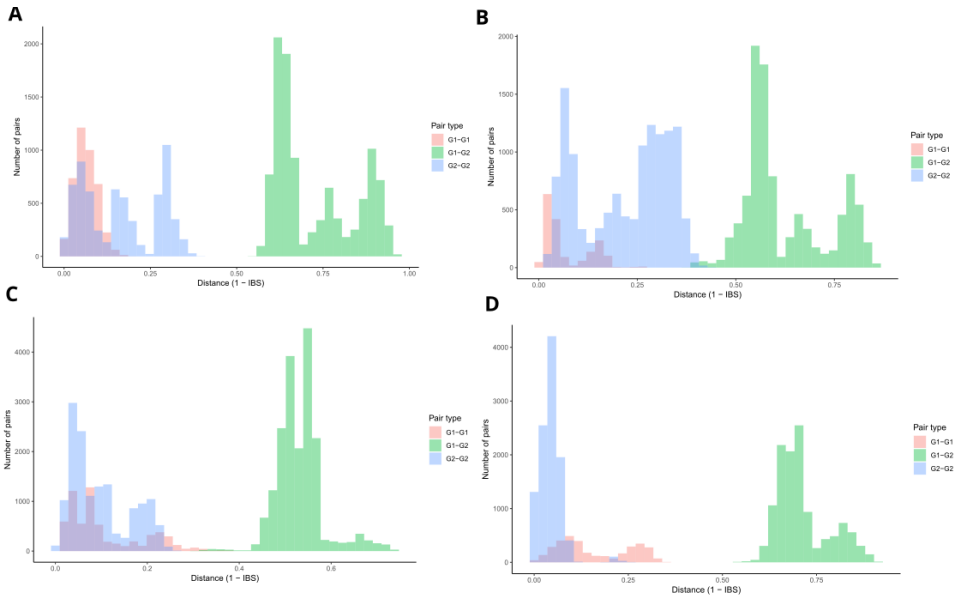
Genome-wide distribution of XP-EHH scores across chromosomes, based on SNP data from Buso et al., 2025. Grey points represent all SNPs included in the XP-EHH analysis. Colored points indicate SNPs overlapping with structural variants, with blue

representing SNPs overlapping BayesTyper_SVs and red representing SNPs overlapping lostructs_SVs. The horizontal dotted line indicates the XP-EHH threshold value, as defined in Buso et al., 2025, used to identify outlier signals.



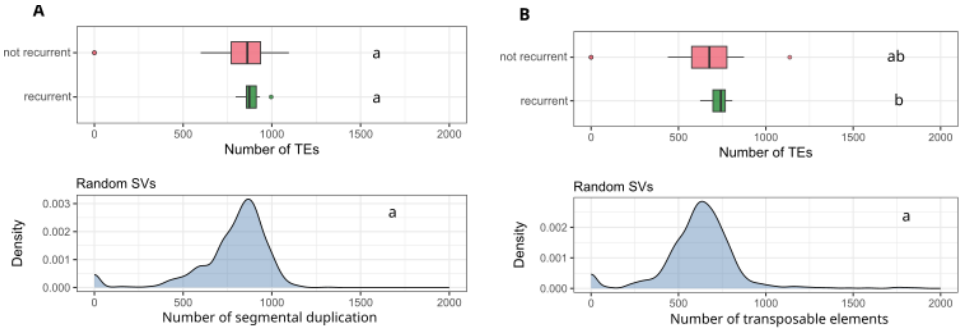
Supplementary figure 7: Distribution of SVs across populations under different simulation scenarios.

Stacked bar plots show the proportion of structural variants (SVs) observed in different populations, stratified by scenario. Labels indicate the evolutionary history of SVs (recurrent or ancient), the mode of selection (directional selection, DirS; balancing selection, BalS), and the recurrence rate (rr2, corresponding to an average of two SVs per simulation). All simulations were performed using a selection coefficient of 1.1.



Supplementary figure 8: Recurrent Lostruct_SVs.

Genetic distance between SNPs in haplotype, G1-G1: AA vs AA; G1-G2:AA vs BB; G2-G2: BB vs BB. Under recurrent SV, the distribution is expected to be multimodal for G1-G2. **A:** ssa05:44020000-44084228; **B:** ssa16:33750000-34020915; **C:** ssa16:60620000-60880000; **D:** ssa21:38333640-38459102



Supplementary Figure 9 Transposable elements enrichment near Lostructs_SVs

Number of repeats near SVs boundaries (+/- 100kb) for the 4 SVs assigned as recurrent (**Supplementary Figure 8**) or the other Lostruct_SVs. These numbers are

compared to the numbers of repeat for 1000 randomly picked location in the genome (+/- 100kb). Letters indicate the significantly different groups **A:** Segmental duplications (not recurrent-recurrent: p-value=0.48; not recurrent-random: p-value=0.13; recurrent-random: p-value=0.15) **B:** Transposable elements (not recurrent-recurrent: p-value=0.13; not recurrent-random: p-value=0.10; recurrent-random: p-value=0.04)

Supplementary table 1: Lostruct_SVs coordinates and overlapping genes

Coordinate	Genes
Ssa01: 92750000-92880000	
Ssa02: 14395000-14450000	ENSSSAG00000041032;ENSSSAG00000045818;ENSSSAG00000040824; ENSSSAG00000047726
Ssa03: 75839027-75979106	ENSSSAG00000001726;ENSSSAG00000001727;ENSSSAG00000099337; ENSSSAG00000095529;ENSSSAG00000083867;ENSSSAG00000090150; ENSSSAG00000004153;ENSSSAG00000057854;ENSSSAG00000009076; ENSSSAG00000004118
Ssa03: 81220000-81261255	ENSSSAG00000000943;ENSSSAG00000000925;ENSSSAG00000088970
Ssa04: 29000000-32100000	ENSSSAG00000005991;ENSSSAG00000041689;ENSSSAG00000109184; ENSSSAG00000113463;ENSSSAG00000045683;ENSSSAG00000086905; ENSSSAG00000049824;ENSSSAG00000049858;ENSSSAG00000089981; ENSSSAG00000111565;ENSSSAG00000059348;ENSSSAG00000041070; ENSSSAG00000005892;ENSSSAG00000120002;ENSSSAG00000005936; ENSSSAG00000005953;ENSSSAG00000041666;ENSSSAG00000041774; ENSSSAG00000103400;ENSSSAG00000043249;ENSSSAG00000111674; ENSSSAG00000043456;ENSSSAG00000045334;ENSSSAG00000045653; ENSSSAG00000010182;ENSSSAG00000005502;ENSSSAG00000045602; ENSSSAG00000084245;ENSSSAG00000047918;ENSSSAG00000107634; ENSSSAG00000049840;ENSSSAG00000085065;ENSSSAG00000036829; ENSSSAG00000040099;ENSSSAG00000041175;ENSSSAG00000041197; ENSSSAG00000005978;ENSSSAG00000006503;ENSSSAG00000043307; ENSSSAG00000063298;ENSSSAG00000096375;ENSSSAG00000045755; ENSSSAG00000049796;ENSSSAG00000088225;ENSSSAG00000102949; ENSSSAG00000049978;ENSSSAG00000054868;ENSSSAG00000054962; ENSSSAG00000057700;ENSSSAG00000041051;ENSSSAG00000041721; ENSSSAG00000049715;ENSSSAG00000049619
Ssa05:44020000-44084228	
Ssa06: 52540000-55000000	ENSSSAG00000100736;ENSSSAG00000032979;ENSSSAG00000107853; ENSSSAG00000003730;ENSSSAG00000003760;ENSSSAG00000007969; ENSSSAG00000104602;ENSSSAG00000094877;ENSSSAG00000067479; ENSSSAG00000120792;ENSSSAG00000120858;ENSSSAG00000007629; ENSSSAG00000114918;ENSSSAG00000007615;ENSSSAG00000002811; ENSSSAG00000115651;ENSSSAG00000003941;ENSSSAG00000100701; ENSSSAG00000008124;ENSSSAG00000015636;ENSSSAG00000087136; ENSSSAG00000067588;ENSSSAG00000072139;ENSSSAG00000121038; ENSSSAG00000007650;ENSSSAG00000007605;ENSSSAG00000103681; ENSSSAG00000022058;ENSSSAG00000072133;ENSSSAG00000088042; ENSSSAG00000094083
Ssa09: 140185732-140202636	ENSSSAG00000057167;ENSSSAG00000057206
Ssa12: 63150000-63800000	ENSSSAG00000069806;ENSSSAG00000077210;ENSSSAG00000119026; ENSSSAG00000069809;ENSSSAG00000094905;ENSSSAG00000077213; ENSSSAG00000079782;ENSSSAG00000097016;ENSSSAG00000070160; ENSSSAG00000098839
Ssa12: 64000000-65000000	ENSSSAG00000082416;ENSSSAG00000084198;ENSSSAG00000094754; ENSSSAG00000080980;ENSSSAG00000080100;ENSSSAG00000080864; ENSSSAG00000104465;ENSSSAG00000080857;ENSSSAG00000080876; ENSSSAG00000080924;ENSSSAG00000080995;ENSSSAG00000081022;

	ENSSSAG00000081065;ENSSSAG00000084657;ENSSSAG00000080247; ENSSSAG00000080851;ENSSSAG00000008488;ENSSSAG00000081034; ENSSSAG00000081196
Ssa13: 51710000-52150000	ENSSSAG00000111810;ENSSSAG00000070051;ENSSSAG00000070458; ENSSSAG00000073513
Ssa13: 76000000-76800000	ENSSSAG00000109016;ENSSSAG00000058260;ENSSSAG00000100170; ENSSSAG00000058310;ENSSSAG00000102970;ENSSSAG00000058323; ENSSSAG00000058336;ENSSSAG00000058211
Ssa14: 35000000-35400000	ENSSSAG00000022880;ENSSSAG00000090634;ENSSSAG00000117177; ENSSSAG00000032035;ENSSSAG00000048190;ENSSSAG00000101138; ENSSSAG00000048567;ENSSSAG00000028979;ENSSSAG00000048728
Ssa14: 56242199-56291923	ENSSSAG00000055927
Ssa16: 33750000-34020915	ENSSSAG00000050779;ENSSSAG00000069422;ENSSSAG00000108830; ENSSSAG00000069428
Ssa16: 51650000-51820000	ENSSSAG00000007201;ENSSSAG00000007205;ENSSSAG00000007185; ENSSSAG00000121435;ENSSSAG00000112886;ENSSSAG00000006625
Ssa16: 60620000-60880000	ENSSSAG00000086157;ENSSSAG00000039914
Ssa17: 35400419-35510000	ENSSSAG00000111008;ENSSSAG00000110934;ENSSSAG00000111335; ENSSSAG00000083564;ENSSSAG00000111627
Ssa18: 67200000-67510731	
Ssa18: 72307638-72471162	ENSSSAG00000066453;ENSSSAG00000066459;ENSSSAG00000101816; ENSSSAG00000040226;ENSSSAG00000102714;ENSSSAG00000040378; ENSSSAG00000040458;ENSSSAG00000040529;ENSSSAG00000092702; ENSSSAG00000040184;ENSSSAG00000040291;ENSSSAG00000040351; ENSSSAG00000062190;ENSSSAG00000112668;ENSSSAG00000062249; ENSSSAG00000103344;ENSSSAG00000062102;ENSSSAG00000117465
Ssa19: 31600000-32200000	ENSSSAG00000072920;ENSSSAG00000072912;ENSSSAG00000072943; ENSSSAG00000115609;ENSSSAG00000091716
Ssa21: 38333640-38459102	ENSSSAG00000059086;ENSSSAG00000059799
Ssa25: 43856943-43882324	ENSSSAG00000114775
Ssa27: 11940000-12240000	ENSSSAG00000068404;ENSSSAG00000082142;ENSSSAG00000040147; ENSSSAG00000040125;ENSSSAG00000039984;ENSSSAG00000039965; ENSSSAG00000084331;ENSSSAG00000120981;ENSSSAG00000107011; ENSSSAG00000115878;ENSSSAG00000068332
Ssa29: 5670000-5900000	

Supplementary table 2: Genes overlapping structural variants highlighted in Figure S6

List of annotated genes located within structural variant (SV) regions that overlap with XP-EHH outlier signals shown in **Supplementary figure 6**. Gene identifiers, gene names, and corresponding SV genomic coordinates are shown.

Gene	Gene name	SV coordinate
ENSSSAG00000038615	TatD DNase domain containing 1 (*tatdn1*)	ssa27: 11940000-12240000
ENSSSAG00000068332	GA binding protein transcription factor subunit beta 2a (*gabpb2a*)	ssa27: 11940000-12240000
ENSSSAG00000115878	Unknown	ssa27: 11940000-12240000
ENSSSAG00000107011	guanine nucleotide-binding protein G(s) subunit alpha-like	ssa27: 11940000-12240000
ENSSSAG00000120981	Unknown	ssa27: 11940000-12240000
ENSSSAG00000084331	MLLT11 transcription factor 7 cofactor	ssa27: 11940000-12240000

ENSSSAG00000039965	CDC42 small effector 1 (*CDC42SE1*)	ssa27: 11940000-12240000
ENSSSAG00000039984	BCL2/adenovirus E1B 19 kDa protein-interacting protein 2	ssa27: 11940000-12240000
ENSSSAG00000040125	Prune exopolyphosphatase 1 (*PRUNE1*)	ssa27: 11940000-12240000
ENSSSAG00000040147	MINDY lysine 48 deubiquitinase 1 (*MINDY1*)	ssa27: 11940000-12240000
ENSSSAG00000082142	DNA-binding protein RFX5-like	ssa27: 11940000-12240000
ENSSSAG00000068404	phosphatidylinositol 4-kinase beta	ssa27: 11940000-12240000
ENSSSAG00000005991	Helt bHLH transcription factor (*helt*)	ssa04:29000000_32100000
ENSSSAG00000005936	Cilia and flagella associated protein 97 (*cfap97*)	ssa04:29000000_32100000
ENSSSAG00000005953	Solute carrier family 25 member 4 (*slc25a4*)	ssa04:29000000_32100000
ENSSSAG00000005978	EF-hand calcium-binding domain-containing protein 6	ssa04:29000000_32100000
ENSSSAG00000006503	Acyl-CoA synthetase long chain family member 1a (*acsl1a*)	ssa04:29000000_32100000
ENSSSAG00000058310	CCR4-NOT transcription complex subunit 8 (*cnot8*)	ssa13:76000000_76800000
ENSSSAG00000102970	Gem (nuclear organelle) associated protein 5 (*gemin5*)	ssa13:76000000_76800000
ENSSSAG00000058323	Family with sequence similarity 114 member A2 (*FAM114A2*)	ssa13:76000000_76800000
ENSSSAG00000039914	Cholecystokinin B receptor a (*cckbra*)	ssa16: 60620000_60880000

Bibliography:

Almarri MA, Bergström A, Prado-Martinez J, Yang F, Fu B, Dunham AS, et al. (2020). Population Structure, Stratification, and Introgression of Human Structural Variation. Cell 182: 189-199.e15.

Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, et al. (2020). Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. Cell 182: 145-161.e23.

Aqil A, Speidel L, Pavlidis P, Gokcumen O (2023). Balancing selection on genomic deletion polymorphisms in humans. eLife 12: e79111.

Berdan EL, Aubier TG, Cozzolino S, Faria R, Feder JL, Giménez MD, et al. (2024). Structural Variants and Speciation: Multiple Processes at Play. Cold Spring Harb Perspect Biol 16: a041446.

Berdan EL, Blanckaert A, Butlin RK, Bank C (2021). Deleterious mutation accumulation and the long-term fate of chromosomal inversions (A Buerkle, Ed.). PLoS Genet 17: e1009411.

Bertolotti AC, Layer RM, Gundappa MK, Gallagher MD, Pehlivanoglu E, Nome T, et al. (2020). The structural variation landscape in 492 Atlantic salmon genomes. Nat Commun 11: 5176.

Boettger LM, Salem RM, Handsaker RE, Peloso GM, Kathiresan S, Hirschhorn JN, et al. (2016). Recurring exon deletions in the HP (haptoglobin) gene contribute to lower blood cholesterol levels. Nat Genet 48: 359–366.

Bradbury IR, Lehnert SJ, Kess T, Van Wyngaarden M, Duffy S, Messmer AM, et al. (2022). Genomic evidence of recent European introgression into North American farmed and wild Atlantic salmon. Evolutionary Applications 15: 1436–1448.

Burssed B, Zamariolli M, Bellucco FT, Melaragno MI (2022). Mechanisms of structural chromosomal rearrangement formation. Mol Cytogenet 15: 23.

Buso P, Diblasi C, Manousi D, Kwak JS, Vera-Ponce De Leon A, Stenløkk K, et al. (2025). Parallel Selection in Domesticated Atlantic Salmon from Divergent Founders Including on Whole-Genome Duplication-derived Homeologous Regions (B Fraser, Ed.). Genome Biology and Evolution 17: evaf063.

Carvalho CMB, Lupski JR (2016). Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet 17: 224–238.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaSci 4: 7.

Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics 32: 1220–1222.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. (2011). The variant call format and VCFtools. Bioinformatics 27: 2156–2158.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. (2021). Twelve years of SAMtools and BCFtools. GigaScience 10: giab008.

Emanuel BS, Shaikh TH (2001). Segmental duplications: an 'expanding' role in genomic instability and disease. Nat Rev Genet 2: 791–800.

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: 305–314.

Falker-Gieske C, Bennewitz J, Tetens J (2022). Structural variation and eQTL analysis in two experimental populations of chickens divergently selected for feather-pecking behavior. Neurogenetics 24: 29–41.

Faria R, Chaube P, Morales HE, Larsson T, Lemmon AR, Lemmon EM, et al. (2019). Multiple chromosomal rearrangements in a hybrid zone between Littorina saxatilis ecotypes. Molecular Ecology 28: 1375–1393.

Ferguson S, Jones A, Murray K, Andrew RL, Schwessinger B, Bothwell H, et al. (2024). Exploring the role of polymorphic interspecies structural variants in reproductive isolation and adaptive divergence in Eucalyptus. GigaScience 13: giae029.

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci USA 117: 9451–9457.

Glover KA, Otterå H, Olsen RE, Slinde E, Taranger GL, Skaala Ø (2009). A comparison of farmed, wild and hybrid Atlantic salmon (Salmo salar L.) reared under farming conditions. Aquaculture 286: 203–210.

Göktay M, Fulgione A, Hancock AM (2021). A New Catalog of Structural Variants in 1,301 A. thaliana Lines from Africa, Eurasia, and North America Reveals a Signature of Balancing Selection at Defense Response Genes (A Larracuente, Ed.). Molecular Biology and Evolution 38: 1498–1511.

Gong T, Hayes VM, Chan EKF (2021). Detection of somatic structural variants from short-read next-generation sequencing data. Briefings in Bioinformatics 22: bbaa056.

González JR, Cáceres A, Esko T, Cuscó I, Puig M, Esnaola M, et al. (2014). A Common 16p11.2 Inversion Underlies the Joint Susceptibility to Asthma and Obesity. The American Journal of Human Genetics 94: 361–372.

Goubert C, Craig RJ, Bilat AF, Peona V, Vogan AA, Protasio AV (2022). A beginner's guide to manual curation of transposable elements. Mobile DNA 13: 7.

Gray YHM (2000). It takes two transposons to tango:transposable-element-mediated chromosomal rearrangements. Trends in Genetics 16: 461–468.

GTEx Consortium, Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, et al. (2017). The impact of structural variation on human gene expression. Nat Genet 49: 692–699.

Haller BC, Messer PW (2023). SLiM 4: Multispecies Eco-Evolutionary Modeling. The American Naturalist 201: E127–E139.

Hämälä T, Moore C, Cowan L, Carlile M, Gopaulchan D, Brandrud MK, et al. (2024). Impact of whole-genome duplications on structural variant evolution in Cochlearia. Nat Commun 15: 5377.

Hämälä T, Wafula EK, Guiltinan MJ, Ralph PE, dePamphilis CW, Tiffin P (2021). Genomic structural variants constrain and facilitate adaptation in natural populations of Theobroma cacao , the chocolate tree. Proc Natl Acad Sci USA 118: e2102914118.

Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, et al. (2017). Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nat Commun 8: 14061.

Jombart T (2008). adegenet : a R package for the multivariate analysis of genetic markers. Bioinformatics 24: 1403–1405.

Karageorgiou C, Gokcumen O, Dennis MY (2024). Deciphering the role of structural variation in human evolution: a functional perspective. Current Opinion in Genetics & Development 88: 102240.

Kidd JM, Graves T, Newman TL, Fulton R, Hayden HS, Malig M, et al. (2010). A Human Genome Structural Variation Sequencing Resource Reveals Insights into Mutational Mechanisms. Cell 143: 837–847.

King TL, Verspoor E, Spidle AP, Gross R, Phillips RB, Koljonen M, et al. (2007). Biodiversity and Population Structure. In: Verspoor E, Stradmeyer L, Nielsen J (eds) The Atlantic Salmon, Wiley, pp 117–166.

Knaus BJ, Grünwald NJ (2017). VCFR : a package to manipulate and visualize variant call format data in R. Molecular Ecology Resources 17: 44–53.

Knief U, Hemmrich-Stanisak G, Wittig M, Franke A, Griffith SC, Kempenaers B, et al. (2016). Fitness consequences of polymorphic inversions in the zebra finch genome. Genome Biol 17: 199.

Korf I, Yandell M, Bedell J (2003). BLAST: an essential guide to the Basic Local Alignment Search Tool, 1. ed. O'Reilly: Beijing Köln.

Korunes KL, Samuk K (2021). PIXY : Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. Molecular Ecology Resources 21: 1359–1368.

Kosugi S (2019). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. : 18.

Krkosek M (2010). Host density thresholds and disease control for fisheries and aquaculture. Aquacult Environ Interact 1: 21–32.

Lehnert SJ, Kess T, Bentzen P, Barson N, Lien S, Dempson JB, et al. (2025). Large haplotypes linked to climate and life history variation in divergent lineages of Atlantic salmon (Salmo salar).

Lehnert SJ, Kess T, Bentzen P, Clément M, Bradbury IR (2020). Divergent and linked selection shape patterns of genomic differentiation between European and North American Atlantic salmon (Salmo salar). Mol Ecol 29: 2160–2175.

Li H, Durbin R (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. : 7.

Li BJ, Li HL, Meng Z, Zhang Y, Lin H, Yue GH, et al. (2017). Copy Number Variations in Tilapia Genomes. Mar Biotechnol 19: 11–21.

Li H, Ralph P (2019). Local PCA Shows How the Effect of Population Structure Differs Along the Genome. Genetics 211: 289–304.

Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, et al. (2016). The Atlantic salmon genome provides insights into rediploidization. Nature 533: 200–205.

Lin Y-L, Gokcumen O (2019). Fine-Scale Characterization of Genomic Structural Variation in the Human Genome Reveals Adaptive and Biomedically Relevant Hotspots (B Chang, Ed.). Genome Biology and Evolution 11: 1136–1151.

Lindtke D, Lucek K, Soria-Carrasco V, Villoutreix R, Farkas TE, Riesch R, et al. (2017). Long-term balancing selection on chromosomal variants associated with crypsis in a stick insect. Molecular Ecology 26: 6189–6205.

Lubieniecki KP, Jones SL, Davidson EA, Park J, Koop BF, Walker S, et al. (2010). Comparative genomic analysis of Atlantic salmon, Salmo salar, from Europe and North America. BMC Genet 11: 105.

Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ (2019). Structural variant calling: the long and the short of it. Genome Biol 20: 246.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297–1303.

Mérot C, Berdan EL, Cayuela H, Djambazian H, Ferchaud A-L, Laporte M, et al. (2021). Locally Adaptive Inversions Modulate Genetic Variation at Different Geographic Scales in a Seaweed Fly (D Charlesworth, Ed.). Molecular Biology and Evolution 38: 3953–3971.

Mérot C, Oomen RA, Tigano A, Wellenreuther M (2020). A Roadmap for Understanding the Evolutionary Significance of Structural Genomic Variation. Trends in Ecology & Evolution 35: 561–572.

NHGRI Centers for Common Disease Genomics, Abel HJ, Larson DE, Regier AA, Chiang C, Das I, et al. (2020). Mapping and characterization of structural variation in 17,795 human genomes. Nature 583: 83–89.

Payne RH, Child AR, Forrest A (1971). Geographical Variation in the Atlantic Salmon. Nature 231: 250–252.

Penso-Dolfin L, Man A, Mehta T, Haerty W, Di Palma F (2020). Analysis of structural variants in four African cichlids highlights an association with developmental and immune related genes. BMC Evol Biol 20: 69.

Porubsky D, Sanders AD, Höps W, Hsieh P, Sulovari A, Li R, et al. (2020). Recurrent inversion toggling and great ape genome evolution. Nat Genet 52: 849–858.

Quan C, Li Y, Liu X, Wang Y, Ping J, Lu Y, et al. (2021). Characterization of structural variation in Tibetans reveals new evidence of high-altitude adaptation and introgression. Genome Biol 22: 159.

Quinlan AR, Hall IM (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842.

Rodriguez OL, Safonova Y, Silver CA, Shields K, Gibson WS, Kos JT, et al. (2023). Genetic variation in the immunoglobulin heavy chain locus shapes the human antibody repertoire. Nat Commun 14: 4419.

Saitou M, Resendez S, Pradhan AJ, Wu F, Lie NC, Hall NJ, et al. (2021). Sex-specific phenotypic effects and evolutionary history of an ancient polymorphic deletion of the human growth hormone receptor. Sci Adv 7: eabi4476.

Schaal SM, Haller BC, Lotterhos KE (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 377: 20210200.

Scott AJ, Chiang C, Hall IM (2021). Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. Genome Res 31: 2249–2257.

Stenløkk K, Saitou M, Rud-Johansen L, Nome T, Moser M, Árnyasi M, et al. (2022). The emergence of supergenes from inversions in Atlantic salmon. Phil Trans R Soc B 377: 20210195.

Stuart K, Oomen R, Tigano A, Wellenreuther M, Wold J, Field DL, et al. (2025). A Beginner's Guide to Structural Variants in Eco-Evolutionary Population Genomics: Everything You Wanted to Know.

The Danish Pan-Genome Consortium, Sibbesen JA, Maretty L, Krogh A (2018). Accurate genotyping across variant classes and lengths using variant graphs. Nat Genet 50: 1054–1059.

Traherne JA, Martin M, Ward R, Ohashi M, Pellett F, Gladman D, et al. (2010). Mechanisms of copy number variation and hybrid gene formation in the KIR immune gene complex. Human Molecular Genetics 19: 737–751.

Wellenreuther M, Mérot C, Berdan E, Bernatchez L (2019). Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification. Mol Ecol 28: 1203–1209.

Xu D, Pavlidis P, Taskent RO, Alachiotis N, Flanagan C, DeGiorgio M, et al. (2017). Archaic Hominin Introgression in Africa Contributes to Functional Salivary MUC7 Genetic Variation. Molecular Biology and Evolution 34: 2704–2715.

Yan H, Bombarely A, Li S (2020). DeepTE: a computational method for de novo classification of transposons with convolutional neural network (A Valencia, Ed.). Bioinformatics 36: 4269–4275.

Zhang Y, Yang Z, He Y, Liu D, Liu Y, Liang C, et al. (2024). Structural variation reshapes population gene expression and trait variation in 2,105 Brassica napus accessions. Nat Genet 56: 2538–2550.

Zhou Y, Minio A, Massonnet M, Solares E, Lv Y, Beridze T, et al. (2019). The population genetics of structural variants in grapevine domestication. Nat Plants 5: 965–979.

ISBN: 978-82-575-2335-0
ISSN: 1894-6402

Norwegian University
of Life Sciences

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

P1017038 / MEDIEHUSET-ANDVORD.NO