Norwegian University
of Life Sciences

Master's Thesis 2022 **60 ECTS**

Faculty of Biosciences

Evolution of Alternative Splice Variation and Exon Usage Following Whole Genome Duplication

Ronja Marlonsdotter Sandholm

M.Sc. Biotechnology (Molecular Biology)

Acknowledgments

This thesis marks the end of my master's degree in biotechnology from the Norwegian University of Life Sciences (NMBU). It is part of the REWIRED project at the Center for Integrative Genetics (CIGENE), which is funded by FRIMEDIBO, The Research Council of Norway and is led by Professor Simen Rød Sandve. The project's goal is to gain a better understanding of the role of whole genome duplication and how it shapes the evolution of vertebrates.

I want to thank my main supervisor Professor Simen Rød Sandve, and my co-supervisors Associate professor Matthew Peter Kent and Ph.D. candidate Marius Andre Strand for their guidance and sound advice in connection with my thesis, as well as a special thanks to Chief engineer Mariann Arnyasi for all her assistance with my lab work. Lastly, I want to thank my fiancée Christoffer Aass for his encouragement and support during this challenging yet rewarding process.

Ås, May 2022

Ronja Marlonsdotter Sandholm

Abstract

Whole genome duplication and alternative splicing are two mechanisms that contribute to protein diversity. Whole genome duplication doubles the genetic material in an organism, providing raw material for adaptation and evolution of novel traits. Alternative splicing contributes to the proteome complexity, as it gives a gene the ability to produce several mRNA isoforms by alternatively splicing gene transcripts. While both processes are important factors in increasing protein diversity, their relationship is not well understood. The two primary aims of this thesis were to use Oxford Nanopore long-read RNA sequencing to better characterize the isoform diversity in Atlantic salmon and look for patterns of alternative splicing evolution following the salmonid-specific whole genome duplication event.

With the long-read RNA sequences, we found that the majority (75%) of isoforms that mapped to known genes in the Atlantic salmon reference genome were previously unannotated; however, the annotated isoforms were more highly expressed. The diversity of isoforms was then used to test the models of alternative splicing evolution following whole genome duplication: the independent model, the function-sharing model, and the accelerated alternative splicing model. Our results did not support either the accelerated alternative splicing or function-sharing model, indicating no strong relationship between alternative splicing evolution and genes duplicated in the salmonid-specific whole genome duplication event.

Sammendrag

Helgenomeduplikasjon og alternativ spleising er to mekanismer som øker proteindiversitet. WGD fordobler en organismes genetiske materiale, noe som gir råmaterial for evolusjon av nye egenskaper, og for tilpasningsdyktighet. Alternativ spleising bidrar til å øke proteindiversiteten ettersom et gen kan produsere flere mRNA-isoformer ved å alternativt spleise transkripter. Selv om begge prosessene er viktige bidrag til økt proteindiversitet, er forholdet mellom dem ikke godt forstått. De to hovedmålene med denne masteroppgaven var å bruke Oxford Nanopore long-read RNA-sekvensering for å bedre karakterisere isoformdiversitet i atlanterhavslaks, og å se etter mønstre i evolusjonen av alternativ spleising som følge av den salmonid-spesifikke helgenomduplikasjonen.

Ved å bruke long-read RNA-sekvenser fant vi ut at flertallet (75%) av isoformene med opphav fra et kjent gen i atlanterhavslaksens referansegenom var tidligere ikke annotert. De kjente isoformene, derimot, var høyere uttrykt. Isoformmangfoldet ble deretter brukt for å teste modellene for evolusjon av alternativ spleising: den uavhengige modellen, deling-av-funksjon-modellen og akselerert alternativ spleising-modellen. Våre resultater støttet hverken akselerert alternativ spleising- eller deling-av-funksjon-modellen, noe som indikerer at det ikke er et sterkt forhold mellom evolusjon av alternativ spleising og gener duplisert i den salmonid-spesifikke helgenomduplikasjonen.

Table of Contents

1 Introduction

When the human genome sequencing project started in 1990, the goal was to uncover the sequences of the 100 000 genes within our genome (NHGRI, 1990). They based this gene number on the estimated number of proteins, and the hypothesis first proposed by Beadle and Tatum in 1941: one gene = one polypeptide. Upon finishing the project in 2003, only 20 000 – 25 000 genes were found (NHGRI, 2004). We now know that many layers of gene regulation and post-transcriptional mechanisms contribute to proteome complexity in eukaryotes. A single protein can be modified by posttranslational incorporations of chemical groups such as methyl-, phosphate-, or acetyl groups, as well as small regulatory proteins called ubiquitin. Furthermore, proteins can be combined and assembled into complexes with other proteins. Finally, each gene can encode several proteins through a process called alternative splicing (AS), whereby different parts of the coding sequence of mRNA molecules can be joined into unique protein variants (Harper & Bennett, 2016).

1.1 Alternative splicing

The discovery of RNA splicing came from working on adenovirus 2 in 1977 (Berget et al., 1977). Based on RNA-DNA hybridization, Berget and colleagues (1977) determined that the 5' terminus of the mRNA coding for a virus capsid protein was not complementary to the gene but to three different segments of the virus DNA, indicating that this mRNA had three different splice variants. Now, more than 40 years later, modern Nanopore long-read sequencing has established that the adenovirus genome, which is ~36 000 bp long and around the size of a mammalian gene, produces over 900 alternative splice variants (Westergren Jakobsson et al., 2021). When it was discovered, AS was not thought to be common, with around 5% of genes being alternatively spliced (Stamm et al., 2005). With the help of high-throughput mRNA sequencing, this estimate is now thought to be as high as ~95% of all multi-exon human genes (Pan et al., 2008).

Most protein-coding eukaryotic genes contain coding regions called exons, interspaced with non-coding regions called introns. Before mRNAs are translated into proteins, the introns are spliced out by the spliceosome. The spliceosome is a protein complex that consists of several small nuclear ribonucleoproteins (snRNPs) that assemble on the newly transcribed pre-mRNA. The components recognize splice sites and perform the catalytic reactions, which cleave the pre-mRNA to excise the introns and join the exons. The spliceosome recognizes exon-intron boundaries through specific sequences at the 5' and 3' splice sites, and splicing is

aided by a third sequence within the intron – the branchpoint site. The spliceosome first cuts the mRNA at the 5' splice site and joins the free end of the intron to the branchpoint site, creating a loop. The 3' end of the free exon is brought close to the 3' splice site and joined with the downstream exon, excising the intron. The canonical sequences of the 5' and 3' splice sites are highly conserved dinucleotides at the start and end of the intron, which are GT and AG, respectively (Burset, 2000). Other splice site sequences are not as frequently used but are thought to be more common in alternatively spliced transcripts, with 76% of non-canonical and 40% of canonical splice sites being used in AS events in humans (Parada et al., 2014). During the splicing process, exons can be excluded, introns retained, or alternative 5' or 3' splice sites can be used to create alternative splice isoforms (Stamm et al., 2005).

AS is regulated as a response to changing environmental conditions, is induced by stress, and is important in cell specialization and development (Baralle & Giudice, 2017; Ule & Blencowe, 2019). Alternatively spliced mRNAs, referred to as isoforms, increase the number of proteins produced by a single gene. The translated isoforms can have subtle differences in protein function or have a new function altogether (Stamm et al., 2005).

1.1.1 Types of alternative splicing

Alternative splicing events can be categorized into five main types (Figure 1.1): exon skipping, alternative 5' splice sites, alternative 3' splice sites, intron retention, and mutually exclusive exons.

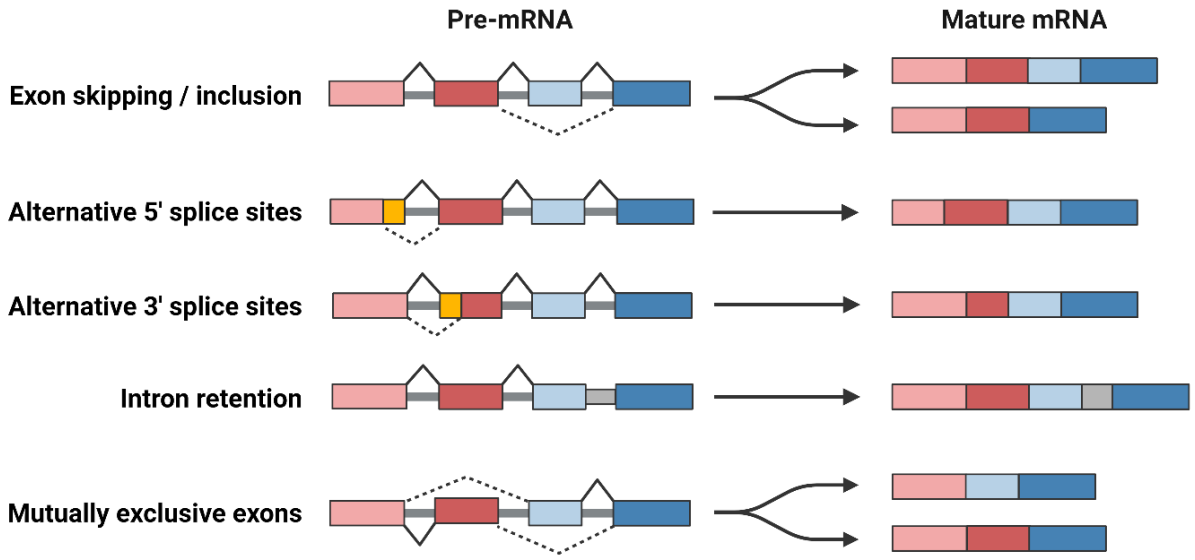

Figure 1.1: Main types of alternative splicing patterns. *Figure created with BioRender (https://biorender.com/).*

The most common type of AS is exon skipping, or inclusion, where a single or multiple exons are left out from the mature transcript. At alternative 5' and 3' splice sites, the pre-mRNA is spliced at alternative splice sites, located within the exon. Alternative 5' splice sites change the 3' boundary of the downstream exon and alternative 3' splice sites change the 5' boundary of the downstream exon. In intron retention, part of an intron is retained in the mature transcript. For mutually exclusive exons, the inclusion of one exon excludes the other (Stamm et al., 2005). An alternatively spliced mRNA can include more complicated splice patterns, where more than one type of AS event is included.

1.1.2 Nonsense-mediated decay

While AS is a mechanism to increase protein diversity, is can also play a role in gene regulation. Not all mRNA transcripts are translated into proteins, and AS can produce transcripts that contain premature termination codons (PTCs), which are thought to play a role in gene regulation through nonsense-mediated decay. Nonsense-mediated decay is the mechanism that recognizes and degrades these transcripts, as they can result in truncated proteins with potentially deleterious functions if translated (Shi et al., 2015). Nonsense-mediated decay is a surveillance mechanism to prevent aberrantly spliced mRNAs from being translated. On the other hand, PTC-containing mRNAs can act as a regulatory mechanism for gene expression. The exact nature of this gene regulation is still unclear, but it is suggested that nonsense-mediated decay maintains the homeostasis of and cross-regulation between RNA-binding proteins through auto-regulation feedback loops (Hamid & Makeyev, 2014; Watabe et al., 2021).

1.2 Whole genome duplication in vertebrates

Mutations are the source of variation that the forces of evolution act upon. They range from point mutations, where one base is altered, to structural variations that affect whole chromosomes. The most extreme form of structural variation comes in the form of whole genome duplication (WGD). WGD can result from autopolyploidy or allopolyploidy, where autopolyploidy is caused by unreduced gametes or chromosomal doubling of somatic cells in the early stages of embryonic development. On the other hand, allopolyploidy is the result of combining the hybridization of two species and chromosomal doubling (Spoelhof et al., 2017). The doubling of the genetic material provides raw material for developing new traits, thereby increasing genetic variation, adaptation to changing environments, and speciation through reciprocal gene losses (Sémon & Wolfe, 2007a, 2007b).

The 2R hypothesis, stating that two ancient WGD events occurred early in vertebrate evolution, was subject to much debate when first proposed by Susumu Ohno in 1970. The issue was still unresolved in the 90s, with opponents arguing that the observed gene duplicates resulted from multiple single-gene or segmental duplications rather than WGDs (Makałowski, 2001). However, with more genomes sequenced and better bioinformatic tools at our disposal, 2R has become broadly accepted in the following decades (Dehal & Boore, 2005; Moriyama & Koshiba-Takeuchi, 2018).

In addition to the 2R events, which occurred ~550-450 Mya (Dehal & Boore, 2005), several other known WGDs have taken place in the vertebrate lineages: a third WGD event in the lineage of teleost fishes (Ts3R) ~350-320 Mya (Meyer & van de Peer, 2005) which was followed by a fourth round in the lineage of salmonids (Ss4R) ~103-88 Mya (Macqueen & Johnston, 2014). Another known round of WGD is the allopolyploidization event in the frog *Xenopus laevis*, which happened ~54-21 Mya (Sémon & Wolfe, 2008).

1.2.1 The fate of duplicated genes after whole genome duplication

After a WGD event, the genome starts its slow journey to reversion, from tetraploid and back to diploid. The rediploidization process takes millions of years and is characterized by gene loss through fractionation or pseudogenization and chromosomal rearrangements (Berthelot et al., 2014; Schubert & Lysak, 2011). However, some gene duplicates (called ohnologs) are retained over hundreds of millions of years after WGD. How duplicated genes escape the fate of nonfunctionalization has been debated since Ohno published the book *Evolution by Gene Duplication* in 1970, and there are three main models to explain the fate of the gene duplicates: subfunctionalization, neofunctionalization, and nonfunctionalization (Figure 1.2).

Gene loss, or nonfunctionalization, is the most likely scenario in the short term, but the proportion of retained ohnologs varies between species (McGrath et al., 2014). Nonfunctionalization occurs when the lack of selective pressure to preserve genetic function in one of the ohnologs leads to an accumulation of mutations that render the gene non-functional (Dehal & Boore, 2005).

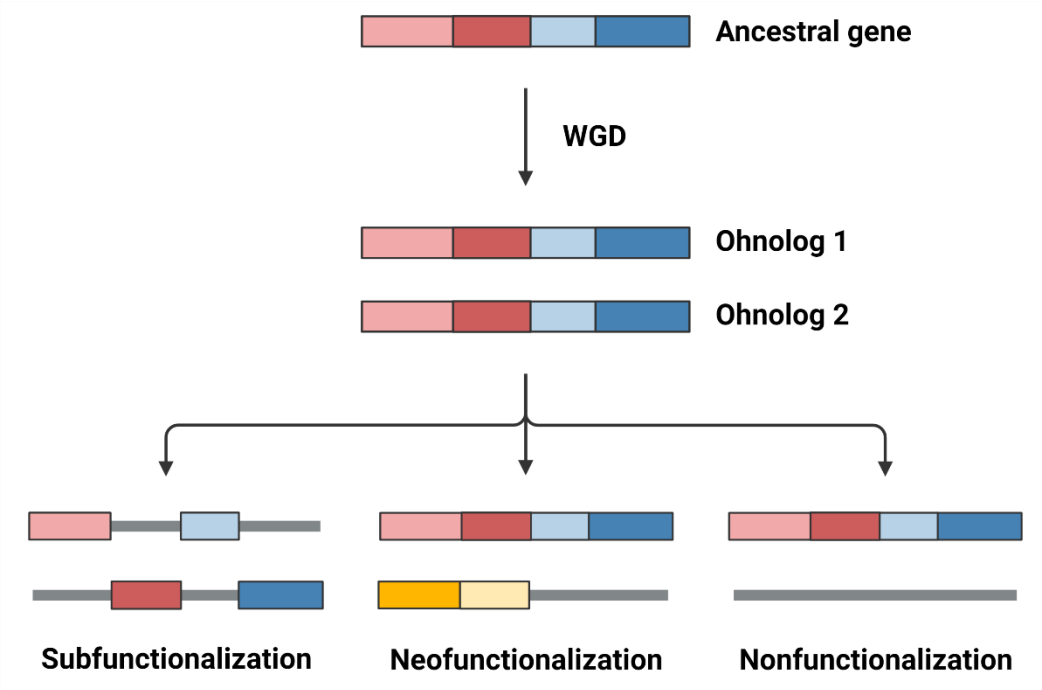

Figure 1.2: Fate of genes following whole genome duplication. *Under the subfunctionalization model, the ancestral functions are subdivided into the ohnologs. The neofunctionalization model states that one ohnolog retains the ancestral function while the other acquires a new function. Under the nonfunctionalization model, one copy is retained, and the other is lost. Figure created with BioRender (https://biorender.com/).*

The subfunctionalization model states that the ancestral function of the gene is partitioned in the ohnologs. Subfunctionalization covers the duplication, degeneration, complementation (DDC) model (Force et al., 1999) and the escape from adaptive conflict (EAC) model (Hittinger & Carroll, 2007). Under the DDC model, mutations that lead to the loss of a subset of the functions of one gene are not deleterious because the other copy still retains that function. Both copies are retained because one copy compensates for the loss of function in the other. Under the EAC model, each copy specializes in different fancestral gene functions through adaptive mutations and happens when the optimization of either ancestral function would be at the expense of the other (Conant & Wolfe, 2008).

In the neofunctionalization model, one gene retains the ancestral function while the other acquires a new function through mutations. The retention of both duplicates can result from positive selection of the neofunctionalized copy or simply genetic drift (Conant & Wolfe, 2008). Beneficial mutations that alter the protein function are relatively rare, and it is hypothesized that most cases of neofunctionalization are due to changes in gene regulation rather than alterations in the protein-coding sequence (Kassahn et al., 2009).

1.2.2 Whole genome duplication in Atlantic salmon

The Atlantic salmon, a species in the salmonid family, has been an integral part of our lives throughout human history. We have consumed these fishes from ancient times, and they are still a staple in our diet. The salmonids underwent a fourth round of WGD ~103-88 Mya (Macqueen & Johnston, 2014), with the Atlantic salmon being one of the extant species that has evolved from the common ancestor of this event. In addition to being an important aquaculture species, the Atlantic salmon genome has shed light on the evolution of vertebrates after WGD (Lien et al., 2016).

As the Ss4R event happened more recently compared to the older WGDs in the vertebrate lineage, and enough time has passed to study the long-term rediploidization process, the Atlantic salmon genome has shed light on the mechanisms of ohnolog retention (Lien et al., 2016), as well as evolution of gene regulation (Gillard et al., 2021). While diploidy is not fully reestablished, the Atlantic salmon genome has retained ~55% of the ohnolog pairs from Ss4R as functional copies, with >60% of pairs showing signs of divergent tissue regulation (Lien et al., 2016). The predominant fate of retained copies post-Ss4R is regulatory neofunctionalization (Sandve et al., 2018). Ohnolog pairs follow different patterns of regulatory evolution, mostly through one copy with conserved expression and another with downregulated expression, or through symmetrical downregulation of both copies. Ohnolog pairs where one copy has evolved lower expression through relaxed purifying selection most likely follow the path to pseudogenization (Gillard et al., 2021; Lien et al., 2016). While we know a lot about the evolution of regulation of gene expression after a WGD event, we still don't know much about the evolution of AS.

1.3 Alternative splicing after gene duplication

WGD, as well as single gene or segmental duplications (SGD), and AS are two phenomena that lead to the diversification of protein function: duplications provide the raw material for genetic novelties to evolve, and AS generates multiple proteins from a single gene. The evolutionary relationship between these processes remains unclear (Iñiguez & Hernández, 2017), but there are three proposed models of AS evolution following WGD: the independent model, the function-sharing model, and the accelerated AS model (Figure 1.3).

The independent model states that no relationship exists between the evolution of isoforms and duplication, meaning that duplicates and singletons would have the same number of isoforms. Under the function-sharing model, duplicates subdivide the ancestral isoforms, similarly to subfunctionalization (Figure 1.2), decreasing the number of isoforms per gene

(Kopelman et al., 2005; Su et al., 2006). One or both duplicates accumulate isoforms in the accelerated AS model, giving them on average more isoforms than singletons due to relaxed functional constraints, which facilitates the gain of new functions through novel isoforms (Jin et al., 2008).

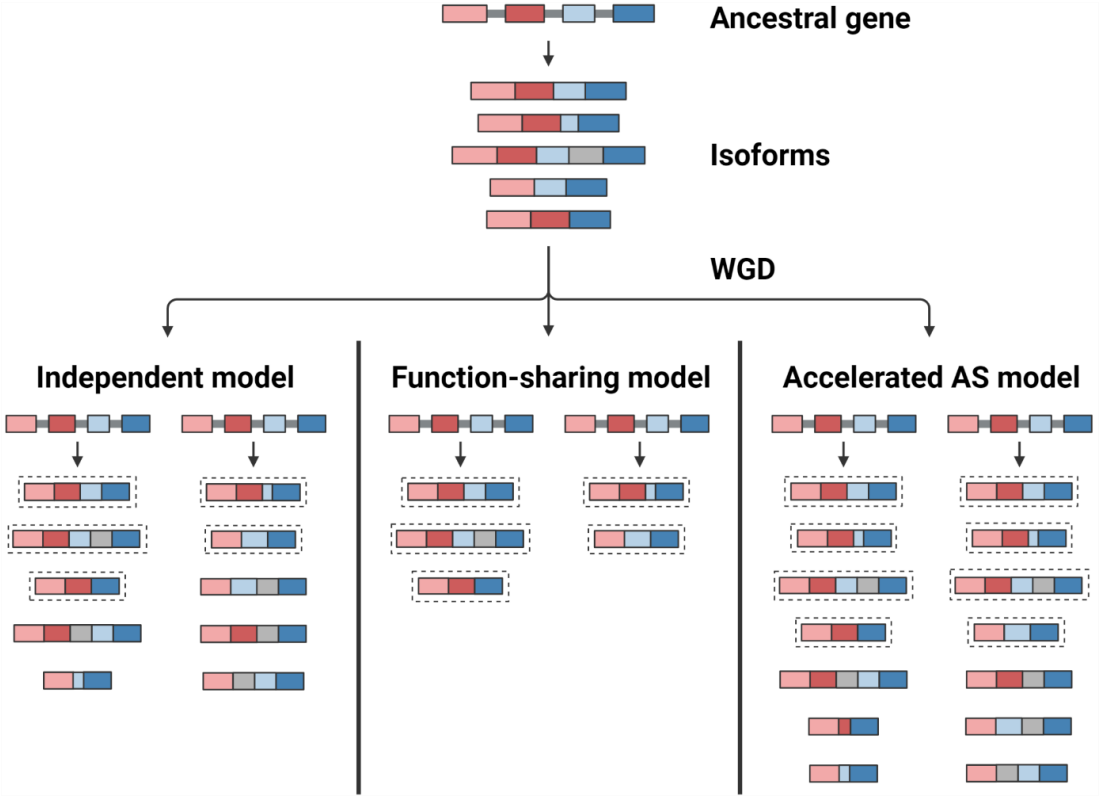

Figure 1.3: Models for AS evolution following gene duplication. *Colored boxes represent exons, and grey boxes represent introns. Isoforms marked by dashed lines represent ancestral isoforms. Figure is created using BioRender (https://biorender.com/) and is adapted from Iñiguez and Hernández (2017), as permitted by the Creative Commons license (https://creativecommons.org/licenses/by/4.0/).*

There are contrasting hypotheses surrounding the evolutionary relationship between AS and gene duplication. One hypothesis argues that the function-sharing model (Figure 1.3) can explain the inverse correlation between gene family size and AS. Smaller gene families have more AS events than large gene families to compensate in terms of functionality (Kopelman et al., 2005; Su et al., 2006), leading duplicated genes to have fewer isoforms than singletons. An opposing idea, the duplicability-age hypothesis, explains this inverse correlation by stating that older gene duplicates have acquired more AS events than more recent duplications, and that genes with low levels of AS tend to duplicate more frequently (Roux & Robinson-Rechavi, 2011). This hypothesis argues that time is a crucial factor in acquiring novel

AS forms, and that duplicates tend to accumulate isoforms over time due to relaxed purifying selection.

Most studies conducted on the subject have used gene family size to measure duplication. By using gene family size, the focus has primarily been on SGDs, and there have been few studies exploring the models of AS evolution following WGD (Wang & Guo, 2021). The type of gene duplication, be it WGD or SGD, could be an important factor in AS evolution (Iñiguez & Hernández, 2017).

1.4 New opportunities through long-read transcriptome sequencing

Over the past decades, the standard transcriptome profiling method has been massive parallel short-read sequencing, also referred to as next-generation sequencing (NGS), with platforms such as Illumina being one of the most popular (Kanzi et al., 2020). All NGS methods rely on fragmenting the RNA strands during library preparation and assembling the fragments after sequencing (Behjati & Tarpey, 2013). Read lengths for the fragments are typically 75-150bp (Besser et al., 2018). Although these short reads can be efficiently mapped to the genome and be used to quantify the transcript levels of whole genes or exons, reconstructing the complete transcriptome, including the diversity of alternatively spliced isoforms and low abundance transcripts, has been difficult (Garber et al., 2011).

More recently, new sequencing technologies that have read lengths surpassing 100 Kb (Besser et al., 2018) have resolved some of the shortcomings of short-read technologies when it comes to transcript isoform analyses. With one read covering the entire transcript, this eliminates the need to assemble after sequencing, and we can distinguish between splice variants and determine exon connectivity from these long, single-molecule reads (Tang et al., 2020). Therefore, long reads will provide a more accurate picture of the isoform diversity than short reads.

While long-read RNA sequencing has some apparent advantages over NSG sequencing, one drawback has been the basecalling error rate of the long reads. NSG platforms have error rates of ~0.1% (Kchouk et al., 2017), and long-read sequencers have had error rates of >10% (Amarasinghe et al., 2020). However, this number is steadily decreasing as the methods are improved. The current estimates for error rates are <1% for the PacBio sequencing platforms (Wenger et al., 2019) and <5% for the Oxford Nanopore Technologies sequencing platforms (Jain et al., 2018).

1.4.1 Oxford Nanopore sequencing

The idea of using a transmembrane protein pore to sequence DNA and RNA molecules was conceived more than 40 years ago. However, the first Nanopore sequencing device, the hand-held MinION manufactured by Oxford Nanopore Technologies (referred to as ONT), was not commercially available until 2014 (Deamer et al., 2016). The higher throughput sequencing device PromethION was launched four years later (Oxford Nanopore Technologies, 2022a).

All ONT sequencing devices use flow cells during sequencing. These flow cells have an array of chambers with a nanopore embedded in an electrically resistant membrane, with each chamber connected to a sensor chip. A single strand of DNA or RNA is sequenced by threading the strand through a protein pore embedded in the membrane (Figure 1.4). When applying an electric current, the DNA or RNA strands pass through the nanopore and disrupt the current, creating a characteristic signal called a “squiggle”, which is measured using the sensor chip. Base-calling algorithms determine the sequence of the strand in real-time based on the squiggle (Oxford Nanopore Technologies, 2022b).

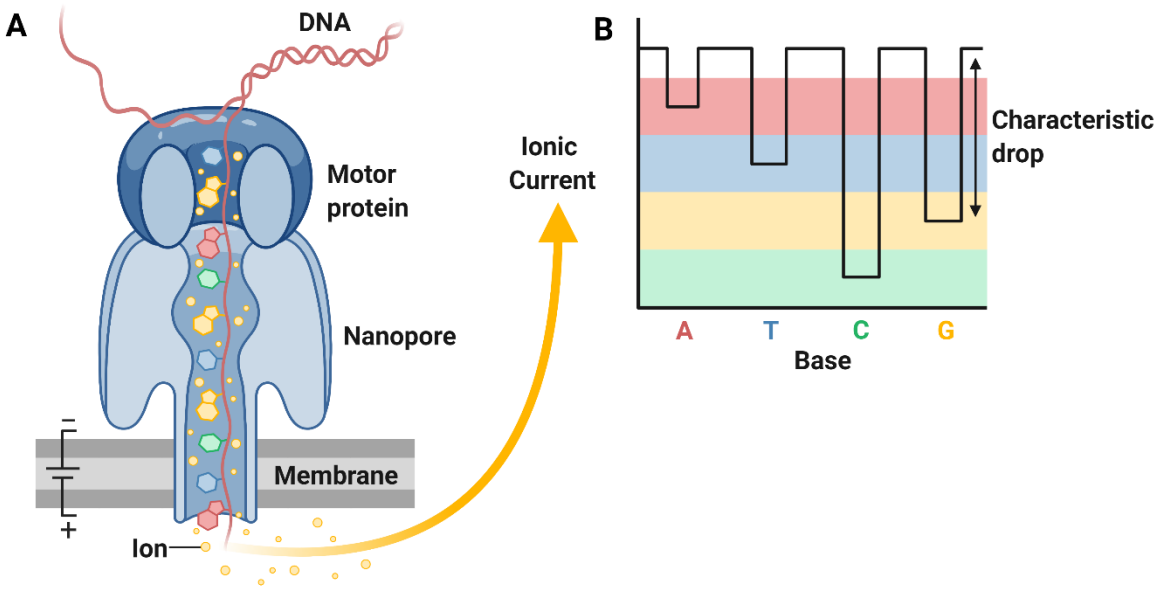

Figure 1.4: Oxford Nanopore sequencing. A) *A strand of nucleic acid is unwound and translocated through the nanopore by a motor protein. The membrane is electrically resistant, which causes the flow of ions to pass through the nanopore with the nucleic acid.* ***B)*** *Each base gives a characteristic reduction in the ionic current, which is measured. Figure created with BioRender (https://biorender.com/).*

In sequencing nucleic acid when using Nanopore, there are two sources of error: the sequencing itself and basecalling. Because multiple nucleotides occupy the space within the pore simultaneously and the strand translocation speed varies, the translation from squiggle to sequence is complex. Repeated regions of nucleotides further complicate the basecalling process because it is difficult to determine the length of these regions (Rang et al., 2018). Refinements in flow cell chemistry and basecalling algorithms further reduce these sources of error, improving the sequence quality.

ONT has three kits available for RNA sequencing: direct RNA sequencing, direct cDNA sequencing, and cDNA-PCR sequencing (Oxford Nanopore Technologies, 2022c). The direct RNA sequencing kit allows for the detection of base modifications but requires a high input of RNA (500 ng of poly-A RNA) and has lower yields than the cDNA kits. Both cDNA kits include the reverse transcription of RNA to cDNA prior to sequencing. The poly-A RNA inputs are 1 ng and 100 ng for the cDNA-PCR and direct cDNA kits, respectively. The PCR step is the main difference between the cDNA-PCR and the direct cDNA kits, and while sequencing using direct cDNA eliminates PCR bias, the higher amounts of RNA needed for this kit limit its use to the amount of RNA available.

1.5 Thesis aims and objectives

This master project is part of the REWIRED project, which aims to better our understanding of genome evolution after WGD events by using salmonid fishes as model species to study the impact of WGDs on novel gene functions and adaptation.

This thesis has two main objectives. First, to characterize the diversity of isoforms in Atlantic salmon using ONT long-read RNA sequencing. Secondly, utilizing the discovered isoforms to test the predictions of the AS evolution models, thereby uncovering if there is an apparent relationship between AS and WGD.

2 Methods and materials

2.1 RNA samples from salmon tissues

The RNA samples used in this master thesis came from salmon individuals used in the AQUA-FAANG project funded by EU (https://www.aqua-faang.eu/), where the aim is to generate functional genome annotations of important aquaculture species in Europe. RNA was isolated from males and females from five tissues: brain, gill, head kidney, liver, and muscle (Table 2.1).

Table 2.1: Overview of RNA samples. *12 RNA samples collected from brain, gill, head kidney (HK), liver, and muscle tissues from salmon. BA RIN values (Bioanalyzer RNA integrity number) from the AQUA-FAANG project is reported.*

Tissue	***Individual***	***BA RIN***
Brain	A19	8.5
	A22	10
Gill	A14	9.4
	A19	9.5
Head kidney	A19	9.8
	A20	8.8
Liver	A19	8.2
	A20	8.0
	A22	8.1
Muscle	A19	9.6
	A20	9.7
	A22	9.7

2.2 Sequencing

We used the PCR-cDNA Barcoding protocol (Oxford Nanopore Technologies, 2019) to sequence the transcriptome on the ONT PromethION sequencing machine (Figure 2.1). The protocol was chosen to accommodate the amount of RNA left in each sample and eliminate the need for upconcentration. The approach is described in the sections Library preparation and Sequencing on PromethION.

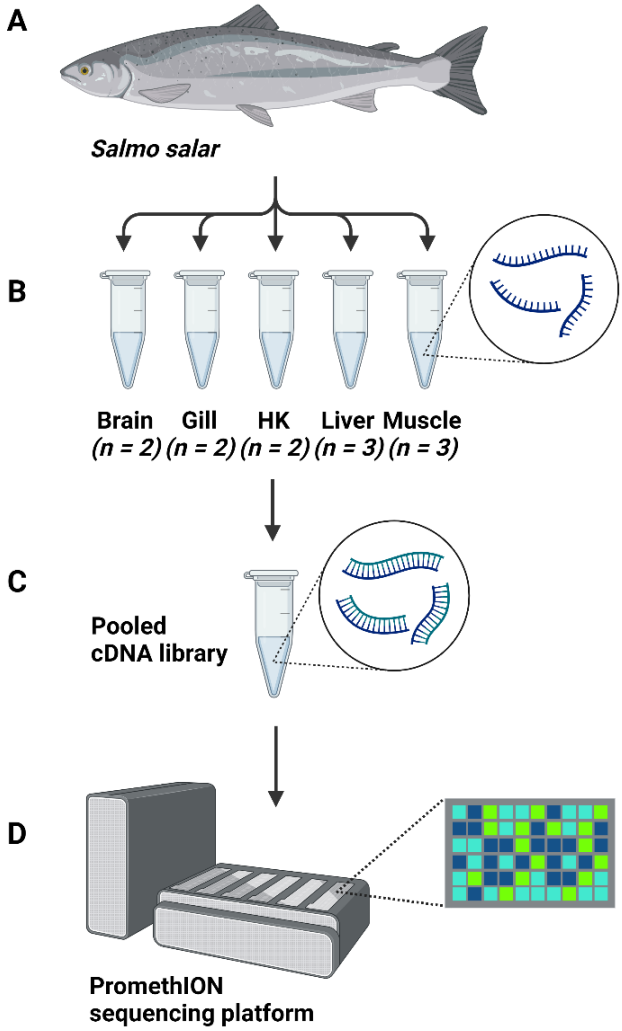

Figure 2.1: Overview of mRNA sequencing using ONT. A) *Samples used in this thesis came from RNA samples collected from Atlantic salmon for the AQUA-FAANG project.* ***B)*** *12 RNA samples were used in library preparation: 2 brain, 2 gill, 2 head kidney, 3 liver and 3 muscle.* ***C)*** *The 12 RNA samples were prepared using the PCR-cDNA barcode protocol from ONT* (Oxford Nanopore Technologies, 2019) *and subsequently pooled into one library.* ***D)*** *The pooled cDNA library was run on PromethION and simultaneously basecalled and demultiplexed. Figure created with BioRender (https://biorender.com/).*

2.2.1 Library preparation with PCR-cDNA barcoding protocol

Consumables used in library preparation are listed in Appendix 1, PCR-cDNA Barcoding Sequencing Kit (SQK-PCB109) contents (referred to as the kit from now on) are listed in Appendix 2, and the Flow Cell Priming Kit (EXP-FLP002) contents are listed in Appendix 3.

We performed library preparation of the RNA using the PCR-cDNA barcoding protocol from ONT (Oxford Nanopore Technologies, 2019). The input needed for library preparation was ~60 ng RNA from each sample. To measure the initial concentration, we used Qubit fluorometer, and we diluted the samples in RNase-free water to get a volume of less than 9 µl (Table 2.2).

Table 2.2: Concentrations, dilutions, and final volume for RNA samples. *Initial concentrations were measured using a Qubit fluorometer, and samples were diluted in nuclease-free water. The volume for each diluted RNA sample used in library preparation.*

Tissue	***Individual***	***Qubit (ng/µl)***	***Dilution ratio***	***Volume (µl)***
Brain	A19	117	1:10	5.1
	A22	120	1:10	5.0
Gill	A14	266	1:10	2.3
	A19	150	1:10	4.0
Head kidney	A19	394	1:10 → 1:2	3.0
	A20	400	1:10 → 1:2	3.0
Liver	A19	876	1:10 → 1:2	1.4
	A20	314	1:10 → 1:2	3.8
	A22	798	1:10 → 1:2	1.5
Muscle	A19	125	1:10	4.8
	A20	125	1:10	4.8
	A22	89	1:10	6.8

We mixed each diluted RNA sample (volumes in Table 2.2) with 1 µl VN Primers from the kit to select mRNA from the RNA samples, as the VN Primers anneal to the polyA-tail of mRNAs. 1 µl of 10mM dNTPs were also added to provide nucleotides for reverse transcription and PCR amplification in later steps. This was mixed with RNase-free water in 0.2 ml PCR tubes to get a total volume of 11 µl for each sample before incubation at 65° C for 5 minutes in a Thermal Cycler for primer annealing. We prepared a buffer for each sample by mixing 2 µl 10 µM Strand-Switching Primers from the kit with 1µl nuclease-free water, 4 µl 5x RT Buffer (Appendix 1), and 1µl RNaseOUT, and then added it to the annealed mRNA samples for a total volume of 19 µl. The samples were incubated at 42° C for 2 minutes to anneal the Strand-

Switching primers before adding 1 µl Maxima H Minus Reverse Transcriptase to each sample. We incubated the samples at 42° C for 90 minutes to perform strand-switching (from RNA to cDNA) by reverse transcription, then at 85° C for 5 minutes to inactivate the Reverse Transcriptase.

Table 2.3: Barcodes for each sample. *Barcodes from the kit (BP01-12) were added to each sample.*

Tissue	***Individual***	***Barcode***
Brain	A19	01
	A22	02
Gill	A14	03
	A19	04
Head kidney	A19	05
	A20	06
Liver	A19	07
	A20	08
	A22	09
Muscle	A19	10
	A20	11
	A22	12

To prepare for the PCR amplification step, we created a PCR mix for each sample containing 25 µl LongAmp Taq polymerase, 18.5 µl nuclease-free water and 5 µl cDNA sample, adding 1.5 µl of a different barcode primer from the kit (Table 2.3) to each sample for a total volume of 50 µl. Cycling conditions were set to one round of initial denaturation at 95° C for 1 minute, then 14 cycles that included denaturation at 95° C for 15 seconds, annealing at 62° C for 15 seconds, and extension at 65° C for 6 minutes, followed by a final extension at 65° C for 6 minutes. To the amplified cDNA samples, we added 1 µl Exonuclease I, followed by two rounds of incubation: first at 37° C for 15 minutes to degrade single-stranded cDNA in the sample, followed by 80° C for 15 minutes to inactivate the exonuclease. The barcoded samples were transferred to 1.5 ml Eppendorf DNA LoBind tubes.

To clean the amplified cDNA samples, we first resuspended the AMPure XP beads by vortexing before adding 40 µl of the beads to each sample. We incubated the samples for 5 minutes at room temperature on a Hula mixer at 10 rounds per minute to make the cDNA adhere to the beads and prepared 500 µl of 70% fresh ethanol for each sample by mixing ethanol with nuclease-free water. The samples were pelleted on a magnetic rack before pipetting off the supernatant. While on the magnet, we used 200 µl 70% ethanol to wash the

beads, and the ethanol was pipetted off before we repeated the step. After taking the samples off the magnetic rack, we added 12 µl Elution Buffer from the kit to each sample to elute the cDNA off the beads. We incubated the samples for 10 minutes at room temperature on a Hula mixer at 10 rounds per minute before pelleting them again on the magnetic rack. The eluates containing the cDNA libraries were removed and transferred to new 1.5 ml Eppendorf DNA LoBind tubes, one for each sample.

We performed quality control of the cDNA libraries by measuring concentration on the Qubit fluorometer and estimating fragment length using the TapeStation System. The samples were diluted in Elution Buffer from the kit so that each sample contained ~120 fmol cDNA (Table 2.4). All samples were then pooled together in a 1.5 Eppendorf DNA LoBind tube (Table 2.4) containing a final cDNA library. 3 µl Elution Buffer was added to the pooled samples to get a total volume of 23 µl, then we added 1 µl of Rapid Adapter from the kit containing the motor protein used in sequencing. Lastly, we incubated the library for 5 minutes at room temperature for adapter attachment.

Table 2.4: Quality control of cDNA library. *Concentrations for each cDNA library were measured using a Qubit fluorometer. Fragment length estimates were measured using the TapeStation System. Volume extracted from each sample when pooling together.*

Tissue	***Individual***	***Qubit (ng/µl)***	***Length (bp)***	***Dilution ratio***	***Volume (µl)***
Brain	A19	7.46	1912	1:2	1.8
	A22	2.74	1847	1:1	2.1
Gill	A14	10.40	1705	1:3	2.2
	A19	11.70	1538	1:3	1.8
Head kidney	A19	23.20	1639	1:3	1.2
	A20	31.00	1379	1:5	1.2
Liver	A19	16.30	1731	1:4	1.9
	A20	23.00	1867	1:4	1.6
	A22	22.20	1794	1:4	1.8
Muscle	A19	27.00	1851	1:4	1.5
	A20	31.80	1879	1:4	1.4
	A22	23.60	1684	1:4	1.7

2.2.2 Sequencing on PromethION

We sequenced the cDNA library on a R9.4.1 PromethION flow cell (FLO-PRO002). The flow cell was stored at 4° C before use but inserted into the PromethION sequencing platform and acclimated for 30 minutes at room temperature before loading the library. To prime the flow cell, we made a flow cell priming mix by mixing 30 µl of Flush Tether into a tube of Flush Buffer, both from the kit. The Flush Tether brings the cDNA library close to the flow cell membrane containing the nanopores. To eliminate air bubbles, we drew out a small buffer volume from the flow cell by inserting a pipette into the inlet port and dialing the wheel on the pipette from 200 µl to 240µl. We flushed 500 µl of the Priming Mix through the inlet port of the flow cell, and after waiting 5 minutes, we flushed the flow cell again. Before loading the library on the flow cell, we mixed the pooled cDNA library with 75 µl Sequencing Buffer and 51 µl Loading Beads, both from the kit. We then loaded the library on the flow cell through the inlet port and let it sit at room temperature for 30 minutes before sequencing to settle the library onto the flow cell.

We washed the flow cell and loaded the library a second time. The first run lasted 45 hours and 25 minutes, while the second run lasted 72 hours. Both runs were basecalled and demultiplexed by Guppy v5.0.17.

2.3 Data analysis

All sequences with the same barcode were saved as one FASTQ-file. We concatenated the FASTQ files with identical barcodes from each run, creating one file for each of the 12 samples.

2.3.1 Isoform detection

The python implemented FLAIR v1.5 (Full-Length Alternative Isoform analysis of RNA) pipeline was used (Tang et al., 2020) to identify isoforms. This tool is developed to identify isoforms from long sequencing reads with high error rates, such as Nanopore reads (Tang et al., 2020). To generate a final high-confidence isoform reference (HCIR) for annotation and quantification, the FLAIR pipeline includes four main steps: (i) extracting splice junctions from short reads, (ii) alignment and correction, (iii) collapsing and quantification, and (iv) prediction of productivity (i.e., functionality) (Figure 2.2).

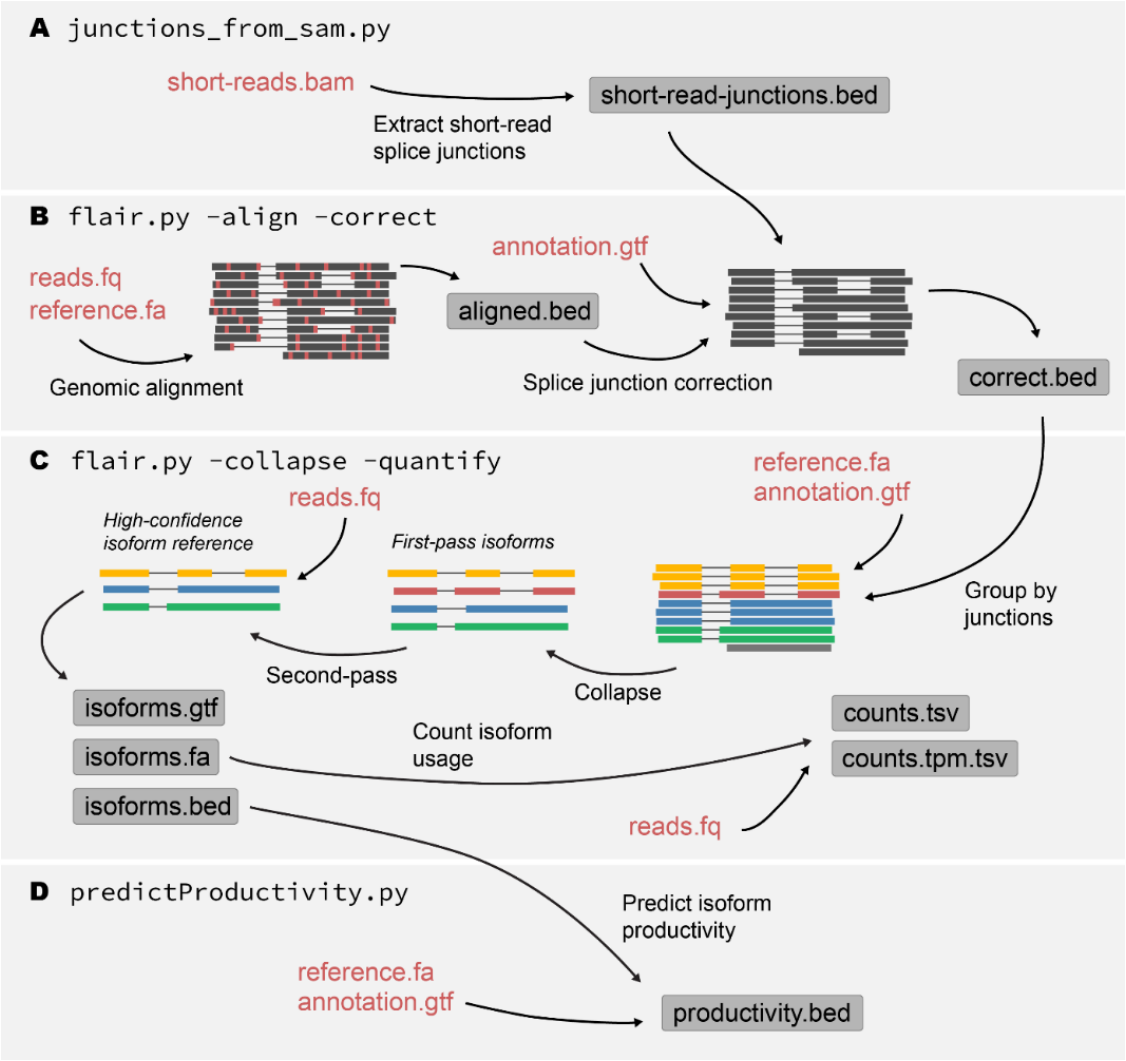

Figure 2.2: FLAIR pipeline for isoform detection*. Input files marked in red, output marked by grey boxes.* ***A)*** *Splice junctions from aligned short reads were extracted.* ***B)*** *Modules align and correct were run together. First, mRNA sequences are aligned to a reference genome, with errors shown in red. The aligned sequences are corrected by annotated splice junctions and short-read splice junctions.* ***C)*** *The corrected reads are grouped by splice junctions and are collapsed into a first-pass isoform set. Isoforms with 3 or more supporting reads are kept as a high-confidence isoform reference. The reference is used to quantify the number of reads per isoform for each sample.* ***D)*** *The productivity for each isoform is predicted based on the presence of start and termination codons in open reading frames. Figure adapted from Tang and colleagues* (2020)*, as permitted by the Creative Commons license (https://creativecommons.org/licenses/by/4.0/).*

The first step, extracting splice junctions from aligned short reads (Figure 2.2 A), is optional. Splice junctions from short reads are used to improve the confidence of the borders between splice junctions. The stand-alone script `junctions_from_sam.py` was run using default parameters to extract short-read splice junctions.

To execute the second step of alignment and correction, the command '`flair.py -align -correct`' was run using default parameters with the added parameter to include short-read junctions (Figure 2.2 B). We used the ICSASG_v2 genome assembly (GenBank accession ID: GCA_000233375.4) from Ensembl as a reference genome. The reads are aligned to the reference genome, producing a file containing the aligned reads. Because the reads contain a high number of sequencing errors, they are corrected using the annotation file from the ICSASG_v2 genome assembly and the splice junctions we extracted from the short reads and produces a file containing the corrected reads.

To create an HCIR, we ran the third step of the pipeline (Figure 2.2 C), '`flair.py -correct -quantify`' using default parameters, as well as a parameter to include the annotation file from the ICSASG_v2 genome assembly for naming known isoforms using Ensembl gene and transcript IDs, as well as a parameter to enable TPM normalization of the read counts. In this part of the pipeline, isoforms are grouped by splice junctions, and each isoform group is collapsed into a set of first-pass isoforms. The raw reads are aligned to the first-pass isoforms to filter out any isoforms with less than 3 supporting reads. This step produces an annotation file, a BED file, and a FASTA file containing the sequences of the isoforms. The reads are aligned to the isoform FASTA file resulting in two types of isoform quantifications: absolute read counts and normalized read counts using transcript per millions (TPM). TPM was estimated as follows:

$$TPM_i = \frac{q_i/l_i}{\sum_j (q_j/l_j)} * 10^6 \qquad \textit{Equation 2.1}$$

In Equation 2.1 q_i is the denotation for reads mapped to transcripts, l_i is the transcript length and $\sum_j (q_j/l_j)$ is the sum of the mapped reads divided by transcript length (Zhao et al., 2021).

The third step was run using the command '`predictProductivity.py`' with default parameters (Figure 2.2 D). This module predicts isoform productivity by comparing coding sequence and translation frame information from the original reference annotation with the discovered isoforms. The output of this step is a BED file containing the predicted productivity of each isoform assigned to one of four categories: productive (*PRO*), no start codon (*NSC*), no termination codon (*NTC*), or premature termination codon (*PTC*).

2.3.2 Isoform analysis

We performed various analyses using RStudio v2021.9.2.382 (RStudio Team, 2022) and visualized results using the *ggplot2* (Wickham, 2016) and the *pheatmap* (Kolde, 2019) packages.

The FLAIR pipeline assigned each isoform to a gene and transcript ID, which we used to determine if isoforms matched previously annotated genes or transcripts. Isoforms that matched annotated genes in the reference genome were assigned the corresponding Ensembl gene ID, and isoforms that matched annotated transcripts were assigned the corresponding Ensembl transcript ID. Isoforms that did not map to an annotated transcript were assigned FLAIR transcript IDs based on the splice junction chain of that isoform. Isoforms that did not map to annotated genes in the reference genome were assigned chromosomal coordinates as the novel gene ID. Isoforms with a match to an Ensembl gene ID were categorized as *Known gene,* and the isoforms without were annotated as *Intergenic*. Isoforms with an Ensembl transcript ID were annotated as *Known isoform*, while isoforms with a FLAIR transcript ID were annotated as *Novel isoform*.

2.3.3 Comparing ohnologs

To identify isoforms that mapped to ohnologs and singletons, we used a published dataset of ohnolog and singleton classifications (Bertolotti et al., 2020). Wilcoxon rank-sum test was performed in R using the *wilcox.test()* function to test if there was a difference in the median number of isoforms between ohnolog and singleton genes.

To find the difference in the number of isoforms between ohnologs, we calculated the absolute difference in isoform numbers:

$$\Delta_{\#i} = |\#i_1 - \#i_2| \qquad \textit{Equation 2.2}$$

In Equation 2.2, $\#i_1$ is the isoform count for one gene in an ohnolog pair, and $\#i_2$ is the isoform count for the other gene.

We determined the tissue specificity of each isoform by using the *tispec* package in R (Galbi, 2019). The *tispec* package calculates tissue specificity for each isoform using the tau (τ) algorithm:

$$\tau = \frac{\sum_{i=1}^{N}(1 - \hat{x_i})}{n-1};\ \hat{x_i} = \frac{x_i}{\max\limits_{1 \le i \le n}(x_i)}$$
$$\tau \geq 0.85 = Tissue\ specific$$
$$\tau < 0.85 = Broadly\ expressed$$

Equation 2.3

In Equation 2.3, n is the number of tissues, x_i is the TPM normalized read count of the isoform in tissue i (Kryuchkova-Mostacci & Robinson-Rechavi, 2016). x_i was $\log_2$-transformed prior to calculating τ. To determine whether an isoform was tissue-specific or broadly expressed, we assigned the cut-off value to 0.85 (Kryuchkova-Mostacci & Robinson-Rechavi, 2016; Yanai et al., 2005). Isoforms with $\tau \geq 0.85$ were labeled *Specific*, and isoforms with $\tau < 0.85$ were labeled *NonSpecific*. The tissue-specific isoforms were assigned with the tissue each isoform was specific to (i.e., *Brain*, *Gill*, *HK*, *Liver*, and *Muscle*).

3 Results

After counting the number of uniquely identified isoforms (Figure 3.2), the absolute read counts for each sample (Figure 3.3), the mean number of isoforms (Figure 3.4) and distribution of expression (Figure 3.5) and the length of each isoform (Figure 3.6) in the *Known gene* and *Intergenic* categories, we discarded all isoforms from the *Intergenic* category. After identifying the ratio of uniquely identified *Known isoforms* and *Novel isoforms* for each category of estimated protein productivity (Figure 3.7), we proceeded with the productive (*PRO*) isoforms when testing the evolutionary models of AS after WGD (Figure 1.3).

3.1 Sequencing runs

The cDNA sequencing libraries were loaded twice on a single PromethION flow cell. The first and second runs un yielded 84 and 12.3 million reads, respectively. However, the sequence read length distribution (i.e., N50) was the same for both runs (Table 3.1).

Table 3.1: Statistics for each sequencing run on PromethION*. Passed bases are the number of reads that passed basecalling with sufficient signal quality, while failed bases could not be confidently determined.*

	First run	**Second run**
N50 (Kb)	1.1	1.1
Passed bases (Gb)	51.2	5.5
Failed bases (Gb)	14.8	3.3
Reads (M)	84.0	12.3

The expected number of reads when using the PCR-cDNA barcoding protocol and sequencing on PromethION is >60 million (Oxford Nanopore Technologies, 2022c), which was achieved in the first run. An average read length of 1 Kb is also expected, which is close to the estimated N50.

3.2 Isoform diversity

3.2.1 Data quality

To evaluate the data quality and detect potential outliers, we performed hierarchical clustering of the normalized isoform read counts and visualized data as a heat map (Figure 3.1). From Figure 3.1, we see that tissues cluster together, indicating that the expression profile of each tissue is more similar to each other than they are to other tissues, indicating good data quality.

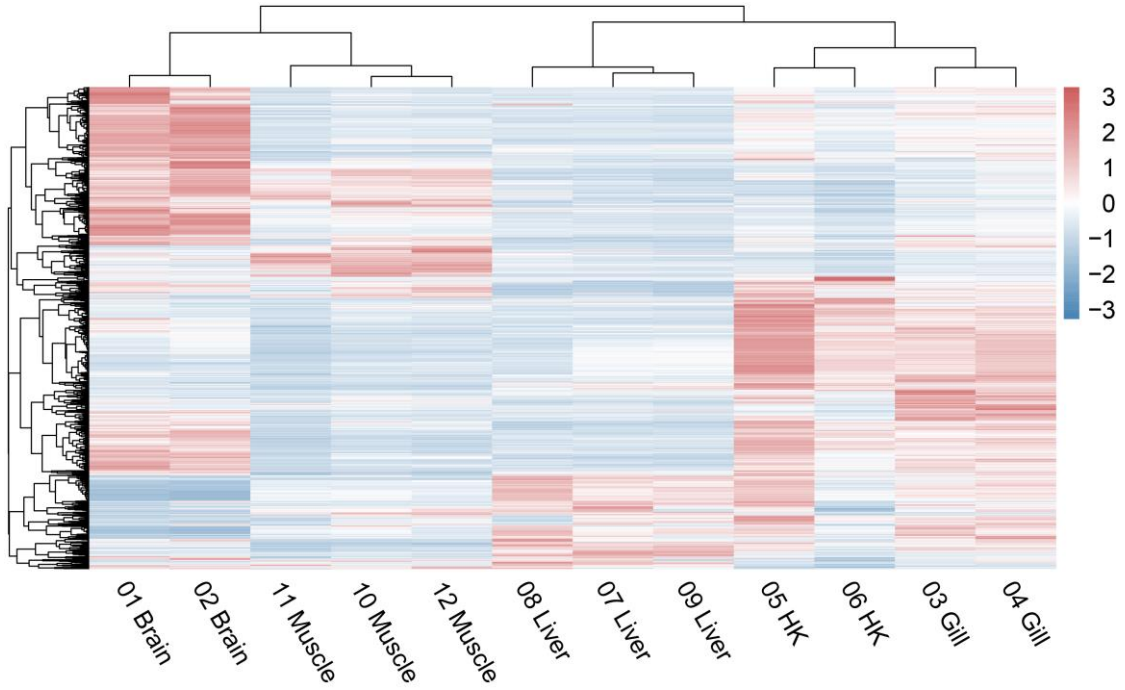

Figure 3.1: Heatmap of TPM normalized read counts. *Row-normalized read counts are plotted, and Pearson correlation distances are used for the hierarchical clustering of rows and columns. HK = head kidney.*

3.2.2 Isoform category distribution

Based on the annotation from section 2.3.1, Figure 3.2 shows the number of uniquely identified isoforms from the sequenced RNA samples. Across all samples, we found 326 544 uniquely identified isoforms. 255 109 of these isoforms originated from 227 910 different intergenic regions (i.e., not previously annotated as genes), and 71 435 of the isoforms originated from 27 967 known genes. Only 25% were previously annotated isoforms in the latter category, and 75% were novel isoforms.

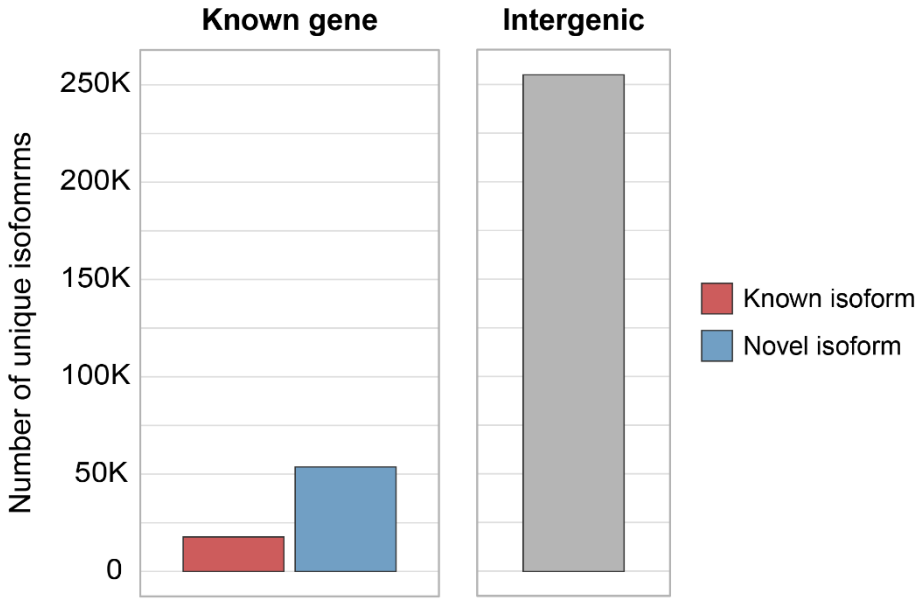

***Figure 3.2: Number of uniquely identified isoforms**. All isoforms originating from intergenic regions are novel isoforms.*

Read count quantification based on out total set of isoforms showed that 23% (22.4M) of the ONT reads could be assigned to an isoform. Out of these, 7.3M (31%) of the reads were from isoforms from intergenic regions (Figure 3.3). Among the reads mapping to known genes, 15.6M (69%) were from known isoforms, and 6.8M (31%) of the reads were novel isoforms.

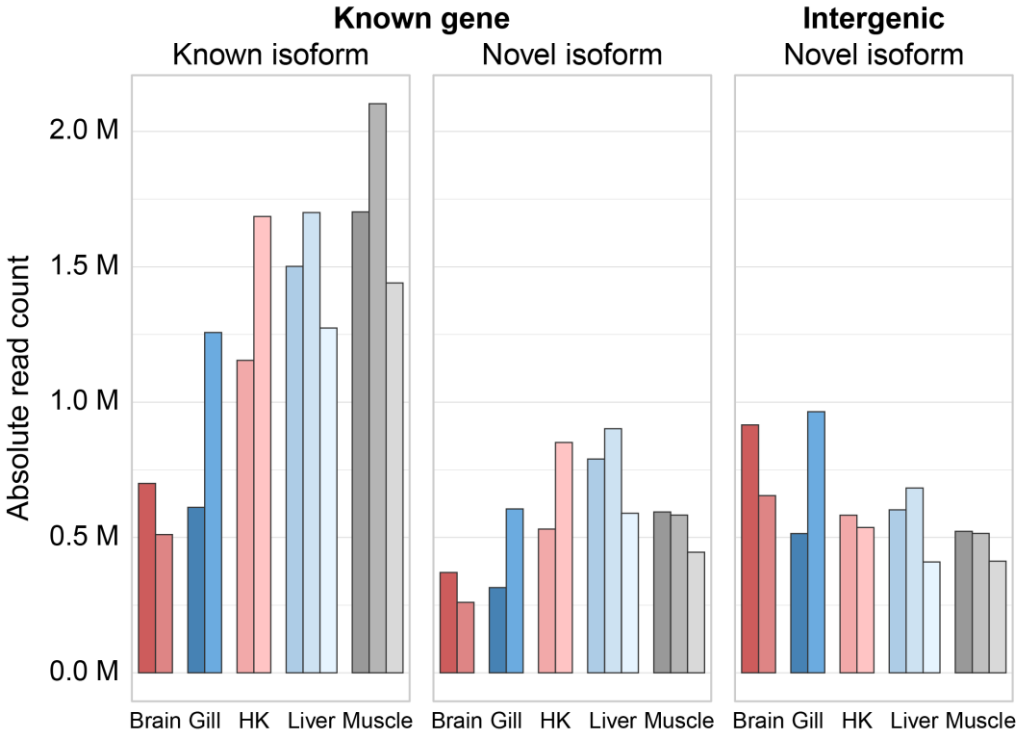

Figure 3.3: Absolute read counts per sample. *Each bar of the same shade of each color represents reads from the same sample. HK = head kidney.*

The proportion of reads originating from novel isoforms differed among the tissues (Figure 3.3). Muscle had for example the highest number of reads originating from known isoforms; however, this was not the trend for novel isoforms. The other tissues had similar proportions of known and novel isoforms.

The variation in isoform number per gene (known genes) ranged from 1-220, with an average of 2 isoforms per gene, or 1.1 known and 2.7 novel isoforms (Figure 3.4). Of the unique isoforms originating from known genes, most were novel isoforms. However, the known isoforms were more highly expressed compared to the novel isoforms (Figure 3.5). Isoforms that did not map to a known gene had isoforms with the lowest mean and median expression (Figure 3.5), and an average of 1.1 isoforms per locus (Figure 3.4).

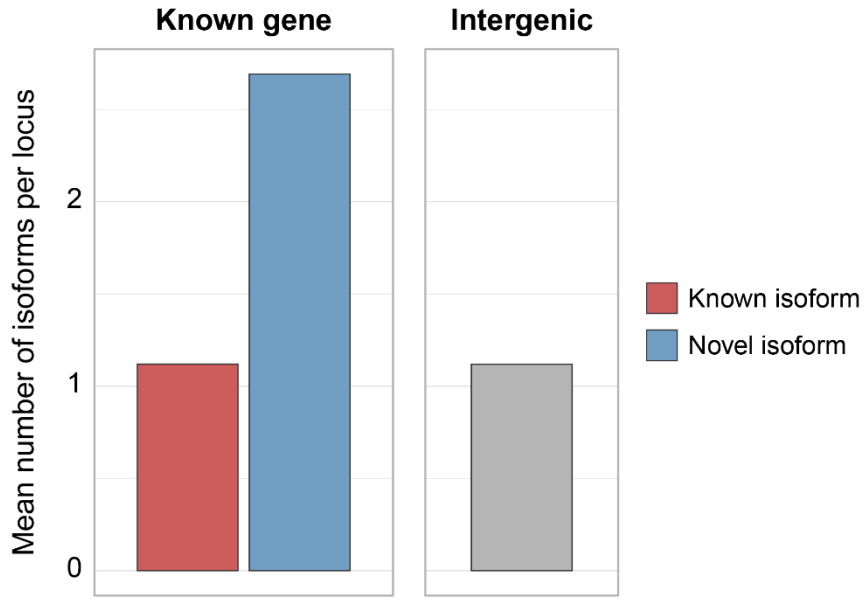

Figure 3.4: Mean number of isoforms per locus. *The average number of known and novel isoforms per gene was 2.7 and 1.1, respectively. Intergenic regions with mapped isoforms had an average number of 1.1 isoforms.*

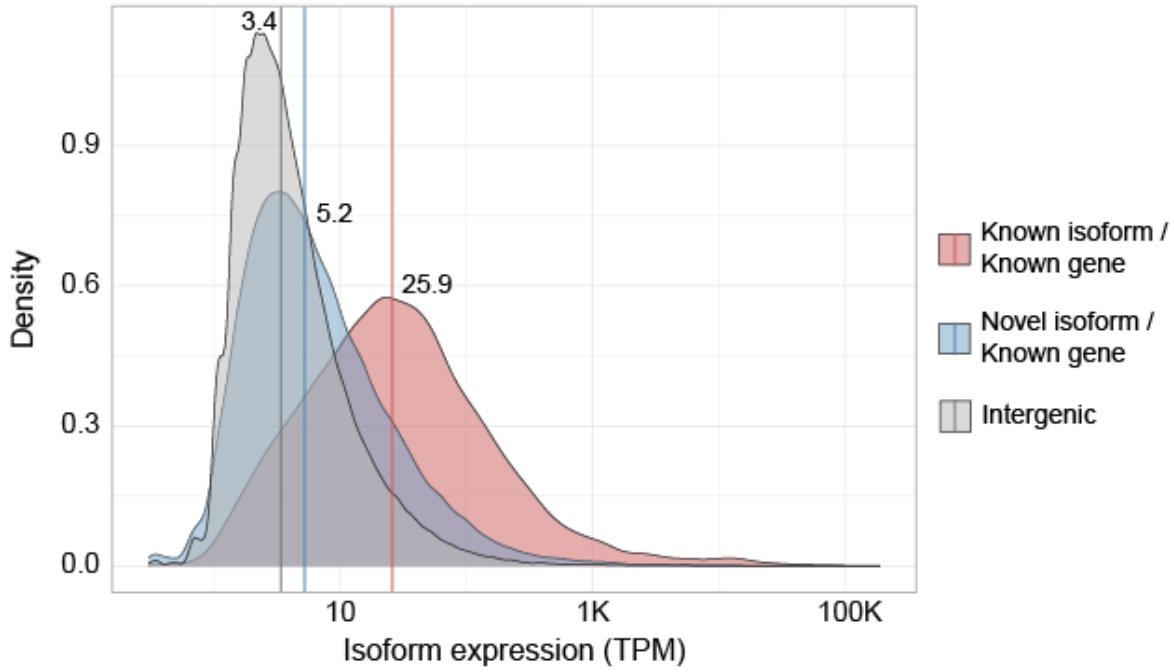

Figure 3.5: Distribution of expression of isoforms, measured in TPM. *Medians are marked with lines, and the mean expression for known and novel isoforms originating from known genes was 345.7 and 50.5, respectively. Isoforms that mapped to intergenic regions had a mean expression of 12.4.*

3.2.3 The protein-coding potential of isoforms

While AS increases the number of proteins a gene can produce, both the transcription and splicing processes are prone to error (Gordon et al., 2015; Hsu & Hertel, 2009). We can expect a significant proportion of the detected isoforms to be non-coding, and therefore we assessed the number of productive (i.e., predicted to encode a complete protein sequence) and non-productive isoforms using the output from the FLAIR pipeline.

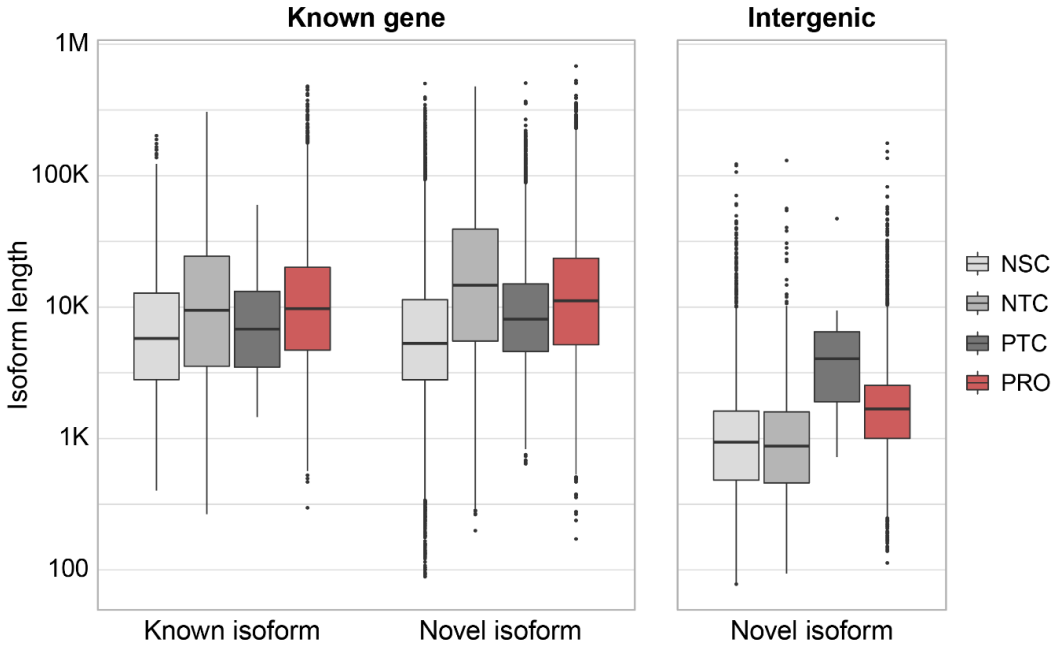

Figure 3.6: Length distribution of the isoforms by productivity. *Productivity categories are no start codon (NSC), no termination codon (NTC), premature termination codon (PTC) and productive (PRO).*

Isoforms originating from intergenic regions had shorter median lengths than isoforms from known genes (Figure 3.6). Most (97%) of these isoforms were estimated to have no start codon. Some isoforms were estimated to be protein-coding, but this was only a minority of cases (2.5%).

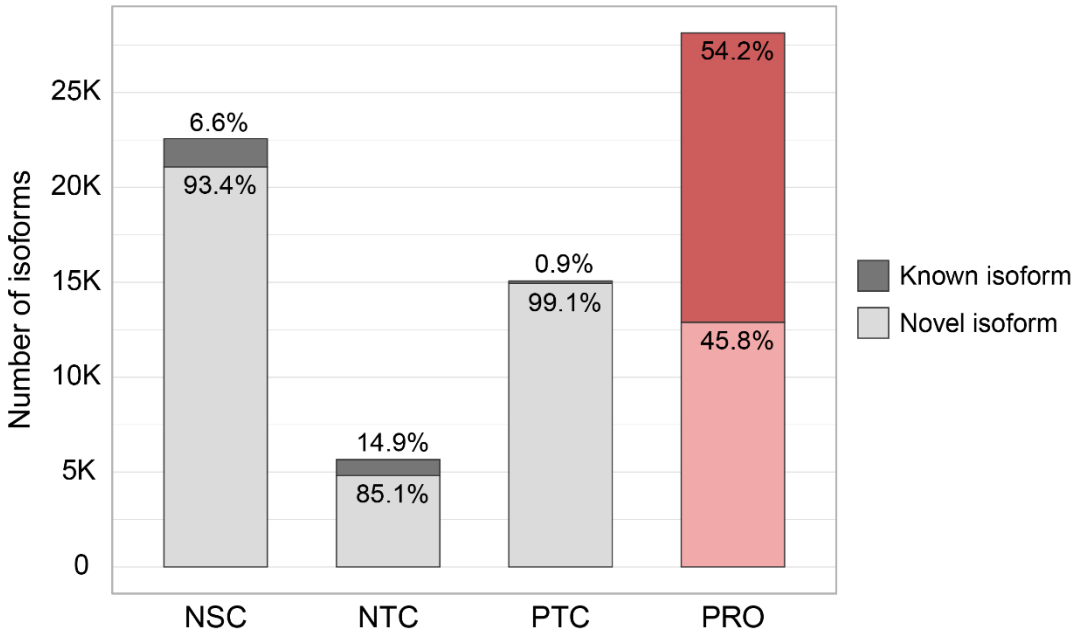

Figure 3.7: Estimated productivity for known and novel isoforms. *Includes all uniquely identified isoforms that mapped to a known gene. Productivity categories are no start codon (NSC), no termination codon (NTC), premature termination codon (PTC) and productive (PRO).*

The FLAIR pipeline uses annotated start codons to predict open reading frames for each isoform and uses this information to predict whether isoforms are productive or non-productive. Considering all isoforms that mapped to known genes (Figure 3.7), 39% were protein-coding. Of the non-productive isoforms, 94% were novel. While most known isoforms (86%) were predicted as productive, we would expect this number to be closer to 100%. This could be the algorithm of the FLAIR tool, due to its limitations, mistakenly classifying protein-producing isoforms as non-productive.

3.3 Expression of isoforms between ohnologs from Ss4R

A long-standing question in genome evolution has been the impact of gene and genome duplication on the evolution of alternative splicing (Iñiguez & Hernández, 2017). One of the thesis aims was to utilize our long-read transcript data to characterize the evolution of AS following WGD in salmonids by exploring the predictions of the proposed models of AS evolution (Figure 1.3).

3.3.1 Difference in isoform number

One expectation from the function-sharing hypothesis (Figure 1.3) is that gene duplication should lead to fewer isoforms for each individual gene. For the accelerated AS model (Figure 1.3), the expectation is that duplication leads to more isoforms in the duplicated genes. To test these predictions, we considered all ohnologs and singletons expressed in the samples and counted the number of productive isoforms for each gene (Figure 3.8). The mean and median number for of isoforms in ohnologs were similar to that of singletons, and we found no statistical difference the median number of isoforms between ohnolog and singleton genes ($Wilcoxon\ rank\ sum,\ P \approx 0.97$).

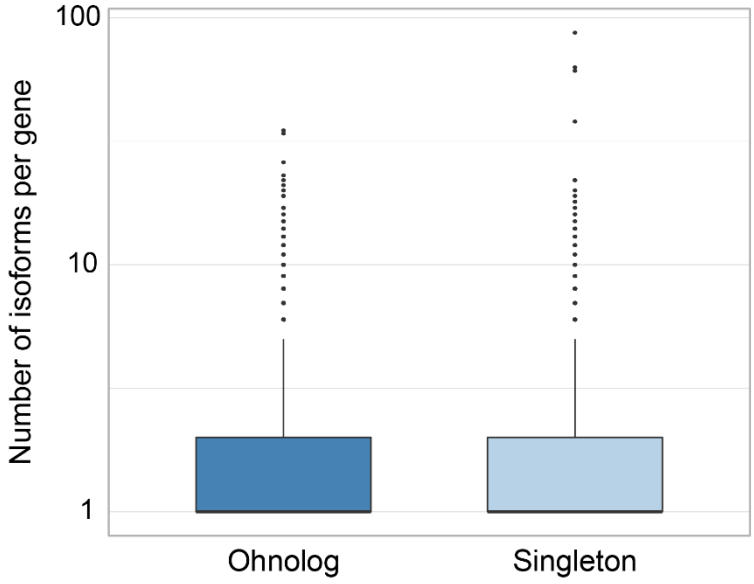

Figure 3.8: Number of isoforms per gene in ohnologs and singletons. *Mean isoform number for ohnolog and singleton genes was 1.6 and 1.7, respectively, and the median was 1 for both ohnolog and singleton genes.*

Asymmetric evolution of regulation and protein-coding sequence among duplicate pairs have been observed across many vertebrate WGDs (Sandve et al., 2018). For gene expression this is characterized by one duplicated gene copy retaining the ancestral regulation while the other evolves under relaxed purifying selection or undergoes functional specialization (Gillard et al., 2021; Sandve et al., 2018). In the context of AS, an expectation under this model, and the accelerated AS model (Figure 1.3), is that one duplicated gene copy retains different numbers of functioning isoforms. To test this, we compared the number of isoforms between the 2 386 ohnolog pairs expressed in the samples.

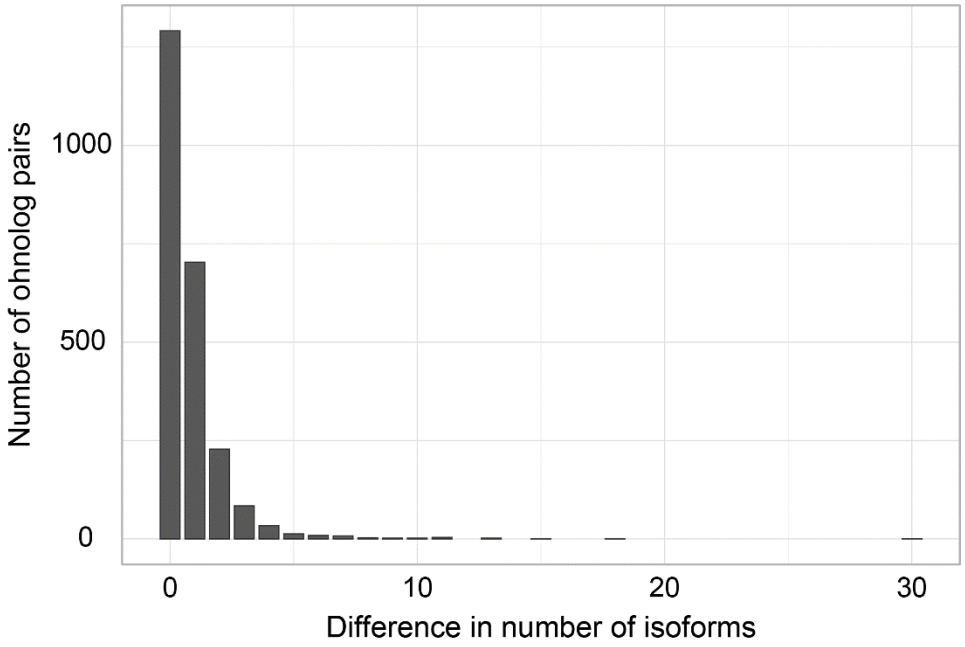

Figure 3.9: Difference in the number of isoforms between ohnologs.

Most ohnolog pairs had no difference in the number of isoforms per gene (Figure 3.9), with 54% having no difference, 29.5% of pairs having one gene with 1 more isoform, and 9.6% having one gene with 2 more isoforms. Only 6.9% of ohnolog pairs had a difference of 3 or more isoforms. Of the ohnolog pairs with no difference in isoform number, 83% had 1 isoform in each copy. Of the ohnolog pairs with differing numbers of isoforms, 71% had one gene with only 1 isoform, while the other copy had more than 1 isoforms. Figure 3.9 shows that, while most ohnologs have no difference in isoform number, 46% have at least one or more in the difference in isoform number.

3.3.2 Difference in tissue specificity

Because many alternatively spliced isoforms are important in maintaining specialized cells (Ule & Blencowe, 2019), a way to consider the function-sharing model (Figure 1.3) is to compare tissue-specific isoforms between ohnolog pairs. If the isoforms for each gene are specific to different tissues, it would suggest that the ohnologs have evolved to subdivide the ancestral isoforms in a tissue-specific manner.

We compared the tissue specificity of all ohnolog pairs with tissue-specific isoforms to determine whether they had the same tissue specificity. Figure 3.10 shows that 66% of ohnolog pairs have isoforms with the same tissue specificity. This indicates that function-sharing through the division of tissue-specific isoforms is not very common. The high number of ohnolog pairs with the same tissue-specificity of isoforms could potentially point to the conservation of tissue-regulated AS.

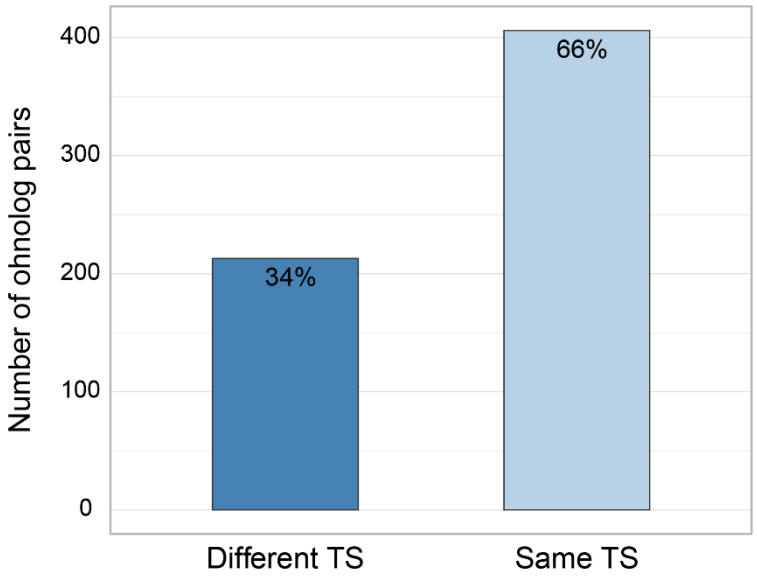

Figure 3.10: Difference in tissue specificity (TS). *Comparison between ohnologs with one or more tissue-specific isoforms in both genes. Ohnolog pairs with different TS had isoforms that were tissue-specific but specific to different tissues. Pairs with the same TS had isoforms that were specific to the same tissue in both genes.*

To look for examples of function-sharing through tissue-specific isoforms, we compared the expressed ohnolog pairs with two sets of orthologous genes in Northern pike. One set had 6 orthologs with both brain and muscle-specific exons, and the other set had 127 orthologous genes with brain-specific exons. We found 2 expressed ohnolog pairs from our samples with brain-muscle-specific orthologs in pike (Figure 3.12) and 18 expressed ohnolog pairs with

brain-specific orthologs in pike (Figure 3.13). If function-sharing through dividing tissue-specific isoforms between ohnolog pairs happens, we would expect to see one copy with brain-specific isoforms and the other with muscle-specific isoforms for the first set of orthologs (Figure 3.11 A). For the second set of orthologs, we expect to see brain-specific isoforms in one ohnolog copy and broadly expressed isoforms in the other copy (Figure 3.11 B).

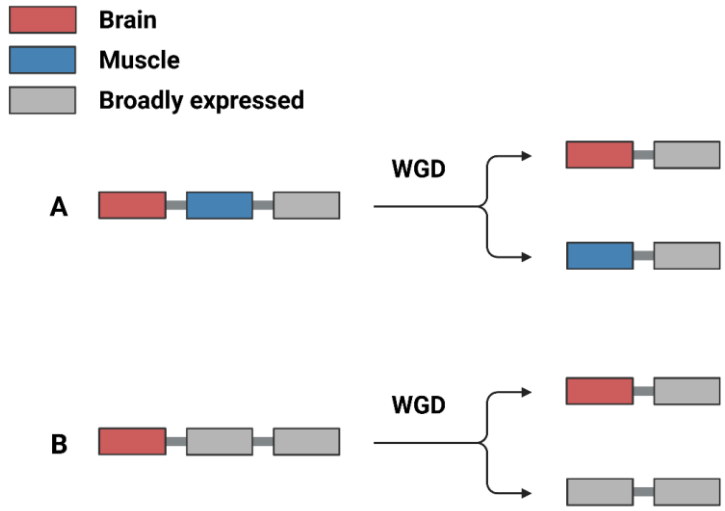

Figure 3.11: Expected division of tissue-specific isoforms after WGD under the function-sharing model. A) *The Northern pike orthologs have exons that are specific to brain (red) and muscle (blue). If AS evolves through function-sharing, the expected scenario after WGD is one ohnolog with brain-specific isoforms, and the other with muscle-specific isoforms.* ***B)*** *The Northern pike orthologs have one brain-specific isoform. We would expect one ohnolog to have brain-specific isoforms and the other to have broadly expressed isoforms.*

The isoforms of the first set of ohnolog pairs did not have the expected division of tissue-specific isoforms (i.e., one copy with brain-specific isoforms and the other with muscle-specific isoforms). One pair has one copy with both muscle and brain-specific isoforms, while the other copy has only brain-specific isoforms (Figure 3.12 A). The brain-specific isoform in the latter ohnolog was lowly expressed compared to the isoforms in its copy. The second expressed ohnolog pair had no isoforms with the expected ancestral tissue specificity (Figure 3.12 B).

The second set of ohnolog pairs showed some division of tissue specificity (Figure 3.13). However, this was only a minority of cases, as only 11% of the expressed ohnolog pairs from our samples follow the expected pattern of tissue-specific isoforms (Figure 3.11 B).

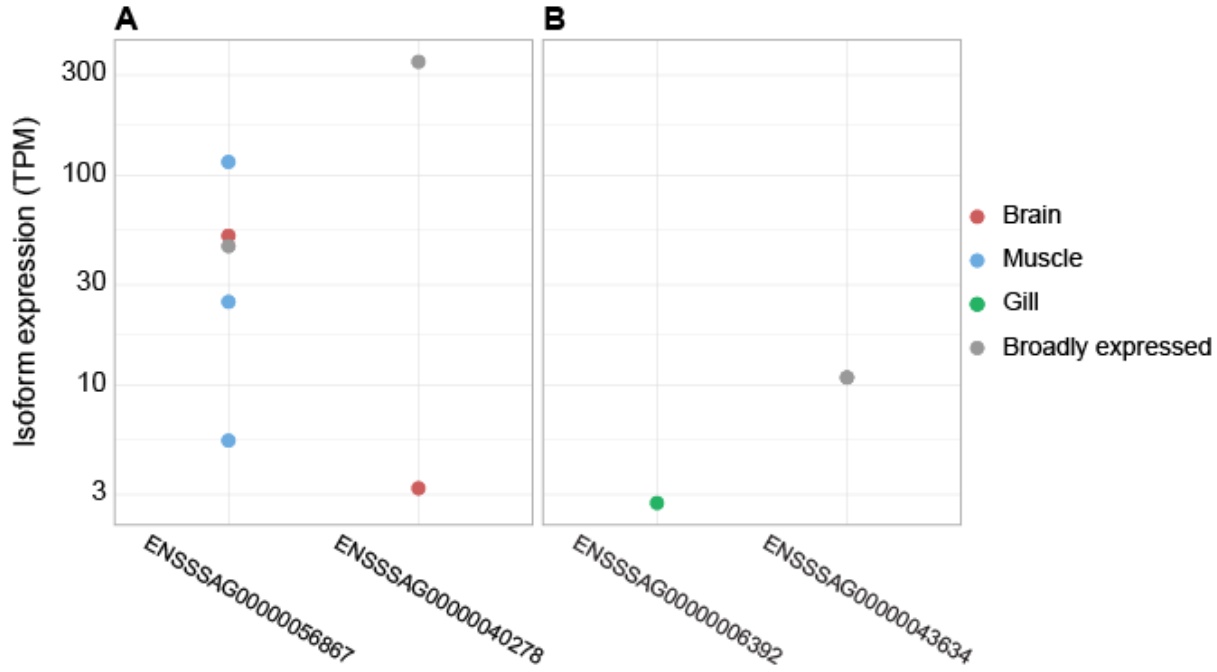

Figure 3.12: Expressed isoforms in ohnolog pairs with brain and muscle-specific exons pike orthologs. A) *One ohnolog has both brain and muscle-specific isoforms and a broadly expressed isoform, while the other has one brain-specific isoform and a broadly expressed isoform.* ***B)*** *One ohnolog has a gill-specific isoform, and the other has a broadly expressed isoform.*

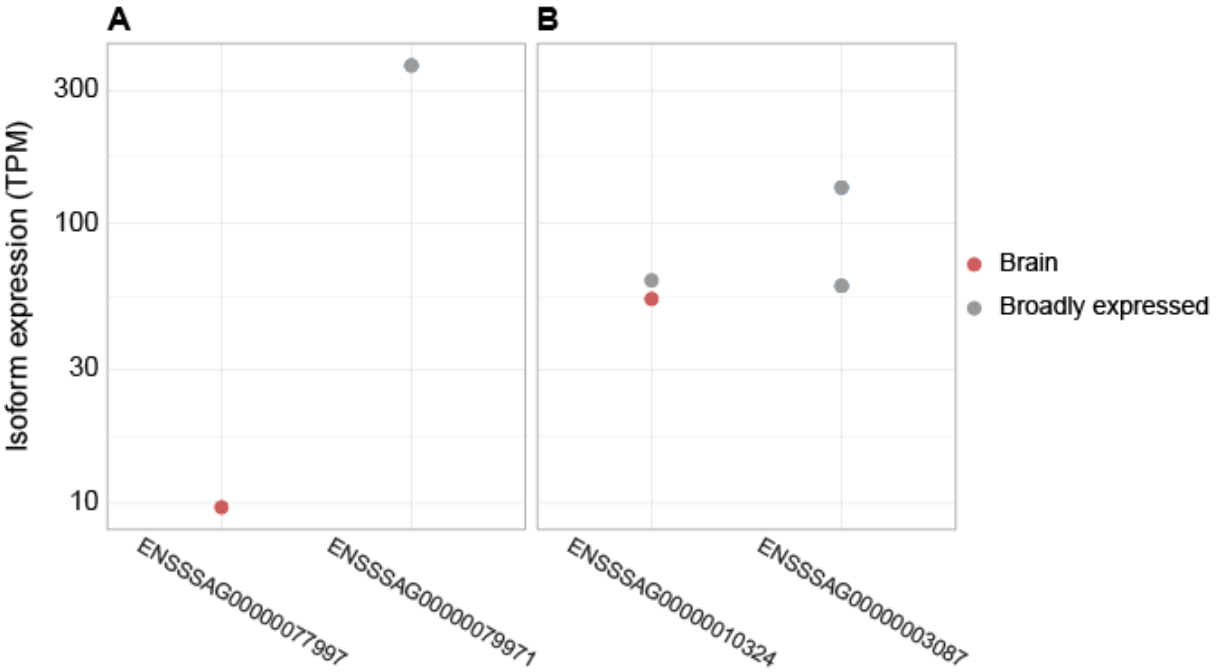

Figure 3.13: Expressed isoforms in ohnolog pairs with brain-specific exons in pike orthologs. *2 of 18 (11%) pairs adhere to the expectation of function-sharing, with both* ***A)*** *and* ***B)*** *having one ohnolog with brain-specific isoforms and the other with broadly expressed isoforms.*

4 Discussion

The two main objectives of this master thesis were to (i) better describe the diversity of splice isoforms and (ii) look for patterns of AS evolution following WGD in salmonids. First, this was done by using ONT long-read sequencing to capture entire isoforms, followed by bioinformatic analyses to compare the current transcriptome assembly with our data. Secondly, to test the predictions of the models of AS evolution following WGD, we looked at the diversity of isoforms originating from ohnolog and singleton genes and the differences in tissue-specific isoforms between ohnolog pairs. In the following discussion, we will highlight the importance of long-read sequencing in transcriptome assembly and the feasibility of the models of AS evolution.

4.1 Transcriptome assembly using full-length RNA sequencing

The availability of high-throughput, cost-efficient NSG platforms has led to a massive increase in transcriptome assemblies from short reads. *De novo* assemblies of the transcriptome do not rely on having an available reference genome, making it possible to assemble the transcriptomes of non-model organisms (Hölzer & Marz, 2019). There are some inherent challenges when reconstructing the transcriptome from short reads. Multi-exon genes can produce several splice isoforms that share exons, leading to reads being incorrectly assigned to isoforms, and lowly expressed isoforms can be discarded because they are mistakenly classified as sequencing errors (Haas & Zody, 2010).

The transcriptome annotation published by the International Cooperation to Sequence the Atlantic Salmon Genome (ICSASG) was assembled using *in silico* predictions and expressed sequence tags (ESTs) in combination with available short-read RNA sequences mapped to the reference genome (Cunningham et al., 2022). While these methods are useful for determining the exon-intron structure of genes, they cannot identify the full diversity of isoforms. Because of the Ss4R event, the Atlantic salmon genome contains many ohnolog genes. This complicates the transcriptome assembly when using short-read RNA seq data (Ramberg et al., 2021), as short reads mapped to the reference genome risk being misplaced on orthologs with high sequence similarity (Leong et al., 2010). In this thesis project, we used ONT long-read sequencing to obtain full-length transcripts for a more complete characterization of isoform diversity.

4.1.1 Long-read transcriptomics uncover novel isoforms

A striking result from this study is the large number of novel isoforms detected (Figure 3.2) compared to the previous Atlantic salmon gene annotation (GenBank accession ID: GCA_000233375.4), which was based on information from short-read RNAseq data. Our results on proportions of novel (75%) versus known (25%) isoforms are consistent with the recent findings from Ramberg and colleagues (2021). In this study, the authors sequenced full-length transcripts across 3 tissues (gill, head kidney, and liver) from Atlantic salmon using the sequencing platform PacBio. They found that 75% of the isoforms that mapped to the Atlantic salmon were novel. Similar studies in other non-salmonid species have also discovered large proportions of novel isoforms when using long-read RNA sequencing. For example, the long-read transcriptomes of the European rabbit (Chao et al., 2018) and ryegrass (Xie et al., 2020) revealed that 66% and 67% of the sequenced isoforms were previously not annotated, respectively. These results highlight the need for integrating long-read transcriptome sequencing as an integral part of genome annotation efforts. We also found that while each gene had fewer known isoforms (Figure 3.2), they were more highly expressed than the novel isoforms (Figure 3.5). This reflects that the lowly expressed isoforms can be challenging to annotate when using short-read RNA sequencing (Garber et al., 2011).

A large portion of the identified isoforms originated from regions of the genome that were not annotated (Figure 3.2). These isoforms were, however, more lowly expressed (Figure 3.5) and had an average of 1.1 isoforms per locus (Figure 3.4), as well as being considerably shorter (Figure 3.6) compared to isoforms from known genes. While some of these isoforms could originate from unannotated genes, genomes of higher eukaryotes produce an abundance of non-coding transcripts, many of which are transcribed at low levels (Mattick & Makunin, 2006). These non-coding RNAs can have biological roles in gene regulation, and have been linked to development, differentiation, and the regulation of immune responses in teleost fishes (M. Wang et al., 2018). This, combined transcription being a noisy process (Gordon et al., 2015), could account for the high number of isoforms that mapped to unannotated, intergenic regions (Figure 3.2).

4.1.2 Low number of predicted protein-coding isoforms

Although the number of annotated isoforms greatly increases by generating long-read transcriptome data into gene annotation (Figure 3.2), it is still unclear what proportion of these isoforms are actually functional. Based on analyses of translation frames, FLAIR predicted isoforms' functionality (i.e., protein-coding ability) and categorized 28 138 isoforms as protein-coding coming from 17 673 genes. In total, 39% of all isoforms were predicted to encode

complete proteins (Figure 3.7). Several factors could lead to this low estimate of protein-coding transcripts. First, it is likely that the FLAIR tool produces an underestimate of “functional” isoforms. For example, isoforms with premature termination codons could be functionally important, as alternatively spliced truncated isoforms have been shown to regulate gene expression through nonsense-mediated decay (Watabe et al., 2021). Secondly, the isoforms without start codons could also potentially be protein-coding, as Kearse and Wilusz (2017) found that non-AUG initiated translation is more common than previously thought. Third, many genes and isoforms have a tissue-specific or tissue-biased regulation (Ule & Blencowe, 2019), and by sampling only five tissues, we could be missing functional isoforms.

4.2 Evolution of alternative splicing in ohnologs

A recent study (Wang & Guo, 2021) explored the divergence of isoform usage between ohnologs in zebrafish, medaka, and stickleback. These three species are part of the teleost family, which underwent the third teleost-specific WGD event. They found a combined scenario of accelerated AS and function-sharing and that the predominant model of AS evolution was species-specific. A critical limitation of this study was its use of short-read RNAseq data, which is known to be challenging to use to estimate isoform diversity (Abdel-Ghany et al., 2016). Here, we used long reads to identify isoforms and to test the predictions of the evolutionary models of AS (Figure 1.3) following a WGD in Atlantic salmon.

4.2.1 Accelerated or asymmetric divergence of splice isoforms

Under the accelerated AS model (Figure 1.3), duplicated genes acquire novel splice variants in either one or both copies (Jin et al., 2008), leading to the expectation that ohnolog genes have more isoforms than singleton genes. In our study, we found no significant difference in the median ($p > 0.05$) number of isoforms between ohnolog and singleton genes (Figure 3.8), not in support of the accelerated AS model of evolution.

Asymmetric evolution of splice variants, where one ohnolog has more isoforms than the other, is another possible outcome of AS evolution predicted by the accelerated AS model (Jin et al., 2008). Ohnolog pairs with asymmetric numbers of isoforms have been associated with functional divergence of gene regulation following several ancient WGDs in vertebrates (Wang & Guo, 2021). In their study of AS evolution after Ts3R, Wang and Guo (2021) found between 2.8-6.1% ohnolog pairs with significant differences in isoform numbers. They found that these asymmetric losses or gains in isoform numbers among the ohnologs were significantly linked with functional activities, particularly in neural tissues (Wang & Guo, 2021).

Of the ohnolog pairs we found, 6.9% had differences in isoform numbers that were 3 or higher (Figure 3.9), which could be putative cases of functional divergence through the gain of isoforms. Obtaining matching data from a closely related, non-salmonid species would better quantify potential asymmetric losses or gains of isoforms in Ss4R ohnologs. GO enrichment analyses could then be performed to test the prediction of functional divergence through asymmetric AS evolution.

4.2.2 Is function-sharing supported by our data?

A much-cited model of AS evolution following gene duplication is the function-sharing model (Figure 1.3), where the ancestral isoforms are subdivided into each copy, reducing the number of isoforms, or possibly partitioning tissue-specific exon regulation to one of the ohnolog copies. We did not observe any significant difference in isoform numbers between Atlantic salmon ohnologs and singletons, which does not support the function-sharing hypothesis (Figure 3.9). However, retained duplicated genes from the Ss4R are not random in terms of function (Lien et al., 2016), which can confound our analyses and interpretation. Therefore, a better approach would be to infer ancestral isoform diversity from orthologs in a non-salmonid species and compare this to present isoform diversity in ohnologs.

Regarding the evolution of tissue-specific isoforms, the majority (66%) of ohnolog pairs have the same isoform tissue specificity (Figure 3.10). This indicates that regulation of tissue-specific splicing is mostly conserved between ohnolog pairs. Moreover, ohnologs that have a likely tissue-specific exon regulation rarely (11%) showed a tendency to subdivide tissue-specific isoforms between the ohnolog pairs (Figure 3.13), or not at all (Figure 3.12). Even though several papers repost patterns of function-sharing (Abascal et al., 2015; Su & Gu, 2012; Wang & Guo, 2021), we find little evidence to support this model as a major evolutionary route for AS.

4.2.3 Independent evolution of alternative splicing

For ohnolog pairs retained in the Atlantic salmon genome, the overall trend has been sown to be adaptive regulatory evolution. Most ohnolog pairs have evolved asymmetrically, with one copy being downregulated, possibly leading to pseudogenization (Gillard et al., 2021). The relaxed purifying selection pressure needed for accumulating several new isoforms (Jin et al., 2008; Roux & Robinson-Rechavi, 2011) has led one copy down the path of pseudogenization rather than acquiring new splice isoforms. This could explain why we found no evidence to support widespread accelerated AS (Figure 3.8).

Mutations in splice sites or within the exonic or intronic regions lead to the loss of splice variants (Abramowicz & Gos, 2018). For reciprocal loss of isoforms to occur, as predicted by the function-sharing model, these mutations must happen in both copies. This scenario appears more unlikely for ohnologs with high sequence evolution constraints, such as the ohnolog pairs with symmetrical downshifts in regulation found by Gillard and colleagues (Gillard et al., 2021), and could also account for the seemingly conserved tissue-specificity of isoforms between ohnolog pairs (Figure 3.10).

Contrary to previous observations of function-sharing and accelerated AS (Jin et al., 2008; Su & Gu, 2012; Wang & Guo, 2021), we found no strong relationship between the models of AS evolution and retained ohnologs in the Atlantic salmon genome based on our data. This could be due to several factors, one of which is the limitations in the methods used in this thesis, as we mainly considered the number of isoforms and tissue specificity in our analyses. Additionally, we did not have access to a long-read transcriptome in a non-salmonid species to use as a proxy for the ancestral state. Another consideration is that AS evolution could be more prevalent after SGDs. Genes duplicated in a WGD event in pear have been shown to generally evolve more slowly than duplicates arising from SGDs, based on the ratio of non-synonymous and synonymous substitution rates (Qiao et al., 2018). Ohnologs retained in older WGDs, such as the 1R/2R and Ts3R events, could therefore be more prone to AS evolution than ohnologs from the more recent Ss4R.

5 Concluding remarks and further perspectives

In this thesis project, we used ONT long-read sequencing to obtain full-length transcripts, improve isoform annotation, and better characterize isoform diversity across five tissues. Our findings highlight the need for including long-read sequencing in transcriptome assemblies, as short-read data do not easily characterize isoform diversity. To better assess the models of AS evolution after Ss4R, the next step would be to assemble a long-read transcriptome in a non-salmonid species, such as the Northern pike, to infer ancestral isoform diversity and function from ortholog genes. Rather than looking for global patterns, we could examine AS evolution on a gene-by-gene basis by comparing isoform diversity in Atlantic salmon to Northern pike orthologs.

References

Abascal, F., Tress, M. L., & Valencia, A. (2015). The Evolutionary Fate of Alternatively Spliced Homologous Exons after Gene Duplication. *Genome Biology and Evolution*, *7*(6), 1392–1403. https://doi.org/10.1093/gbe/evv076

Abdel-Ghany, S. E., Hamilton, M., Jacobi, J. L., Ngam, P., Devitt, N., Schilkey, F., Ben-Hur, A., & Reddy, A. S. N. (2016). A survey of the sorghum transcriptome using single-molecule long reads. *Nature Communications*, *7*(1), 11706. https://doi.org/10.1038/ncomms11706

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, *21*(1), 30. https://doi.org/10.1186/s13059-020-1935-5

Abramowicz, A., & Gos, M. (2018). Splicing mutations in human genetic disorders: examples, detection, and confirmation. *Journal of Applied Genetics*, *59*(3), 253–268. https://doi.org/10.1007/s13353-018-0444-7

Baralle, F. E., & Giudice, J. (2017). Alternative splicing as a regulator of development and tissue identity. *Nature Reviews Molecular Cell Biology*, *18*(7), 437–451. https://doi.org/10.1038/nrm.2017.27

Behjati, S., & Tarpey, P. S. (2013). What is next generation sequencing? *Archives of Disease in Childhood - Education & Practice Edition*, *98*(6), 236–238. https://doi.org/10.1136/archdischild-2013-304340

Berget, S. M., Moore, C., & Sharp, P. A. (1977). Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences*, *74*(8), 3171–3175. https://doi.org/10.1073/pnas.74.8.3171

Berthelot, C., Brunet, F., Chalopin, D., Juanchich, A., Bernard, M., Noël, B., Bento, P., da Silva, C., Labadie, K., Alberti, A., Aury, J.-M., Louis, A., Dehais, P., Bardou, P., Montfort, J., Klopp, C., Cabau, C., Gaspin, C., Thorgaard, G. H., … Guiguen, Y. (2014). The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nature Communications*, *5*(1), 3657. https://doi.org/10.1038/ncomms4657

Bertolotti, A. C., Layer, R. M., Gundappa, M. K., Gallagher, M. D., Pehlivanoglu, E., Nome, T., Robledo, D., Kent, M. P., Røsæg, L. L., Holen, M. M., Mulugeta, T. D., Ashton, T. J., Hindar, K., Sægrov, H., Florø-Larsen, B., Erkinaro, J., Primmer, C. R., Bernatchez, L., Martin, S. A. M., … Macqueen, D. J. (2020). The structural variation landscape in 492 Atlantic salmon genomes. *Nature Communications*, *11*(1), 5176. https://doi.org/10.1038/s41467-020-18972-x

Besser, J., Carleton, H. A., Gerner-Smidt, P., Lindsey, R. L., & Trees, E. (2018). Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clinical Microbiology and Infection*, *24*(4), 335–341. https://doi.org/10.1016/j.cmi.2017.10.013

Burset, M. (2000). Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Research*, *28*(21), 4364–4375. https://doi.org/10.1093/nar/28.21.4364

Chao, Y., Yuan, J., Li, S., Jia, S., Han, L., & Xu, L. (2018). Analysis of transcripts and splice isoforms in red clover (Trifolium pratense L.) by single-molecule long-read sequencing. *BMC Plant Biology*, *18*(1), 300. https://doi.org/10.1186/s12870-018-1534-8

Conant, G. C., & Wolfe, K. H. (2008). Turning a hobby into a job: How duplicated genes find new functions. *Nature Reviews Genetics*, *9*(12), 938–950. https://doi.org/10.1038/nrg2482

Cunningham, F., Allen, J. E., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Austine-Orimoloye, O., Azov, A. G., Barnes, I., Bennett, R., Berry, A., Bhai, J., Bignell, A., Billis, K., Boddu, S., Brooks, L., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., … Flicek, P. (2022). Ensembl 2022. *Nucleic Acids Research*, *50*(D1), D988–D995. https://doi.org/10.1093/nar/gkab1049

Deamer, D., Akeson, M., & Branton, D. (2016). Three decades of nanopore sequencing. *Nature Biotechnology*, *34*(5). https://doi.org/10.1038/nbt.3423

Dehal, P., & Boore, J. L. (2005). Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. *PLoS Biology*, *3*(10), e314. https://doi.org/10.1371/journal.pbio.0030314

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y., & Postlethwait, J. (1999). Preservation of Duplicate Genes by Complementary, Degenerative Mutations. *Genetics*, *151*(4), 1531–1545. https://doi.org/10.1093/genetics/151.4.1531

Galbi, Y. (2019). *tispec: Calculates Tissue Specificity from RNA-seq Data*. GitHub Repository. https://github.com/roonysgalbi/tispec

Garber, M., Grabherr, M. G., Guttman, M., & Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods*, *8*(6), 469–477. https://doi.org/10.1038/nmeth.1613

Gillard, G. B., Grønvold, L., Røsæg, L. L., Holen, M. M., Monsen, Ø., Koop, B. F., Rondeau, E. B., Gundappa, M. K., Mendoza, J., Macqueen, D. J., Rohlfs, R. v., Sandve, S. R., & Hvidsten, T. R. (2021). Comparative regulomics supports pervasive selection on gene dosage following whole genome duplication. *Genome Biology*, *22*(1), 103. https://doi.org/10.1186/s13059-021-02323-0

Gordon, A. J., Satory, D., Halliday, J. A., & Herman, C. (2015). Lost in transcription: transient errors in information transfer. *Current Opinion in Microbiology*, *24*, 80–87. https://doi.org/10.1016/j.mib.2015.01.010

Haas, B. J., & Zody, M. C. (2010). Advancing RNA-Seq analysis. *Nature Biotechnology*, *28*(5), 421–423. https://doi.org/10.1038/nbt0510-421

Hamid, F. M., & Makeyev, E. V. (2014). Emerging functions of alternative splicing coupled with nonsense-mediated decay. *Biochemical Society Transactions*, *42*(4), 1168–1173. https://doi.org/10.1042/BST20140066

Harper, J. W., & Bennett, E. J. (2016). Proteome complexity and the forces that drive proteome imbalance. *Nature*, *537*(7620), 328–338. https://doi.org/10.1038/nature19947

Hittinger, C. T., & Carroll, S. B. (2007). Gene duplication and the adaptive evolution of a classic genetic switch. *Nature*, *449*(7163), 677–681. https://doi.org/10.1038/nature06151

Hölzer, M., & Marz, M. (2019). De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. *GigaScience*, *8*(5). https://doi.org/10.1093/gigascience/giz039

Hsu, S.-N., & Hertel, K. J. (2009). Spliceosomes walk the line: Splicing errors and their impact on cellular function. *RNA Biology*, *6*(5), 526–530. https://doi.org/10.4161/rna.6.5.9860

Iñiguez, L. P., & Hernández, G. (2017). The Evolutionary Relationship between Alternative Splicing and Gene Duplication. *Frontiers in Genetics*, *08*. https://doi.org/10.3389/fgene.2017.00014

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., Fiddes, I. T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H. E., Pedersen, B. S., Rhie, A., Richardson, H., Quinlan, A. R., … Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, *36*(4), 338–345. https://doi.org/10.1038/nbt.4060

Jin, L., Kryukov, K., Clemente, J. C., Komiyama, T., Suzuki, Y., Imanishi, T., Ikeo, K., & Gojobori, T. (2008). The evolutionary relationship between gene duplication and alternative splicing. *Gene*, *427*(1–2), 19–31. https://doi.org/10.1016/j.gene.2008.09.002

Kanzi, A. M., San, J. E., Chimukangara, B., Wilkinson, E., Fish, M., Ramsuran, V., & de Oliveira, T. (2020). Next Generation Sequencing and Bioinformatics Analysis of Family Genetic Inheritance. *Frontiers in Genetics*, *11*. https://doi.org/10.3389/fgene.2020.544162

Kassahn, K. S., Dang, V. T., Wilkins, S. J., Perkins, A. C., & Ragan, M. A. (2009). Evolution of gene function and regulatory control after whole-genome duplication: Comparative analyses in vertebrates. *Genome Research*, *19*(8), 1404–1418. https://doi.org/10.1101/gr.086827.108

Kchouk, M., Gibrat, J. F., & Elloumi, M. (2017). Generations of Sequencing Technologies: From First to Next Generation. *Biology and Medicine*, *09*(03). https://doi.org/10.4172/0974-8369.1000395

Kearse, M. G., & Wilusz, J. E. (2017). Non-AUG translation: a new start for protein synthesis in eukaryotes. *Genes & Development*, *31*(17), 1717–1731. https://doi.org/10.1101/gad.305250.117

Kolde, R. (2019). *pheatmaps: Pretty Heatmaps*. R Package Version 1.0.12. https://CRAN.R-project.org/package=pheatmap

Kopelman, N. M., Lancet, D., & Yanai, I. (2005). Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. *Nature Genetics*, *37*(6), 588–589. https://doi.org/10.1038/ng1575

Kryuchkova-Mostacci, N., & Robinson-Rechavi, M. (2016). A benchmark of gene expression tissue-specificity metrics. *Briefings in Bioinformatics*, bbw008. https://doi.org/10.1093/bib/bbw008

Leong, J. S., Jantzen, S. G., von Schalburg, K. R., Cooper, G. A., Messmer, A. M., Liao, N. Y., Munro, S., Moore, R., Holt, R. A., Jones, S. J., Davidson, W. S., & Koop, B. F. (2010). Salmo salar and Esox lucius full-length cDNA sequences reveal changes in evolutionary pressures on a post-tetraploidization genome. *BMC Genomics*, *11*(1), 279. https://doi.org/10.1186/1471-2164-11-279

Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Kent, M. P., Nome, T., Hvidsten, T. R., Leong, J. S., Minkley, D. R., Zimin, A., Grammes, F., Grove, H., Gjuvsland, A., Walenz, B., Hermansen, R. A., von Schalburg, K., Rondeau, E. B., di Genova, A., Samy, J. K. A., … Davidson, W. S. (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature*, *533*(7602), 200–205. https://doi.org/10.1038/nature17164

Macqueen, D. J., & Johnston, I. A. (2014). A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1778), 20132881. https://doi.org/10.1098/rspb.2013.2881

Makałowski, W. (2001). Are We Polyploids? A Brief History of One Hypothesis. *Genome Research*, *11*(5), 667–670. https://doi.org/10.1101/gr.188801

Mattick, J. S., & Makunin, I. v. (2006). Non-coding RNA. *Human Molecular Genetics*, *15*(suppl_1), R17–R29. https://doi.org/10.1093/hmg/ddl046

McGrath, C. L., Gout, J.-F., Johri, P., Doak, T. G., & Lynch, M. (2014). Differential retention and divergent resolution of duplicate genes following whole-genome duplication. *Genome Research*, *24*(10), 1665–1675. https://doi.org/10.1101/gr.173740.114

Meyer, A., & van de Peer, Y. (2005). From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *BioEssays*, *27*(9), 937–945. https://doi.org/10.1002/bies.20293

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

NHGRI. (1990). *Understanding our genetic inheritance: the US Human Genome Project, the first five years 1990*. https://www.genome.gov/10001477/human-genome-projects-fiveyear-plan-19911995/

NHGRI. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*(7011), 931–945. https://doi.org/10.1038/nature03001

Oxford Nanopore Technologies. (2019). *PCR-cDNA Barcoding (SQK-PCB109)*.

Oxford Nanopore Technologies. (2022a). *Company History*. https://nanoporetech.com/about-us/history

Oxford Nanopore Technologies. (2022b). *How nanopore sequencing works*. https://nanoporetech.com/how-it-works

Oxford Nanopore Technologies. (2022c). *RNA and gene expression analysis using direct RNA and cDNA sequencing*. https://nanoporetech.com/applications/rna-sequencing

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., & Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, *40*(12), 1413–1415. https://doi.org/10.1038/ng.259

Parada, G. E., Munita, R., Cerda, C. A., & Gysling, K. (2014). A comprehensive survey of non-canonical splice sites in the human transcriptome. *Nucleic Acids Research*, *42*(16), 10564–10578. https://doi.org/10.1093/nar/gku744

Qiao, X., Yin, H., Li, L., Wang, R., Wu, J., Wu, J., & Zhang, S. (2018). Different Modes of Gene Duplication Show Divergent Evolutionary Patterns and Contribute Differently to the Expansion of Gene Families Involved in Important Fruit Traits in Pear (Pyrus bretschneideri). *Frontiers in Plant Science*, *9*. https://doi.org/10.3389/fpls.2018.00161

Ramberg, S., Høyheim, B., Østbye, T.-K. K., & Andreassen, R. (2021). A de novo Full-Length mRNA Transcriptome Generated From Hybrid-Corrected PacBio Long-Reads Improves the Transcript Annotation and Identifies Thousands of Novel Splice Variants in Atlantic Salmon. *Frontiers in Genetics*, *12*. https://doi.org/10.3389/fgene.2021.656334

Rang, F. J., Kloosterman, W. P., & de Ridder, J. (2018). From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biology*, *19*(1), 90. https://doi.org/10.1186/s13059-018-1462-9

Roux, J., & Robinson-Rechavi, M. (2011). Age-dependent gain of alternative splice forms and biased duplication explain the relation between splicing and duplication. *Genome Research*, *21*(3), 357–363. https://doi.org/10.1101/gr.113803.110

RStudio Team. (2022). *RStudio: Integrated Development Environment for R*. http://www.rstudio.com/

Sandve, S. R., Rohlfs, R. v., & Hvidsten, T. R. (2018). Subfunctionalization versus neofunctionalization after whole-genome duplication. *Nature Genetics*, *50*(7), 908–909. https://doi.org/10.1038/s41588-018-0162-4

Schubert, I., & Lysak, M. A. (2011). Interpretation of karyotype evolution should consider chromosome structural constraints. *Trends in Genetics*, *27*(6), 207–216. https://doi.org/10.1016/j.tig.2011.03.004

Sémon, M., & Wolfe, K. H. (2007a). Reciprocal gene loss between Tetraodon and zebrafish after whole genome duplication in their ancestor. *Trends in Genetics*, *23*(3), 108–112. https://doi.org/10.1016/j.tig.2007.01.003

Sémon, M., & Wolfe, K. H. (2007b). Consequences of genome duplication. *Current Opinion in Genetics & Development*, *17*(6), 505–512. https://doi.org/10.1016/j.gde.2007.09.007

Sémon, M., & Wolfe, K. H. (2008). Preferential subfunctionalization of slow-evolving genes after allopolyploidization in Xenopus laevis. *Proceedings of the National Academy of Sciences*, *105*(24), 8333–8338. https://doi.org/10.1073/pnas.0708705105

Shi, M., Zhang, H., Wang, L., Zhu, C., Sheng, K., Du, Y., Wang, K., Dias, A., Chen, S., Whitman, M., Wang, E., Reed, R., & Cheng, H. (2015). Premature termination codons are recognized in the nucleus in a reading-frame-dependent manner. *Cell Discovery*, *1*(1), 15001. https://doi.org/10.1038/celldisc.2015.1

Spoelhof, J. P., Soltis, P. S., & Soltis, D. E. (2017). Pure polyploidy: Closing the gaps in autopolyploid research. *Journal of Systematics and Evolution*, *55*(4), 340–352. https://doi.org/10.1111/jse.12253

Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D., Thanaraj, T. A., & Soreq, H. (2005). Function of alternative splicing. *Gene*, *344*, 1–20. https://doi.org/10.1016/j.gene.2004.10.022

Su, Z., & Gu, X. (2012). Revisit on the evolutionary relationship between alternative splicing and gene duplication. *Gene*, *504*(1). https://doi.org/10.1016/j.gene.2012.05.012

Su, Z., Wang, J., Yu, J., Huang, X., & Gu, X. (2006). Evolution of alternative splicing after gene duplication. *Genome Research*, *16*(2), 182–189. https://doi.org/10.1101/gr.4197006

Tang, A. D., Soulette, C. M., van Baren, M. J., Hart, K., Hrabeta-Robinson, E., Wu, C. J., & Brooks, A. N. (2020). Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-15171-6

Ule, J., & Blencowe, B. J. (2019). Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution. *Molecular Cell*, *76*(2). https://doi.org/10.1016/j.molcel.2019.09.017

Wang, M., Jiang, S., Wu, W., Yu, F., Chang, W., Li, P., & Wang, K. (2018). Non-coding RNAs Function as Immune Regulators in Teleost Fish. *Frontiers in Immunology*, *9*. https://doi.org/10.3389/fimmu.2018.02801

Wang, Y., & Guo, B. (2021). The divergence of alternative splicing between ohnologs in teleost fishes. *BMC Ecology and Evolution*, *21*(1), 98. https://doi.org/10.1186/s12862-021-01833-6

Watabe, E., Togo-Ohno, M., Ishigami, Y., Wani, S., Hirota, K., Kimura-Asami, M., Hasan, S., Takei, S., Fukamizu, A., Suzuki, Y., Suzuki, T., & Kuroyanagi, H. (2021). A-mediated alternative splicing coupled with nonsense-mediated mRNA decay regulates SAM synthetase homeostasis. *The EMBO Journal*, *40*(14). https://doi.org/10.15252/embj.2020106434

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.-S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., … Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, *37*(10), 1155–1162. https://doi.org/10.1038/s41587-019-0217-9

Westergren Jakobsson, A., Segerman, B., Wallerman, O., Bergström Lind, S., Zhao, H., Rubin, C.-J., Pettersson, U., & Akusjärvi, G. (2021). The Human Adenovirus 2 Transcriptome: an Amazing Complexity of Alternatively Spliced mRNAs. *Journal of Virology*, *95*(4). https://doi.org/10.1128/JVI.01869-20

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analyisis*. Springer-Verlag New York.

Xie, L., Teng, K., Tan, P., Chao, Y., Li, Y., Guo, W., & Han, L. (2020). PacBio single-molecule long-read sequencing shed new light on the transcripts and splice isoforms of the perennial ryegrass. *Molecular Genetics and Genomics*, *295*(2), 475–489. https://doi.org/10.1007/s00438-019-01635-y

Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A., Horn-Saban, S., Safran, M., Domany, E., Lancet, D., & Shmueli, O. (2005). Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, *21*(5), 650–659. https://doi.org/10.1093/bioinformatics/bti042

Zhao, Y., Li, M.-C., Konaté, M. M., Chen, L., Das, B., Karlovich, C., Williams, P. M., Evrard, Y. A., Doroshow, J. H., & McShane, L. M. (2021). TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. *Journal of Translational Medicine*, *19*(1), 269. https://doi.org/10.1186/s12967-021-02936-w

Appendix

Library preparation

Appendix 1: *Consumables used in library preparation.*

Consumable	**Manufacturer**
Agencourt AMPure XP Beads	*Beckman Coulter*
LongAmp Taq 2X Master Mix	*New England Biolabs*
Maxima Minus H Reverse Transcriptase (200 U/µl)	*ThermoFisher Scientific*
5x RT buffer	*ThermoFisher Scientific*
RNase OUT™ (40 U/µl)	*New England Biolabs*
10 mM dNTP solution	*New England Biolabs*
Exonuclease I	*New England Biolabs*

Appendix 2: *Contents of PCR-cDNA barcoding sequencing kit (SQK-PCB109).*

Name	**Acronym**
VN Primer	VNP
Strand Switching Primer	SSP
Rapid Adapter	RAP
Sequencing Buffer	SQB
Loading Beads	LB
Elution Buffer	EB
Barcode Primers 1-12	BP01-BP12

Appendix 3: *Contents of flow cell priming kit (EXP-FLP002).*

Name	**Acronym**
Flush Buffer	FB
Flush Tether	FT

Data availability

All BASH-scripts and data used in data analysis are available at https://gitlab.com/RonjaSan/alternative_splicing_salmo_salar.

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway