THE UNIVERSITY *of* EDINBURGH

To Cut a Short Story Long: Development of Full-Length RNA-Sequencing Approaches to Resolve Transcript-Level Expression Dynamics in Atlantic Salmon

THE UNIVERSITY *of* EDINBURGH

A thesis presented for the degree of Doctor of Philosophy
at the University of Edinburgh

2024

Oliver T.H. Eve

BSc. (Hons) – University of Aberdeen

Declaration

I hereby declare that this Thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a higher degree. The work presented was designed and conducted between October 2020 and September 2024 and is my own with the following exceptions.

Dr Perojil-Morata (Imperial College London; formerly University of Edinburgh) collected Atlantic salmon eggs, milt and then conducted ex-vivo artificial inseminations and subsequent embryonic sampling during the AQUA-FAANG project for the purpose of DevMap analyses. I extracted total RNA from these samples and they were used for the Nanopore RNA-seq detailed in Chapter 2.

Dr Shahmir Naseer (University of Aberdeen), Dr Thomas Clark (Roslin, Institute, University of Edinburgh; formerly University of Aberdeen) and Prof. Samuel Martin (University of Aberdeen) conducted the in vivo immunostimulation of Atlantic salmon smolts and head kidney sample harvesting at the Scottish Fish Immunology Research Centre during the AQUA-FAANG project for the purpose of ImmunoMap analyses. Dr Naseer produced the total RNA samples from these samples, which was used for the Nanopore RNA-seq detailed in Chapter 2.

Dr Manu Gundappa (Wageningen University and Research; formerly University of Edinburgh) produced the annotations used for classifying the salmonid ohnologues investigated in Chapter 2.

Whilst I prepared all Nanopore RNA-seq libraries, the actual sequencing was performed by Edinburgh Genomics on their PromethION device, except for early trial runs I performed on the MinION platform, which were not included in the Thesis.

Oliver Eve September 2024

Acknowledgements

They say it takes a village to raise a child and the same rings true for my PhD project, which would not have been possible without the support and encouragement of many individuals, both professionally and personally.

Firstly, I would like to thank my PhD supervisor, Prof. Daniel Macqueen, for the never-ending support and mentorship I have received throughout the entire project. I am extremely grateful for the opportunity he has provided and I would not have developed into the scientist I am today without his encouragement and guidance. Thank you for being there every step of the way.

I must thank Manu Gundappa for his unwavering friendship, kindness and commitment to solving every bioinformatic problem under the sun. What started off as a professional relationship ended as a deeply personal friendship which I will value forever. I miss our lunch chats and hope to get over to the Netherlands soon for a visit!

On a more professional level, I would like to thank the members of the AQUA-FAANG consortium for their assistance in completing this project for without them I would have had no samples! In particular, I am grateful to Prof. Sam Martin and Dr Shahmir Naseer for their patience when explaining how the immune challenges and sample collection detailed in Chapter 3 were carried out.

The connections, friendships and memories I've made at the Roslin Institute will be treasured forever and showed me that a PhD can be a (mostly) enjoyable experience. From lunch banter provided by the wider members of the Roslin aquaculture squad; Thomas, Ambre, Sarah, Clémence, Tim^2, Rob and Nick, to merry evenings at after-work social events, I've had a great time which I attribute to the friendship and support I've received from every one of you, however large or small.

The latter stages of this PhD have ranked amongst the most stressful weeks of my life and I reserve a special thanks for Sarah and Clémence for their

words of encouragement, and chocolatey, cakey, delectable sucrose boosts, during this time.

To Hannah, Laura and Cara, thank you for being my friends and sharing in the trials and tribulations that make up a PhD. Hannah, it's nearly time to cash in those early morning penalty chips!

There are many more people who have supported me during this process but I cannot name them all here and would hate to miss anyone out. As such, I'd like to extend my thanks to anyone that has been there for me in the last four years. I greatly appreciate all the impacts you've had on me.

Finally, I would like to thank my family for being an ever-present, never-ending fountain of encouragement and advice. My Mum and Dad have constantly been there for me, never once doubting my ability to get this over the line and I am deeply grateful for their support. In the inimitable words of a 5-year-old Oliver, "to inferbinity RBO!".

Table of Contents

List of Figures

List of Tables

Abstract

Atlantic salmon is a finfish of significant cultural, ecological and commercial importance, representing the United Kingdom's main aquaculture species. There is currently a great opportunity to apply genomics to improve the sustainability, efficiency and welfare of the aquaculture sector. This includes a current drive to perform functional annotation of genomes to identify genes and other elements that shape the traits of aquaculture species. The AQUA-FAANG consortium aimed to produce comprehensive functional annotations for six key aquaculture species, including Atlantic salmon. The work in this thesis was carried out under the AQUA-FAANG umbrella.

Long-read RNA-sequencing (RNA-seq) technologies are powerful tools for functional annotation of gene expression, with great scope to resolve complex transcript variants that cannot be accurately assessed using traditional short-read methods. However, long-read RNA-seq is yet to be applied and benchmarked in many aquaculture species. As such, my work aimed to develop a robust full-length RNA-seq method in Atlantic salmon using long-read technology to examine transcriptional diversity and conduct expression analyses resolved to individual transcript variants. My work focused on two distinct study systems where extensive transcriptional regulation is applied: 1) embryogenesis, the stage of ontogeny where the adult body plan is established, and 2) immune function in response to acute viral and bacterial stimulation, improving understanding of innate immune function.

I developed a full-length RNA-seq method using the Oxford Nanopore Technologies platform involving the optimisation of total RNA extraction and mRNA isolation protocols, as well as cDNA library generation and subsequent sequencing on the PromethION device. A custom transcriptome assembly pipeline was optimised to generate the first nanopore-based long-read transcriptome for Atlantic salmon, used as the reference for further analyses reported in this Thesis. The long-read transcriptome consisted of 266,222 transcripts and 35,480 genes, with a transcript-to-gene ratio of 7.50 in comparison with 2.65 in the current Ensembl reference annotation

(Ssal_v3.1). Furthermore, 60% of transcript models were deemed to contain a novel splice site, indicating that my full-length RNA-seq method captured extensive novel transcript diversity not annotated in the current reference assembly.

To examine transcript expression dynamics in response to viral and bacterial infection, I developed a differential transcript expression and usage analysis workflow, adapting existing bioinformatic tools. My analysis captured complex dynamics of alternative transcript expression for antiviral and antibacterial genes involved in the interferon-JAK/STAT pathway and proinflammatory responses, respectively. A novel fusion transcript between *pctk2* and an undescribed locus containing a FIP2-like coding sequence was identified to be upregulated in both viral and bacterial response.

A separate pipeline was developed to assess transcript expression during embryogenesis using a complex timecourse design that sampled embryos at six stages (from blastulation to the late-eyed stage). Using a dimensionality reduction technique called self-organising maps (SOM), twinned with a generalised linear model and quasi-likelihood F-test method, I optimised a differential transcript expression workflow and developed an approach to examine differential transcript usage across development stages. This resulted in a comprehensive description of transcript expression throughout early development and the discovery of alternative transcript usage events within individual genes including an exon-chaining event in the coding sequence of *slc25a3b,* (mitochondrial phosphate carrier PiC) causing expression of unique isoforms in blastulation and late-eyed stages of development, whilst a 5’ UTR difference in the *tagl* gene led to different isoforms being expressed in blastulation and somitogenesis.

The full-length sequencing method captured many mono-exonic, or intronless gene and transcript models not present in the reference annotation. Over a third of these models were found to contain a complete or partial ORF indicating they are protein-coding, whilst approximately 25% of mono-exonic transcripts were found to overlap repetitive regions.

Additionally, I identified a previously undescribed retrogene family found to be widespread throughout the genome.

Overall, this thesis reports approaches for robust full-length RNA-seq analysis in a non-model species with a complex genome. This work has furthered our understanding of the transcript-level expression dynamics underpinning early development and immune function in Atlantic salmon, with possible applications in aquaculture research.

Lay Summary

All the information required by an organism to produce the diverse proteins needed to maintain life is stored in a molecule called deoxyribonucleic acid, or simply, DNA. DNA consists of a continuous string of units, called nucleotide bases, which are arranged in a continuous double strand and represented commonly by four letters; A, T, C and G. DNA itself cannot be immediately converted to a usable product; it is instead copied into an intermediate molecule called ribonucleic acid (RNA). Like DNA, RNA is made up of four nucleotide bases, this time A, U (in replacement of T), C and G, but unlike DNA, RNA molecules are single-stranded.

The parts of the DNA that are copied into RNA are called genes. We know that a single gene can code for multiple unique RNA molecules known as transcripts, and thus can code for multiple products, each with potentially different functions. However, not all RNA is used to make proteins. Some types of RNA bind to and work with proteins to support fundamental cellular processes, whilst others prevent or influence the production of other RNA molecules or proteins. The RNA molecules that *are* converted to a protein are termed messenger RNA or mRNA. Studying which mRNA transcripts are expressed, and to what level they are expressed, can give us insights into which proteins and molecules are required under distinct conditions, for example, in different cells, tissues or organs, or in response to an infection or change in environment.

The current main approach used by scientists for studying mRNA is called RNA-sequencing (RNA-seq). This method allows us to identify both the sequence and expression level of mRNA molecules in a sample. The most common method of RNA-seq involves taking an mRNA molecule, cutting it into little chunks, and then sequencing these fragments. The fragmentation of the mRNA and the resulting small chunks of data produced by this technique give rise to its name: short-read RNA-seq. These short reads can be either stitched back together to form a complete strand of mRNA in order to build a database of full-length mRNA sequences, or can be directly compared with pre-existing mRNA databases to quantify expression.

Whilst short-read methods sequence the individual chunks of mRNA accurately and are well-established for expression analyses, the process of stitching the short reads back together again to form full-length mRNA sequences can be difficult, akin to piecing together a jigsaw puzzle. This poses a serious challenge if we are interested in understanding how transcripts with different structures, which may have different functions, are being expressed. Recently, newer RNA-sequencing technologies have been introduced which do not involve cutting the mRNA. These techniques, called long-read RNA-seq, can sequence the mRNA molecule in a single pass, more similar to taking a photograph than piecing together a jigsaw puzzle. This allows us to better characterise the diversity of mRNA produced by each gene and examine the functional consequences of expressing different mRNA transcripts.

In my project, I use a novel long-read RNA-sequencing technique to explore the role of RNA diversity and examine mRNA expression in the immune system and during early development in Atlantic salmon, a key aquaculture species of cultural and commercial significance. To do this, salmon embryos at different stages of development, as well as a tissue involved in fish immunity, were sampled. I sequenced the full-length of the mRNA molecules in these samples, and developed new analysis methods to reveal the expression dynamics involved in each condition. My thesis reveals mRNA expression patterns that would be missed by short-read RNA-seq methods. Additionally, the techniques I developed can be transferred to other organisms and studies, forming a foundation for future RNA research.

Chapter 1: General Introduction

Framing:

In this thesis, I report the first Oxford Nanopore long-read transcriptome for Atlantic salmon and explore transcript diversity expressed during early development and in response to immunostimulation, activation of the immune system, through injection of bacterial and viral mimics. This chapter introduces concepts necessary to contextualise my aims and findings, including the development of long-read sequencing technologies and the importance of genome functional annotation. Atlantic salmon as a species is introduced and I review the current state of knowledge on fish embryogenesis and immune responses. I end the chapter by detailing the specific aims and objectives of my project.

1.1 The Transcriptome

1.1.1 The Central Dogma of Molecular Biology

All the genetic information required for an organism to produce proteins, build cells, construct organs and develop tissues is stored in a molecule called deoxyribonucleic acid (DNA). DNA is made up of four nucleic acid bases; adenine, thymine, cytosine and guanine, often abbreviated A, T, C and G, which are arranged in complementary pairs (A-T, C-G). Base pairs are connected by sugar-phosphate links and form the classic, double-stranded helical structure (Watson & Crick, 1953) displayed in biology classrooms worldwide. Together, these four bases encode all products an organism needs to function and the entire complement of DNA is termed the genome. For DNA to be converted into a usable product, for instance the diverse repertoire of proteins used to perform cellular functions, it is first copied into an intermediate, single-stranded molecule called ribonucleic acid (RNA). RNA and DNA share three of the four bases with thymine being substituted for uracil (U). The flow of information, from DNA to RNA to proteins, the machines underlying all biological functions, is called the ‘Central Dogma of Molecular Biology’ (Crick, 1970; Cobb, 2017).

1.1.2 RNA Transcription & Synthesis

RNA transcription is the process by which DNA is copied into RNA and regions of the genome that are transcribed into RNA are termed genes. In RNA transcription, a DNA helicase unwinds the double-stranded DNA molecule whilst an RNA polymerase binds to the antisense strand, reading it in the 3'-5' direction and synthesises a complementary strand. As a result, a single-stranded RNA molecule is produced which matches the sense strand of the gene.

The rate at which RNA is transcribed is regulated by a variety of mechanisms; promoter regions are sequences of DNA lying upstream of a gene and serve as binding sites for transcription proteins (Haberle & Stark, 2018; Andersson & Sandelin, 2020), transcription factors are proteins which bind to promoter regions and can either enhance or repress transcription via interactions with RNA polymerase (Lambert et al., 2018), whilst other genomic elements like enhancers and silencers either increase or decrease expression by interacting with transcription factors (Shlyueva et al., 2014; Pang et al., 2023).

Not all RNA molecules are directly converted into protein and are thus deemed to be non-protein-coding, or non-coding for short. One of the most abundant classes of non-coding RNA is ribosomal RNA (rRNA), which forms parts of the ribosome - a molecular unit that synthesises protein (Hori et al., 2023). Another class of RNA involved in protein synthesis but not directly translated into protein is transfer RNA (tRNA). tRNAs carry amino acids, the building blocks of proteins, to the ribosome and serve as adapters between RNA that is being translated and the growing polypeptide chain (Berg & Brandl, 2021). Additional classes of non-coding RNAs have recently captured the attention of researchers for their involvement in the regulation of gene expression. These include long non-coding RNA (lncRNA) which have been implicated in the control of multiple transcriptional mechanisms (Tsai et al., 2010; Jathar et al., 2017), microRNA (miRNA), small non-coding RNAs that can inhibit protein synthesis by binding to protein-coding RNAs (Ye et al., 2019), and circular RNA (circRNA), a class of non-coding RNA initially

thought to have function similar to miRNA that has yet to be fully elucidated (Salzman, 2016).

Whilst not as ubiquitous as non-coding RNA, making up only 2-5% of total RNA production, protein-coding messenger RNA (mRNA) is the most studied class of RNA. mRNA is transcribed from genes and then subjected to a suite of post-transcriptional modifications in the nucleus, including the excision of intronic regions, 5' capping, and the addition of a polyA tail to the 3' end of the molecule. Now mature, messenger RNAs exit the nucleus to be translated into proteins by the ribosomal machinery.

1.1.3 Generating mRNA Transcript Diversity

A further layer of complexity is added to mRNA expression via the post-transcriptional modification of immature, or pre-RNA. Pre-RNA consists of segments called introns, which are spliced out of the molecule, and exons, which are retained in the mature RNA. A mechanism called alternative splicing (AS) causes different combinations of exons and introns to be retained in the final mature RNA enabling single genes to produce multiple RNA variations (Figure 1.1; McManus & Graveley, 2011) that may have distinct protein functions or RNA properties (Manuel et al., 2023). AS has been shown to be a fundamental and important process (reviewed in Marasco & Kornblihtt, 2023), involved in events such as the acquisition of tissue functions during early development (Baralle & Giudice, 2017) and linked with diseases such as cancer (Frankiw et al., 2019) or neurodegenerative conditions (Nikom & Zheng, 2023) in humans.

In addition to AS, the use of alternative transcription start (TSS) and termination sites (TTS) contributes further to transcriptional diversity, leading to the formation of transcripts with differing 5' and 3' ends (Figure 1.1), which can influence important biological processes (Reyes & Huber, 2018). Alternative polyadenylation sites and differences in polyA-tail length can also affect gene translation by altering RNA stability (de Klerk & 't Hoen, 2015). Furthermore, additional post-transcriptional modifications can affect the fate of mature RNA. Chemical modifications like m6A methylation may influence

RNA decay and extra-nucleus transport rates (Fu et al., 2014; Nachtergaele & He, 2017), leading to another layer of post-transcriptional regulation.

The sum of all RNA molecules, or transcripts, expressed at any given moment is termed the transcriptome. In addition to defining different tissues and stages of ontogeny, the transcriptome is highly variable in response to stimuli such as physiological stress, or environmental conditions. Studying the transcriptome thus allows the dissection of fundamental pathways and mechanisms associated with organism function including growth and development, cell fate, and disease progression (Reese et al., 2023).

1.1.4 Regulation of RNA Transcription

RNA transcription is a complex process regulated by many different mechanisms. Epigenetic modification can regulate transcription in a variety of ways. Direct modification to histones, proteins around which genomic DNA is wound, such as methylation of H3K27, has been shown to reduce histone accessibility and acts as a gene silencer (Ferrari et al., 2014), whereas acetylation of histones causes conformational changes which allow RNA transcription to occur more readily (Marmorstein & Roth, 2001). Specific regions of DNA that initiate the binding of transcriptional machinery lie upstream of genes and are called promoters (Lin et al., 2017). Promoters can be epigenetically modified via the addition of methyl groups to DNA in these regions. Increased levels of methylation cause gene repression by recruitment of gene repressive proteins or blocking of transcription factor binding sites (Moore et al., 2013). Regulating the transcription of individual genes is also achieved through *cis*-regulatory elements such as enhancers (Bulger & Groudine, 2011) or silencers (Pang et al., 2023), which recruit transcription factor (TF) proteins that influence transcription rates of target genes, with their activity depending on biological context (Spitz & Furlong, 2012).

1.1.5 Early Technologies for Transcriptomics

Introduced in the 1970s, Sanger sequencing, also known as the chain-termination method (Sanger et al., 1977), was the prevailing sequencing technology for the rest of the 20th Century and was first used for

transcriptomics in the 1990s. Developed by Adams et al. (1992), the expressed sequence tag (EST) strategy was the first widespread method for transcriptome analysis. The EST strategy involves the sequencing of cDNA clones via Sanger technology. Whilst useful for identifying novel genes and generating sequence data, the EST method is expensive and laborious with low throughput, limiting scope for quantitative analysis.

Hybridisation-based microarrays were introduced shortly after the EST strategy and allowed gene expression to be measured (Hrdlickova et al., 2017), thus providing a vehicle for quantitative transcriptome analysis. In microarray technology, fluorescently-labelled cDNA copies of transcripts are identified when they bind to a series of immobilised, custom-made nucleotide probes (Schena et al., 1995). Microarrays are widely used to quantify relative gene expression by measuring fluorescence intensity. However, microarray analysis possesses two fundamental issues. First, the microarray probes must be designed from pre-existing genome or transcript information, hindering their ability to detect novel transcripts. Second, high levels of both background binding and cross-hybridisation can limit the detection of lowly expressed transcripts (Okoniewski et al., 2006). The advent of high-throughput RNA-sequencing (RNA-seq) in the mid-2000s helped to alleviate these issues (Wang et al., 2009) and initiated a ‘transcriptomics revolution’.

1.1.6 High-Throughput RNA-Seq and Second-Generation Sequencing

The first human genome was completed using Sanger sequencing in 2004 (International Human Genome Sequencing Consortium, 2004) at an estimated cost of $2.7 billion. An explosion in the development of sequencing technology followed this milestone, thanks to a funding initiative introduced by the National Human Genome Research Institute, which aimed to reduce the cost of sequencing each human genome to <$1000 (Schloss et al., 2020). Now referred to as Next-Generation Sequencing (NGS) or Second-Generation Sequencing (SGS), these new platforms were able to sequence millions of nucleotides in parallel at a fraction of the time and cost of Sanger sequencing (Liu *et al.,* 2012; van Dijk *et al.,* 2014; Reuter *et al.,* 2015). The significant reduction in per-base cost offered by these massively parallel

platforms (hereinafter referred to as SGS) gave small labs and researchers access to sequencing technology, revolutionising genomics studies (van Dijk *et al.,* 2014).

RNA-seq on SGS platforms transformed transcriptomic research through their ability to sequence cDNA. During the early 2000s many SGS technologies were released to the market and used for RNA-seq (reviewed in Metzker, 2010). In general, RNA-seq on SGS platforms shared similar steps to generate sequencing libraries: RNA molecules are either converted to cDNA and fragmented randomly into shorter sections, or fragmented before conversion to cDNA. The small fragments of cDNA are then multiplied to increase their abundance via polymerase chain reaction (PCR) amplification. Sequencing adapters are ligated onto the fragments before being sequenced in a high-throughput manner (Wang et al., 2009). Resulting data is then either aligned to a reference genome or transcriptome database, or assembled into a custom transcriptome *de novo* to show the structure and expression level of each gene and transcript. Today, Illumina dominates the SGS market with their sequencing-by-synthesis (SBS) strategy, which measures the fluorescence of labelled nucleic acid bases when incorporated into a complementary strand during PCR amplification. This method is highly accurate (Q30+: 99.9% per-base accuracy) with their largest machine the NovaSeq X producing up to 8Tb per flowcell (Illumina, 2024), equating to 26 billion reads of 150bp in length.

Another viable SGS method is the MGI-Tech DNBSEQ-T7 platform, which can produce approximately 1 billion paired-end 100bp reads of comparable quality and accuracy to Illumina's platforms (Kim et al., 2021a). Like Illumina, the MGI method leverages fluorescent sequencing-by-synthesis probes, however, MGI uses nucleotides fluorescently labelled with antibodies as opposed to Illumina's fluorescent dye approach. In the MGI approach, cDNA is synthesised into circular structures and amplified to form DNA nanoballs (DNB) which are then hybridised to a flowcell (Drmanac et al., 2010). Sequencing continues in a similar fashion to Illumina, with the fluorescent signatures produced by the addition of labelled nucleotides being recorded in real-time.

The majority of RNA-seq studies to-date use Illumina SBS technology (Stark et al., 2019). However, the fragmentation of RNA molecules to form short reads and mandatory PCR amplification required by SGS methods can introduce major challenges for transcriptomic analysis. All SGS platforms sequence short reads, which are either mapped to reference transcript sequences or computationally stitched together to form longer contiguous sequences post-sequencing. Fragmentation of RNA during library preparation occurs stochastically along the length of the molecule. As a result, when mapping to reference transcripts some reads may span exon-exon junctions, whilst others fall entirely within an exon (Figure 1.2); it will be extremely rare for single short-read sequences to completely capture the full-length of transcripts. This introduces difficulties when trying to locate alternative splice junction start and end sites, identify alternative transcript isoforms, as well as determine intron retention and exonic chaining. These challenges, combined with the fact that reference transcriptomes are often generated through *in-silico* prediction software based on SGS sequencing can lead to incomplete transcriptome annotation (Steijger et al., 2013; Kuo et al., 2017). Additionally, shorter reads can be ambiguously mapped to highly repetitive regions of the genome or similar gene copies (Li & Dewey, 2011) and PCR amplification during library preparation can introduce sequencing bias (Aird et al., 2011).

The advent of long-read sequencing and its rise to prominence over the last decade may help overcome these issues.

1.2 Long-Read Transcriptomics

1.2.1 Advent of Long-Read Sequencing Technology

Short-read platforms remain the current standard for RNA-seq due to their low cost, high per-base accuracy and high throughput (Stark et al., 2019). However, long-read sequencing technologies developed in recent years have proven valuable tools for RNA-seq due to their lack of RNA fragmentation during library preparation. This allows full-length molecules to be sequenced in a single pass which reduces ambiguity when mapping reads to the genome, as well as improving the identification of alternative

splice sites, thus allowing better characterisation of transcript diversity and identification of novel alternative isoforms (Kuo et al., 2017; Seki et al., 2019; Stark et al., 2019).

Introduced by Pacific Biosciences (PacBio) in 2011, Single Molecule Real Time sequencing (SMRT) was the first publicly available long-read technology (McCarthy, 2010). A few years later, Oxford Nanopore Technologies (ONT) released their MinION device to the market in 2015 (Jain et al., 2015). Collectively known as third-generation sequencing (TGS), both technologies were initially characterised by high error rates and low throughput compared with short-read approaches (Quail et al., 2012; Weirather et al., 2017). However, with the introduction of circular consensus sequencing (CCS) in PacBio systems, and improvements to platform chemistry and basecalling software for ONT sequencing, the error rates and throughput of both approaches has improved dramatically since inception (Van Dijk et al., 2018; Liu-Wei et al., 2024).

Despite wide use for transcriptome construction and alternative transcript discovery, long-read RNA-seq is yet to be routinely applied for quantitative analysis. Low throughput and high error rates seen in early iterations of TGS were prohibitive for quantitative analysis (Amarasinghe et al., 2020; Hu et al., 2021). Additional challenges remain for long-read data analysis where few standardised pipelines have been created and validated.

In the following sections I explore in more detail the two distinct methods employed by PacBio and nanopore RNA-seq and the advantages and disadvantages of the two platforms.

1.2.2 PacBio Iso-Seq

Similar to most short-read platforms, PacBio's SMRT approach, also termed Iso-Seq, utilises sequencing-by-synthesis whereby a DNA polymerase synthesises a complementary strand from a sample cDNA template strand. In short, RNA is converted to full-length cDNA using a strand-switching polymerase – no further rounds of PCR amplification are carried out. Hairpin adaptors are ligated onto both ends of the cDNA, creating a circular, single-stranded molecule. A template-polymerase complex is generated for each

read which is subsequently tethered to the bottom of micron-scale wells, termed zero-mode waveguide (ZMW) chambers. As a complementary strand is synthesised, fluorescently-labelled nucleotide bases are incorporated, emitting a pulse of light which is recorded by the machine. Each base emits a different wavelength which is used to then generate the sequence in a process called basecalling

In its early stages of development, the PacBio RS system could only produce a maximum of 1Gb of data, with reads up to 1.5kb in length and an error rate of approximately 13% (Quail et al., 2012; Van Dijk et al., 2018). However, the next generation of the PacBio system, the Sequel II was a drastic improvement. Output was increased 10-fold, producing 5-10Gb of reads up to 15kb in length while reducing the average per-base error rate through CCS. As the template DNA strand is circular, it is possible to synthesise a complementary strand multiple times. CCS (also known as 'HiFi' sequencing) is the process of deriving a consensus sequence from multiple passes (Travers et al., 2010). CCS has reduced raw error rates to <1% (Wenger et al., 2019) and can reach sequencing accuracies of 99.999% after 25 passes (van Dijk et al., 2018). Nevertheless, read lengths on the Sequel were limited to an average of 15kb (Stark et al., 2019). PacBio's current flagship system, the Revio, can produce upwards of 90Gb of HiFi reads in a single SMRT-cell with an accuracy of 99.95% (Q31) at a cost of $995 USD per cell (https://www.pacb.com/revio/). The Iso-Seq data analysis pipeline (available here: https://isoseq.how/) is the most established long-read RNA-seq analysis pipeline for obtaining full-length reads and generating a transcriptome assembly, thus offering a simple workflow for researchers to employ quickly.

In humans, PacBio RNA-seq was used to improve transcriptome annotation, providing evidence for approximately 4,000 novel gene models of which 2,500 were deemed non-coding (Kuo et al., 2020) following sequencing of a universal human cell line (UHRR). Another study in humans described significant pseudogene transcript diversity using PacBio sequencing of foetal and adult tissues in addition to the same UHRR dataset from Kuo et al. (2020) twinned with CRISPR-Cas9 knockout and CAGE-seq validation

(Troskie et al., 2021). Moving away from humans, Ramberg and colleagues (2021) used PacBio RNA-seq with matched short-read error-correction to identify over 71,000 full-length transcripts in Atlantic salmon *Salmo salar*, of which 75% were not present in the reference annotation. This is a significant increase in novel transcript discovery compared with humans (e.g. Kuo et al., 2020) due to the fact that humans have a greater body of work supporting their transcriptome annotation. This highlights the utility of long-read sequencing for profiling the transcriptomes of non-model species. A similar approach was employed in the salmon louse *Lepeophtheirus salmonis* and revealed significant transcript diversity expressed at different life-stages of the louse (Hansen et al., 2023). These studies show the ability of long-read sequencing to catalogue hitherto undescribed transcript diversity and elucidate its role in both model and non-model species.

1.2.3 Nanopore RNA-Seq

Distinct from PacBio and SGS methods, ONT's method reads nucleic acid sequences as they thread through nanopores (Figure 1.3). Pore proteins are embedded in an electrically-resistant lipid membrane separating two compartments filled with electrolyte solution – this forms the flow cell. An ionic current, produced by applying a constant voltage bias across the membrane, causes the nucleic acid molecule to pass through the pore, regulated by a helicase. During this process, different nucleotides induce characteristic disruptions to the ionic signal. This signal is recorded and the disruptions are used to determine the nucleotide sequence of the molecule. Unlike PacBio's SMRT approach, the read lengths produced by nanopore sequencing are theoretically unlimited, restricted only by the length of the input molecules (van Dijk et al., 2018).

ONT currently offers two methods of RNA-seq. The first protocol, like PacBio, involves generating full-length cDNA using reverse transcription and strand-switching, ligating on sequencing adaptors and loading into the flow cell. This method can be carried out with, or without PCR amplification. Whilst the second method does not utilise strand-switching, it does synthesise a complementary strand to form a stable cDNA-RNA hybrid. Only

the native RNA molecule is sequenced however, as sequencing adapters are ligated only to the RNA strand. This is termed direct RNA sequencing (Garalde et al., 2018). Without any PCR, native RNA sequencing preserves epigenetic information such as m6A modifications, which allows epigenetic and sequence information to be captured simultaneously during the run (Workman et al., 2019).

Low throughput and high raw basecalling error rates historically characterised nanopore sequencing. However, improvements in nanopore sequencing chemistry have increased throughput and made differentiating the ionic signal disruptions easier for basecalling software, thus improving raw accuracy (van Dijk et al., 2018). For example, using the now retired R9.4.1 chemistry, a single MinION flow cell could generate over 40Gb sequence data, a 100-fold increase on the previous R7 generation (Brown & Clarke, 2016). According to ONT, the current R10.4.1 sequencing chemistry kits can produce reads with accuracies in excess of 99.5% (Q23), with novel duplex basecalling algorithms producing reads of Q30+ (ONT, 2023).

Furthermore, ONT have a range of sequencing machines to customise throughput to study requirements. The largest, the PromethION, can generate over 200Gb of data per flow cell, scaling up to above 13Tb of sequence data (comparable with Illumina) at maximum flow cell capacity. For RNA-seq, the PromethION can produce greater than 100 million reads for transcript analysis.

Like PacBio, ONT platforms have been used to investigate transcriptomic diversity in a variety of species. Multiplexing samples from 32 tissues on a single PromethION flowcell from Hereford cow (*Bos taurus*) Halstead et al. (2021) generated 99,044 high confidence transcript models, of which 61% were novel from previously annotated genes, revealing extensive tissue-specific diversity. Nanopore long-read RNA-seq was used successfully in humans, capturing 70,000 novel transcripts from known genes (Glinos et al., 2022). Other transcriptome annotations generated with Nanopore sequencing include those in the ovary of Muscovy duck *Cairina moschata* revealing novel transcriptomic diversity expressed during ovulation (Lin et al.,

2021), while >20,000 novel transcripts were discovered in honeybee *Apis mellifera* midgut tissue (Zang, 2024). ONT methods have also been used in a more targeted way, for example, Nanopore RNA-seq has aided the transcriptomic profiling of diseases such as herpes simplex virus type 1 (Depledge et al., 2019) and recently human adenovirus type F 41 (Abebe et al., 2024), and showed expression of widespread transcript variation on the surface of B-cells (Byrne et al., 2017), furthering our understanding of the role of transcript diversity for immune function.

A significant bottleneck for analysing ONT data is basecalling, the process of converting raw signal data into base information. This process often requires considerable computing power and can take multiple days to complete. In addition to sequencing chemistry updates, basecalling software is evolving at a rapid rate, reducing raw error rates and the time and computing power required. The addition of neural networks to ONT basecalling algorithms was shown to reduce raw error rates from 15% to ~5% (Jain et al., 2018; Wick et al., 2019), whilst increasing basecalling speed. The GridION and PromethION currently come with built-in GPUs, which support live-basecalling - reducing the time and computational costs associated with independent basecalling.

1.3 Genome Functional Annotation

The technological advancement in sequencing technology over the past 20 years (sections 1.1 & 1.2) has allowed genome sequencing to be conducted rapidly and at low cost. This has led to a drastic increase in the number of publicly available reference genome assemblies for both model and non-model species. For context, the latest Ensembl Release 112 (6th May 2024: https://www.ensembl.org/index.html) contains 324 reference genome assemblies, a 200-fold increase compared with the 16 assemblies available in 2005 (Hubbard et al., 2005). Ensembl rapid release is updated every two weeks with new genome assemblies and contains 2,591 genome assemblies as of 4th September 2024 (https://rapid.ensembl.org/index.html). Figure 1.4 depicts the increase in the number of eukaryotic genomes in NCBI's GenBank repository from 0 (in the year 2000), to over 41,000 assemblies

today[1]. This rapid advancement in genomic resources is set to continue into the future with projects such as Darwin Tree of Life, which aims to assemble genomes for all eukaryotes in the UK and Ireland (Darwin Tree of Life Project Consortium, 2022), whilst the Earth BioGenome Project aims to sequence all eukaryotic species on Earth (Lewin et al., 2018; Lewin et al., 2022).

Alongside the wealth of new genome sequences, there is a growing ambition to identify which regions of DNA show identifiable functional activity (Giuffra et al., 2019) in a process called genome functional annotation. Annotating functional genomic elements including coding and non-coding genes and their transcript variants, regulatory elements (e.g. promoters and enhancers), dynamic epigenetic modifications (e.g. methylation, chromatin accessibility, histone marks) and long-range DNA-DNA interactions, is essential for understanding the genetic basis for phenotypic variation. This enables a shift from simply describing what DNA elements are present, to being able to predict traits exhibited by the organism based on genomic information. Such approaches allow for elucidation of genes and genetic variants causative for traits, discovery of biomarkers and the construction of complex gene regulation pathways (Ritchie et al., 2015; Giuffra et al., 2019; Woolley et al., 2023).

The Encyclopaedia of DNA Elements (ENCODE) initiative was the first and most high-profile functional annotation project. ENCODE commenced in the early 2000s, with the goal to characterise functionality in 30Mb (~1%) of the human genome (ENCODE Project Consortium, 2007). As the ENCODE project progressed, a rapidly expanding ‘omics toolkit, allowed more ambitious goals to annotate the whole genome (ENCODE Project Consortium, 2020). The most recent iteration of the ENCODE project, ENCODE4, has focussed on long-read sequencing to generate full-length transcriptome annotations for human and mouse (Reese et al., 2023). Since the inception of ENCODE, further related consortia have formed for diverse organisms including the Functional Annotation of Animal Genomes (FAANG) initiative, which focuses on farmed animals (Giuffra et al., 2019), DANIO-

[1] *Data extracted from Hjelmen (2024) and updated using a custom R script on 03/09/2024*

CODE, which annotated functional elements in the model zebrafish *Danio rerio* (Tan et al., 2016; Baranasic et al., 2022), and AQUA-FAANG, which produced comprehensive annotations of functional elements across the genomes of six farmed fish species (see section 1.4.4; Johnston et al., 2024).

1.3.1 Functional Annotation Assays

The increase in technical capability of sequencing technology has led to the development of diverse methods for studying genome functional activity (Giuffra et al., 2019). Most relevant to my project, various flavours of RNA sequencing (RNA-seq) allow gene expression to be quantitatively measured, while also revealing transcriptome structural diversity (see section 1.3.2). Beyond the transcriptome, non-coding elements like enhancers and promoters can be identified by the digestion of DNase I cleavage sites followed by sequencing (DNase-seq; Song & Crawford, 2010). This technique leverages the fact that regions of open chromatin, which host active regulatory elements, are more sensitive to DNase I (ENCODE, Project Consortium, 2007; Boyle et al., 2008; He et al., 2014). The most popular current method to measure chromatin accessibility is the Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) (Buenrostro et al., 2015).

Additionally, it is possible to characterise DNA-protein interaction sites across the genome using Chromatin Immunoprecipitation (ChIP) sequencing (ChIP-seq; Park, 2009). ChIP-seq captures proteins bound to DNA using antibodies, which can identify biochemical modifications to different histone marks across the genome, which are strongly indicative of regulatory activity (Nakato & Sakata, 2021). For example, h3k27me3 marked regions are associated with actively repressed genes, while h3k27ac marked regions are associated with enhancer activity. ChIP-seq can also be used to locate DNA binding sites for transcription factors, proteins which regulate transcription rate through several mechanisms (Lambert et al., 2018). Hi-C is a technique for investigating DNA-DNA interactions (Belton et al., 2012), which can reveal genomic structures of chromosomes (often used in *de novo* genome

assembly; Dudchenko et al., 2017), while also allowing the identification of long-range cis-regulatory elements and their interactions with target genes (Lan et al., 2012). Such functional assays provide evidence of how gene expression is regulated by a variety of mechanisms.

1.3.2 Functional Annotation using RNA-Seq

RNA-seq is a major tool for genome functional annotation, with both short-read and long-read methods widely adopted in recent years (Figure 1.4). RNA-seq allows us to describe gene structure and location within the genome, alongside gene expression in any sample of interest, with scope to resolve alternative transcripts and their potential functions. Differential gene expression (DGE) analysis of mRNAs remains the primary use of RNA-seq (Stark et al., 2019), but there are established tools for transcript-level expression analyses such as differential transcript expression (DTE) which aims to identify significant changes in expression of individual transcripts across all transcripts (Trapnell et al., 2013; Yi et al., 2018), and differential transcript usage (DTU) which looks to determine if the proportions of expressed transcripts change significantly within a single gene (Love et al., 2018; Marques-Coelho et al., 2021). Other common applications are assessing mRNA splicing (Wang et al., 2008) and identifying expression quantitative trait loci (DNA variants impacting gene expression level) (Kukurba & Montgomery, 2015). In addition, RNA-seq has facilitated the analysis of non-coding RNAs. This has led to insights into the complex roles of non-coding RNAs, such as microRNAs and long non-coding RNAs (lncRNA), in regulating gene expression (Morris & Mattick, 2014).

Variations of RNA-seq exist, including CAGE-seq, which allows the accurate identification and quantification of transcription start sites (Adiconis et al., 2018), whist single-cell RNA-seq (scRNA-seq) methods offer the ability to investigate gene expression within a population of individual cells (Luecken & Theis, 2019). As gene expression is not homogeneous, even within the same cell types (Huang, 2009), scRNA-seq allows a higher resolution of gene expression study (Tang et al., 2009; Hwang et al., 2018) and is quickly becoming the gold standard in many fields of study.

For long-read RNA-seq, the ability to capture full-length RNA transcripts in a single read is a powerful approach for alternative transcript identification and quantification (Hardwick et al., 2019) leading to the improvement of many current transcriptome annotations (Sharon et al., 2013; Kuo et al., 2017; Nudelman et al., 2018; Kuo et al., 2020). With the recent improvements in throughput and accuracy of long-read RNA-seq data, these technologies have been shown to be effective for conducting transcript-level quantification analyses such as DTE and DTU (Dong et al., 2021; Wright et al., 2022).

1.4 Aquaculture

The human population is projected to reach up to 10 billion by 2050, placing great strain on global food resources (United Nations, 2022). In particular, there will be a large increase in the demand for animal protein to feed this growing population (Alexandratos & Bruinsma, 2012). For the past 60 years, the average annual growth in global fish consumption has surpassed that of any other animal protein, contributing 15% of global animal protein consumed in 2021 (FAO, 2024a). Aquaculture is the fastest growing food sector (Anderson et al., 2017) and produced 94.4 million tonnes of fish and shellfish in 2022 (FAO, 2024a), surpassing capture fisheries production for the first time, consisting of 51% of total production from aquaculture and capture fisheries combined. With capture fisheries production stagnating and many fish species in decline (Pauly and Zeller, 2016), aquaculture is expected to permanently surpass that of capture fisheries from 2024 (OECD, 2020).

In 2022, total aquaculture production surpassed 130 million tonnes of live weight with an estimated value of USD 312.8 billion, the highest ever recorded (FAO, 2024b). Of this, finfish contributed the greatest proportion of total production weight (61.6 million tonnes; 47.0%), with aquatic plants producing 36.5MT (27.9%), and molluscs and crustaceans contributing 18.91MT (14.4%) and 12.8MT (9.7%) respectively (FAO, 2024b). The rapid growth of aquaculture and projected increase in demand for aquatic protein offer a distinct opportunity for the incorporation of novel genomic technologies to improve production, welfare and sustainability of the industry.

1.4.1 Potential for 'Omics in Aquaculture Research and Practice

Chromosome-level genome assemblies, that is genome assemblies where the majority of sequences are anchored into scaffolds representing individual chromosomes, for key aquaculture species like Atlantic salmon (Lien et al., 2016) and Nile tilapia (Conte et al., 2017) have greatly advanced aquaculture research allowing the use of 'omics to resolve mechanisms underpinning the biology of aquaculture species (Houston & Macqueen, 2019; Houston et al., 2020). Utilising these genomic resources via the rapidly expanding 'omic toolbox is essential for aiding sustainable growth of global aquaculture (Abdelrahman et al., 2017

One such area of interest is to understand the basis of host defence to pathogens as a way of combatting aquatic diseases. Measuring host transcriptomic responses to infection provides knowledge for understanding immune response mechanisms (e.g. Krasnov et al., 2021; Zhong et al., 2023), whilst studying pathogen transcriptomes during infection or treatments provides key insight into host-pathogen interactions e.g. how host defences are evaded (Gallardo-Escarate et al., 2014). Transcriptome profiling has also been applied to investigate responses to alternative feed sources such as plant-based feeds on fish health (e.g. Tawfik et al., 2024). Furthermore, transcriptomics has been used to measure the efficacy of treatments delivered through functional feeds (Cai et al., 2022), leading to a better understanding of the relationship between diet and immune health.

The molecular basis of key traits for aquaculture species remains vastly underexplored in comparison with terrestrial agriculture species. Thus, there exists a great opportunity for the use of 'omics to better improve our understanding of these traits and use that knowledge to inform breeding practices (Bernatchez et al., 2017; Johnston et al., 2024), improve fish health via therapeutic treatments (Nguyen, 2024) and employ genomic selection (Houston et al., 2020) to ultimately increase the sustainability of the aquaculture sector (Abdelrahman et al., 2017).

1.4.2 Atlantic Salmon

Atlantic salmon is a fish species of vast economic and cultural importance (Macqueen et al., 2017). Atlantic salmon aquaculture production generated USD 21.9 billion constituting 13.3% of all finfish revenue in 2022, despite only contributing 7.0% of finfish production weight (FAO, 2024b). Atlantic salmon farming is dominated by a handful of countries, namely Norway, Chile, Scotland and Canada, which together contribute 87% of global salmon production by value (FAO, 2024b). Aside from aquaculture, wild Atlantic salmon have extensive cultural and conservational value, are prized for recreational wild fishing, show fascinatingly diverse life history strategies, and have been used as biological indicators for quality of the aquatic environment (Davidson et al., 2010).

As a result of economic and societal interest, there exists a large body of published research on Atlantic salmon. This includes a focus on immunology to understand immune function and response to pathogens (Magnadottir, 2010; Robertsen, 2018). Moreover, there is an interest in Atlantic salmon physiology and osmoregulation to understand the basis for anadromy - the ability to migrate between freshwater and salt water during the life cycle (Handeland et al., 1998; Vargas-Chacoff et al., 2018). Many studies on nutrition have investigated the impact of functional feeds administered to farmed Atlantic salmon at different life stages (Tacchi et al., 2011), or explored how starvation periods impact immune performance (Martin et al., 2010). To meet this research demand, Atlantic salmon has extensive genomic resources including a high-quality reference genome (Lien et al., 2016) and has been the subject of international efforts to functionally annotate the genome (see section 1.4.4).

The salmonid lineage is also notable for its history of whole genome duplication (WGD) events. The ancestor to jawed vertebrates underwent two separate rounds of WGD, one in the common vertebrate ancestor, and another after the split of jawed and jawless fish (Dehal & Boore, 2005; Holland & Ocampo Daza, 2018; Simakov et al., 2020; Marlétaz et al., 2024). WGD is a process by which all genomic information is doubled and can

occur within species (autopolyploidy) or after two distinct species have hybridised (allopolyploidy). A third round of WGD was experienced by the ancestor of all teleosts around 300 Mya (Teleost-Specific Third Round of WGD [Ts3R]), which increased genomic complexity and provided teleosts the genomic material required for diversification (Santini et al., 2009; Hurley et al., 2007; Glasauer & Neuhass, 2014; Qi et al., 2024). The salmonid ancestor underwent a autopolyploidy WGD event approximately 100 Mya (Macqueen & Johnston, 2014), referred to as the salmonid-specific 4th Round of WGD (Ss4R), with modern descendants, including Atlantic salmon, possessing regions of the genome still displaying signatures of the rediploidisation process and many (co)orthologues of genes found in mammals (Berthelot et al., 2014; Lien et al 2016; Robertson et al., 2017; Gundappa et al., 2022). The functional redundancy provided by WGD enables duplicated gene pairs to evolve with very different constraints than before duplication (Glasauer & Neuhauss, 2014). A variety of fates for gene duplicates, have been proposed, including; 1) non-functionalisation, where one duplicate is lost (Ohno, 1970; Brunet et al., 2006), 2) sub-functionalisation where both duplicates are retained and each carry out part of the ancestral gene function (Force et al., 1999; Des Marais & Rausher, 2008), and 3) neo-functionalisation, where one duplicate retains the ancestral function, whilst the other evolves a new function (Ohno, 1970). Thus, the salmonids represent an interesting system for studying the impacts of WGD on genome evolution and gene expression (Berthelot et al., 2014; Garcia de la Serrana & Macqueen, 2018; Varadharajan et al., 2018; Colgan et al., 2021). On the other hand, the duplicated genes retained after the Ss4R WGD, introduce challenges for ‘omics methods. For example, gene editing in salmonids is challenging due to the need to consider whether gene duplicates should be co- or single-targeted (Cleveland et al., 2018; Blix et al., 2021). Transcriptomic and other omics analysis can prove difficult with some gene duplicates only differing by a few base pairs across their entire length. This can lead to a high prevalence of multi-mapping (where sequencing reads have more than one mapped location with equal probability) when using short-read RNA-seq, a problem which may be alleviated by long-read RNA-seq.

1.4.3 Aquatic Diseases Affecting Salmonid Aquaculture

Aquatic diseases are the largest barrier to expansion for the global aquaculture sector (Stentiford et al., 2015). Due to the marriage of intensive farming and interaction with the open marine environment, salmonid aquaculture is particularly threatened by various aquatic diseases (Pettersen et al., 2015a). Viruses such as salmonid alphavirus (SAV), the causative agent of pancreas disease, have been shown to cause mass mortality in Atlantic salmon farms (Jansen et al., 2017) leading to significant economic losses (Aunsmo et al., 2010; Pettersen et al., 2015b). Piscine myocarditis virus (PMCV), the causative agent for cardiomyopathy syndrome has recently emerged as a virus of concern for both wild and farmed salmon stocks (Garseth et al., 2018), accounting for the greatest number of reported viral detections in Norwegian aquaculture last year (Sommerset et al., 2024).

Ectoparasitic sea lice (*Lepeophtheirus salmonis* and *Caligus* spp.) cause topical skin lesions which reduce growth rates (Torrissen et al., 2013) and increase pathogenicity of other aquatic diseases (Gharbi et al., 2015), whilst complex gill disease, including Amoebic Gill Disease (AGD) caused by *Neoparamoeba perurans,* is emerging as a significant problem for the sector (Boerlage et al, 2020). Several bacterial diseases such as furunculosis (*Aeromonas salmonicida*) have historically been successfully controlled through vaccination (Pettersen et al., 2015a), but some remain problematic. For example, the Chilean Atlantic salmon aquaculture industry attributes the majority of disease-specific mortality to salmonid rickettsial septicaemia (SRS), caused by *Piscirickettsia salmonis* (Maisey et al., 2017). Winter ulcers, caused by *Moritella viscosa* represents the greatest bacterial threat to fish health in Norwegian aquaculture (Sommerset et al., 2024). Finally, the prevalence of disease in aquaculture stocks raises conservation concerns about the potential transmission between farmed salmon and sympatric wild populations (Mordecai et al., 2021). Additionally, disease in aquaculture, and the associated impacts on mortality, fish welfare and the environment, is an emotive issue that is often negatively portrayed in news media, damaging the public perception of the industry.

There has been much effort to reduce disease-specific mortality in salmonid aquaculture (Aunsmo et al., 2023) via various treatment methods. Chemical bath treatments with organophosphates, pyrethroids or hydrogen peroxide are effective at removing fungal and parasitic pathogens (Overton et al., 2019), but cause stress, are costly to administer, and their over-use has led to parasite resistance (Aaen et al., 2015). The use of therapeutic antibiotics for treatment and prevention of aquatic diseases is widespread throughout the aquaculture sector (Lulijwa et al., 2020), however, concerns about antimicrobial resistance (Mog et al., 2020; Schar et al., 2021) and the impact of antibiotic residues on the surrounding environment (Adenaya et al., 2023) have instigated research to identify antibiotic alternatives (Bondad-Reantaso et al., 2023). For example, immune system enhancement through functional feeds or immunostimulant injection has been shown to increase immune function (Tacchi et al., 2011; Kibenge et al., 2012). Recent non-harmful developments in sea cage design and incorporation of physical barriers have been shown to be effective at preventing parasitic infection from sea lice (Stien et al., 2016).

Vaccines for farmed finfish bacterial diseases were introduced in Norwegian salmon aquaculture in the 1980s, drastically reducing antibiotic use and providing long-term resistance to many prolific pathogens (Sommerset et al., 2005). Today, vaccines are an important disease control measure employed in all countries with salmon farming operations and are key to the sustainability of the industry. Particularly effective against bacterial pathogens, administering vaccinations against diseases such as furunculosis (*Aeromonas salmonicida)* and cold-water vibriosis (*Vibrio salmonicida)* has become common practice (Brudeseth et al., 2013) although some intracellular bacterial pathogens have proven to be difficult to protect against (e.g. SRS; Happold et al., 2020). A number of vaccines targeting viral pathogens are available, but these are generally less efficacious than those developed for bacteria (Kibenge et al., 2012, Pettersen et al., 2015b). Whilst ubiquitous and essential for successful aquaculture, vaccines require significant investment to develop and rely on the adaptive immune system of the fish, which remains poorly understood (Ye et al., 2011), making the

production of vaccines for all pathogens a challenging venture (Brudeseth et al., 2013).

There is great potential for genetic improvement in aquaculture species (Gjedrem et al., 2012; Gjedrem & Rye, 2018; Houston et al., 2020) and selective breeding has proven an important tool for disease control. For example, marker-assisted selection (MAS) for a quantitative trait locus (QTL) offering resistance to infectious pancreatic necrosis virus (IPNV; Houston et al., 2010) greatly reduced the prevalence of the disease in Norwegian salmon aquaculture in two generations (Norris, 2017). The integration of genomic technologies into the salmonid aquaculture sector, particularly the uptake of MAS and genomic selection, have improved genetic resistance to disease (Houston, 2017; Houston et al., 2020). Additionally, gene editing for disease resistance is considered to be of significant potential for managing aquatic diseases (Gratacap et al., 2019). However, for such practices to be effective, they rely on access to a high-quality, annotated reference genome, and increasingly on understanding the functional genomic basis for disease resistance and immune system traits.

1.4.4 The AQUA-FAANG Consortium

Following in the footsteps of ENCODE, FAANG was launched as a community effort to annotate functional genomic elements in terrestrial farmed animal species (Archibald et al., 2015; Clark et al. 2020; Section 1.3). An initiative with related aims was later launched for salmonids, called Functional Annotation of All Salmonid Genomes (FAASG) (Macqueen et al. 2017). FAASG was later superseded by AQUA-FAANG (https://www.aqua-faang.eu/), a funded European consortium that produced genome functional annotations for six major aquaculture species, including two key salmonids (Atlantic salmon and rainbow trout), in addition to European seabass, gilthead seabream, common carp and turbot (Johnston et al., 2024). AQUA-FAANG was one of several farmed animal genome functional annotation projects funded in Europe, grouped under the name of EuroFAANG (https://eurofaang.eu/), an initiative that went on to be funded as a formal European Research Infrastructure project.

AQUA-FAANG generated 3,890 functional annotation datasets in common across all six species; 2,183 from a panel of tissues from sexually mature and immature individuals of both sexes (BodyMaps), 1,018 from head kidney tissue from fish challenged with bacterial and viral immunostimulants (ImmunoMaps), and 689 representing key stages of embryogenesis (DevMaps). Three functional annotation assays were used in AQUA-FAANG; ATAC-seq to determine chromatin accessibility, ChIP-seq with different histone marks to identify regulatory elements including enhancers and promoters, and short-read RNA-seq to examine coding and non-coding gene expression.

In this PhD project, I developed and report a full-length RNA-seq method using Atlantic salmon DevMap and ImmunoMap samples provided by AQUA-FAANG. Whilst not directly included within the AQUA-FAANG project (Figure 1.5), my project has benefitted through funding, samples and collaboration with the AQUA-FAANG team (see Declaration).

1.5 Atlantic Salmon Embryogenesis

The AQUA-FAANG consortium identified embryogenesis as a key period of ontogeny for functional annotation, where the maternal to zygotic transition is followed by the development of cells and tissues that ultimately dictate the adult body plan, influencing many aquaculture production traits (Zhang et al., 2019). It has been shown that transcript diversity is important for normal embryogenesis, with extensive alternative splicing observed in mouse embryos (Revil et al., 2010) and differential splicing occurring at successive stages of embryogenesis in zebrafish (Liu et al., 2022c). However, most knowledge about embryonic development has been established in mammals and model species such as zebrafish. As a result, the molecular regulation of embryonic development in Atlantic salmon is poorly understood. One of the aims of my PhD project (section 1.7) was to carry out long-read RNA-seq over an embryonic timecourse covering stages from late blastulation to the beginning of organogenesis, as defined by the AQUA-FAANG project. As context to this work, I below introduce the current state of knowledge on finfish embryogenesis.

1.5.1 Zygote Formation and Blastulation

The process by which a multicellular organism develops from a simple zygote to having a complex body plan with fully differentiated tissues and diverse cell types is called embryogenesis. In finfish, this process starts with the fusion of egg and sperm to form a zygote. Then, rapid cell proliferation occurs as the zygote experiences several rounds of synchronised cleavage that divide the zygote into several blastomeres (Gorodilov, 1996). This process is called blastulation and culminates in the formation of a conglomeration of a few thousand cells called the blastula. The initial cleavages and resulting formation of the blastula are regulated by maternal transcripts and proteins (Lee et al., 2014)

The blastulation period is characterised by the desynchronisation of cell cleavage and elongation of the cell cycle (Kane & Kimmel, 1993), as well as the cessation of maternal transcriptional control and activation of the zygotic genome (Jukam et al., 2017). This process is called the mid-blastula transition or maternal-to-zygote transition. From studies conducted in zebrafish, it has been shown that zygotic genome activation (ZGA) begins just prior to mid-blastulation, with a distinct wave of zygotic gene expression (Vesterlund et al., 2011; Heyn et al., 2014), followed by the main activation of the zygotic genome and the assumption of full zygotic control by late blastulation (Kane and Kimmel, 1993; Wragg and Müller, 2016). Inhibition of ZGA causes continued division and cleavage, but failure to undergo gastrulation (section 1.5.2) (Kane, 1998). Thus, ZGA is a key process during blastulation, essential for further development of the embryo.

1.5.2 Gastrulation

Next, the blastula undergoes gastrulation; embryonic cells envelop the yolk sac and begin to differentiate into distinct lineages. Starting at mid-epiboly, epiboly, the spreading and thinning of the ectoderm (Panousopoulou et al., 2016), continues throughout gastrulation, ceasing when blastoderm cells (epithelial layer surrounding the blastocoel) completely cover the yolk (Kimmel et al., 1995). In fish, cells forming the precursors of the mesoderm, the layer of cells which produces bones, muscles and the circulatory system,

and endoderm, the layer of cells which gives rise to the gastrointestinal tract (Pinheiro & Heisenberg, 2020), are internalised by involution, forming the hypoblast which is a layer of cells that help determine the embryo body axes (Rohde & Heisenberg, 2007). The convergent extension of cells occurs simultaneously to involution and further rearranges these two dermal layers forming progenitors to structures such as somites and the notochord (Kimmel et al., 1990). Gastrulation results in the formation of three layers of germ cells (Solnica-Krezel, 2006) which form the basis of the body axes (Kimmel et al., 1995; Pinheiro and Heisenberg, 2020).

1.5.3 Somitogenesis

Segmentation follows gastrulation. This stage of embryogenesis, termed somitogenesis, gives rise to primary organ precursors, and the embryo begins to elongate (forming a tail bud), while a rudimentary vascular system is established (Kimmel et al., 1995). The presomitic mesoderm formed during gastrulation undergoes segmentation in an anterior-to-posterior direction to form symmetrical pairs of somites positioned either side of the neural tube and notochord (Dequéant & Pourquié, 2008). This process is precisely timed and conserved between species, governed by ‘segmentation clock’ genes belonging to the Notch pathway (Palmeirim et al., 1997; Jiang et al., 2000). As somite pairs are segmented and boundaries established, they immediately start to differentiate, directed by signals provided by surrounding tissue (Yusuf & Brand-Saberi, 2006). Up to 67 somite pairs are observed in Atlantic salmon embryogenesis, with early anterior somites differentiating into optic structures and forming the primordial kidney, heart and gill tissue (Gorodilov, 1996) which continue to develop throughout somitogenesis. Towards the end of Atlantic salmon somitogenesis, segmentation slows, weak muscular myotome contractions can be recorded and the body plan can be clearly seen (Gorodilov, 1996).

1.5.4 Pharyngula Stage

The final stage of embryogenesis is the pharyngula stage or late-eyed stage. This is the considered to be the most conserved period of ontogeny across vertebrates (Irie & Kuratani, 2011). In zebrafish, caudal and pectoral fins

begin to form during the pharyngula period, the embryonic eye becomes pigmented and the formation of the jaws and operculum commences (Kimmel et al., 1995). In salmon, the vitelline plexus begins to grow, eventually covering the entirety of the yolk sac, thus finishing yolk sac vascularisation (Gorodilov, 1996) and further development results in the connection of blood vessels and the formation of the circulatory system. In zebrafish, caudal fin rays and gill filaments begin to develop, pigmentation of the eye progresses and melanophores emerge in the trunk section of the body (Kimmel et al., 1995). Muscular contractions increase, which cause thrashing movements of the embryonic body and the mandibular and hyoid arches are evident, forming the jaw. The embryo has now finished embryogenesis and is ready to hatch, into an alevin in the case of salmonids.

1.5.5 Transcriptional Regulation of Embryogenesis

The differentiation of multiple distinct cell types from a single totipotent cell is a result of significant transcriptional regulation (Farrell et al., 2018; Cao et al., 2019), alongside post-transcriptional regulatory mechanisms. Many unique genes are switched on during early development that are not expressed outside of embryogenesis (Graveley et al., 2011; Xue et al., 2013; Junker et al., 2014; Kleppe et al., 2015; Bayega et al., 2021). Whilst Gorodilov (1996) described the stages of embryogenesis, transcriptional regulation of salmonid embryogenesis remains relatively unexplored, with the majority of knowledge based on studies conducted using zebrafish (e.g. White et al., 2017a) and medaka (Li et al., 2020b). At the transcriptome-level, embryonic timecourse studies are rare in salmonids, and often only focus on a single stage of development or focus on broader life stages such as embryo-to-alevin-to-fry rather than delving into embryonic development (e.g. Abdellaoui & Kim, 2024).

Until the AQUA-FAANG project, there was no embryonic timecourse research investigating molecular regulation of embryogenesis in salmonids. AQUA-FAANG addressed this gap using ATAC-seq, ChIP-seq and short-read RNA-seq. My PhD research builds on AQUA-FAANG through the elucidation of RNA transcript-level regulation, providing significant additional

information and improvements to the functional annotation of the Atlantic salmon genome.

1.6 Salmon Immune Responses

Aquatic diseases represent the biggest challenge to the sustainable growth of the aquaculture sector (section 1.4.3). Given the prevalence of diseases in salmonid aquaculture, there is a drive to understand immune function in these species. Sharing many of the same building blocks of the immune system with mammals, teleosts possess both an innate and adaptive immune response (Rauta et al., 2012). The adaptive immune response is similar to mammals, involving the major histocompatibility complex (MHC), T-cells, B-cells and antibody affinity maturation (Bruce & Brown, 2017), however, in poikilothermic fish, the adaptive immune response is slow and has limited antibody repertoire (Magnadóttir, 2006; Magor, 2015). As such, finfish rely on the innate immune response as their initial defence system against the majority of aquatic pathogens (Magnadóttir, 2006). The AQUA-FAANG consortium (section 1.4.4) aimed to produce functional annotation maps for the innate response to viral and bacterial pathogens using short-read RNA-seq, ChIP-seq and ATAC-seq.

The role of transcript diversity in immune response has been explored previously outside salmonids. For instance, alternative splicing is a key mechanism by which transcriptional control is exercised over viral pathogens (Liao & Garcia-Blanco, 2021). Further, alternative splicing was reported in response to dengue virus vaccination in humans (Kim et al., 2022), while preferential transcript isoforms were expressed in granulomas associated with tuberculosis (Carow et al., 2019). Considering the recognised importance of transcript diversity for immune system function, I carried out long-read RNA-seq on the AQUA-FAANG ImmunoMap samples from Atlantic salmon to examine transcript-level expression during the innate immune response (section 1.7). As background to this work, I introduce the early stages of both antiviral and inflammatory/antibacterial responses below.

1.6.1 Salmonid Immune Organs

Whilst the cellular basis of immune function is conserved between mammals and teleosts - both possess innate and adaptive immunity and employ populations of immune cells including lymphocytes, monocytes and macrophages (Flajnik, 2018) – the most drastic difference exists when examining immune organ arrangement (Bjørgen & Koppang, 2022). The spleen is considered to be a secondary immune organ in teleosts (Flajnik, 2018) involved in antigen concentration and antibody affinity maturation during the adaptive immune response (Fu et al., 2022). In salmonids, the liver produces acute phase proteins during the innate immune response (Bayne & Gerwick, 2001) and harbours many immune cells (Castro et al., 2014; Taylor et al., 2022). The Atlantic salmon intestine also has immune function, harbouring potential lymphatic and mucosal system components, and possessing immune activity second only to spleen and head kidney (Kortner et al., 2024).

The head kidney is the primary haematopoietic organ in fish, acting as a mammalian bone marrow equivalent, providing the primary source of immune cells from undifferentiated stem cells (Press & Evensen, 1999; Zapata & Amemiya, 2000; Løken et al., 2020; Bjørgen & Koppang, 2022). Due to its primary immune functions, the AQUA-FAANG consortium selected the head kidney to generate maps of the functional elements associated with the innate response to bacterial and viral immunostimulation.

1.6.2 Pathogen Recognition

Before pathogens stimulate a cellular response, they must first overcome the physical barrier provided by skin and mucus of fish (Rajme-Manzur et al., 2021). Contained within the mucus are several innate molecules including mucins, lysozymes and antimicrobial peptides showing pathogen killing activity (Fast et al., 2002; Cain & Swan, 2010; Nimalan et al., 2022).

Once pathogens have penetrated physical and mucosal barriers, the pathogen is recognised as non-self by germ-line encoded pattern recognition receptors (PRR; Thompson et al., 2011; Liao & Garcia-Blanco, 2021). PRRs interact with pathogen-associated molecular patterns (PAMPs) such as

bacterial peptidoglycans or viral RNAs and initiate a range of non-specific responses dependent on pathogen class (sections 1.6.3 and 1.6.4). Toll-like receptors (TLRs) are the most well-known PRR family, highly conserved between vertebrates (Aderem & Ulevitch, 2000), and mainly responsible for extracellular PAMP detection, acting as principal inducers of the innate immune system in finfish (Whyte, 2007). Other PRRs exist in teleosts; NOD-like receptors are conserved with mammals (Van der Vaart et al., 2012), whist retinoic acid-inducible gene I (RIG-I)-like receptors are key components of teleost viral recognition (Chen et al., 2017). PRRs initiate both antimicrobial and antiviral responses at the cellular level.

1.6.3 Cellular Innate Antiviral Response

In vertebrates, including teleosts, viral pathogen defence is driven by interferons (IFN; Gan et al., 2020), specialised cytokines which stimulate the production of interferon-stimulated genes (ISGs; Clark et al., 2023). Unlike mammals, salmonids only possess type-I and type-II interferons (Liu et al., 2020), lacking IFN-III (involved in mammalian antiviral responses) (Mesev et al., 2019). Type-I IFNs are involved in the innate antiviral response. IFN-I is produced after recognition of viral pathogens by PRRs (section 1.6.1) activate interferon receptors, initiating a downstream signalling cascade via the Janus kinase/signal transducers and activators of transcription (JAK/STAT) pathway (Rawlings et al., 2004; Zou & Secombes). The JAK/STAT pathway ultimately leads to production of the IFN-stimulated gene factor 3 (ISGF3) complex, which binds to the promotor regions of many ISGs, thus initiating their production (Schoggins et al., 2011; McNab et al., 2015). Unlike type-I IFNs, which can be produced by the majority of nucleated cells, type-II IFNs also known as IFN-γ are produced exclusively by immune cells (Schroder et al., 2004) and play a key role in adaptive cell-mediated immunity (Robertsen, 2006).

Polyinosinic:polycytidylic acid (poly I:C) was the agent used by AQUA-FAANG to simulate viral infection. Poly I:C is an immunostimulant that reacts with toll-like receptor 3 and has been demonstrated to induce the production of type-I IFNs thus leading to a robust innate viral response in the host

(Fortier et al., 2004; Matsumoto & Seya, 2008). Since this discovery, poly I:C has been commonly used to simulate viral infection (Andresen et al., 2020; Komal et al., 2021).

1.6.4 Cellular Antibacterial Response

Teleosts have a host of antibacterial defence mechanisms involving non-specific and specific factors. Since the aim of my project was to study the innate response, only non-specific antibacterial machinery is discussed here. Antimicrobial peptides (AMPs) are an important component of the innate response, produced in response to bacterial infection in teleosts (Valero et al., 2020). Triggered by bacterial pathogens, transcription factor nuclear factor κB (NF- κB) is activated, translocates to the nucleus and leads to the production of pro-inflammatory molecules including AMPs (Gilmore & Wolenski, 2012). The pro-inflammatory response is a strong anti-microbial mechanism involving production of cytokines like TNF-α (Sherwood & Toliver-Kinsky, 2004) that is rapidly elicited in response to bacterial infection (Sun et al., 2022). Expression of AMPs in fish is regulated by a host of different genes including β-defensins (Harte et al., 2020) and piscidin genes (Raju et al., 2021). Another group of AMPs called hepcidins also display powerful antimicrobial activity in teleosts (Álvarez et al., 2014). AMPs combat bacteria by a number of different mechanisms including inhibiting cell wall synthesis, targeting bacterial cell membranes as well as intracellular activities like inhibition of DNA and protein synthesis (Nicolas, 2009).

Other non-specific antibacterial elements include lysozyme expression, and lectins. These responses are both involved in mucosal immunity (section 1.6.2) but are also present in lymphoid tissue, liver and blood plasma (Saurabh & Sahoo, 2008). Lysozymes act on the peptidoglycan layer of the bacterial wall leading to reduced cell wall integrity and eventual bacterial lysis (Song et al., 2021) whilst lectins are potential opsonins of bacteria stimulating increased bacterial uptake and subsequent destruction by macrophages (Ewart et al., 1999).

The AQUA-FAANG used an inactivated strain of *Vibrio anguillarum,* the causative agent of vibriosis (Frans et al., 2011; Lages et al., 2019) to

stimulate an antibacterial proinflammatory response. *V. anguillarum* was chosen as it is a widely-spread pathogen that would elicit a strong proinflammatory response for all species in the AQUA-FAANG project (Frans et al., 2011).

1.7 Project Objectives

The aim of this project was to develop an ONT-based full-length RNA-seq method, including robust analysis pipelines, to uncover transcript-resolved expression dynamics associated with innate immunity and embryogenesis, and to improve genome functional annotation, in Atlantic salmon.

The specific objectives were:

1) To develop a full-length RNA-seq method applied to AQUA-FAANG DevMap and ImmunoMap datasets, generating the first full-length transcriptome for Atlantic salmon using ONT-based RNA-seq. This objective is addressed in Chapter 2.

2) To establish robust bioinformatic pipelines to conduct quantitative analysis of transcript-specific expression in response to viral and bacterial immunostimulation in Atlantic salmon. This objective is addressed in Chapter 3.

3) To elucidate transcript expression dynamics employed during Atlantic salmon embryogenesis. This objective is addressed in Chapter 4.

4) To investigate the high prevalence of mono-exonic transcript models discovered by full-length RNA-seq in an attempt to better understand the biological relevance of these transcripts. This objective is addressed in Chapter 5.

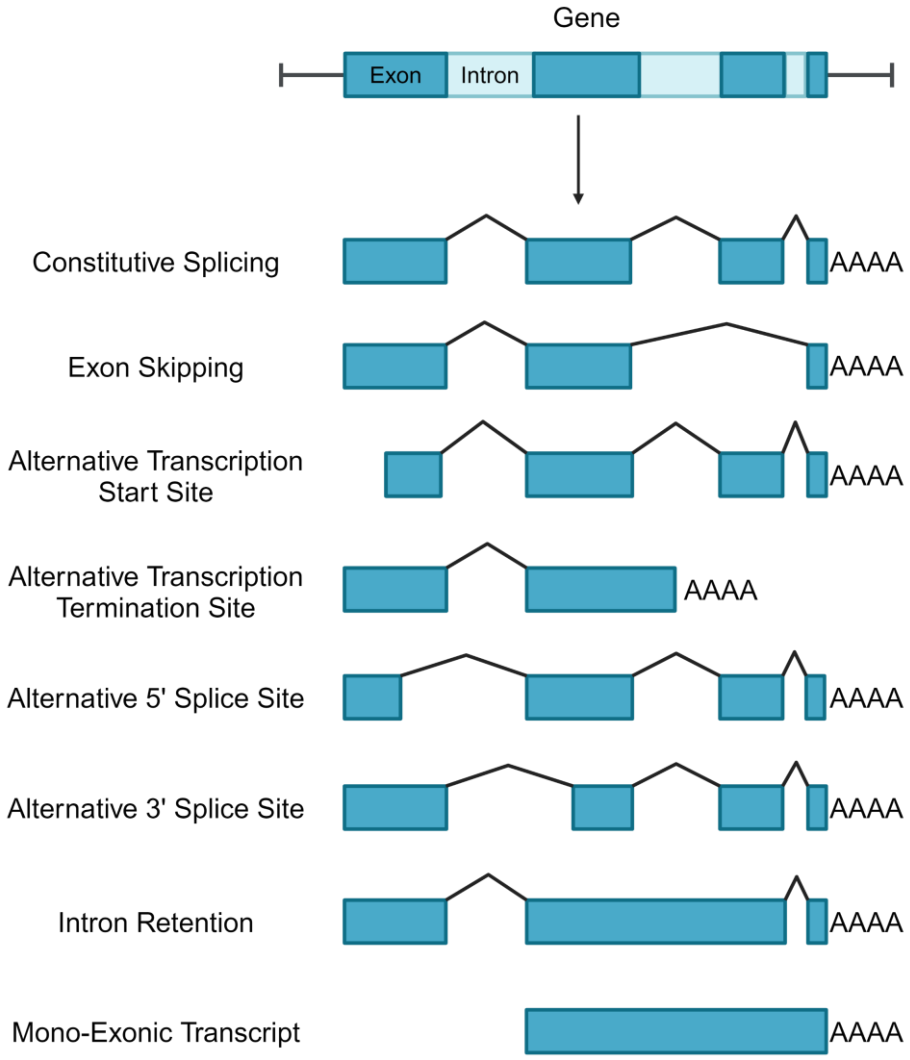


Figure 1.1: Visual summary of the types of transcript diversity arising through a range of transcriptional processing modifications. Created with BioRender.com. Constitutive results in an mRNA molecule with all possible exons included. Exon skipping occurs when mRNA processing excises one or more exons without affecting the splice sites of the other exons. Alternative transcription start and termination sites are alternative splice sites at the 5' or 3' terminus respectively. Alternative splice sites also occur within the mRNA molecule itself leading to different exon start and end sites. Intron retention occurs when introns are not excised by post-transcriptional machinery. A mono-exonic transcript is that which is only formed by a single exon, either through intron retention events leading to an unspliced transcript which contains introns, or through transcription of an intronless gene. PolyA tails are added to the mRNA transcripts at the end of post-transcriptional modification.

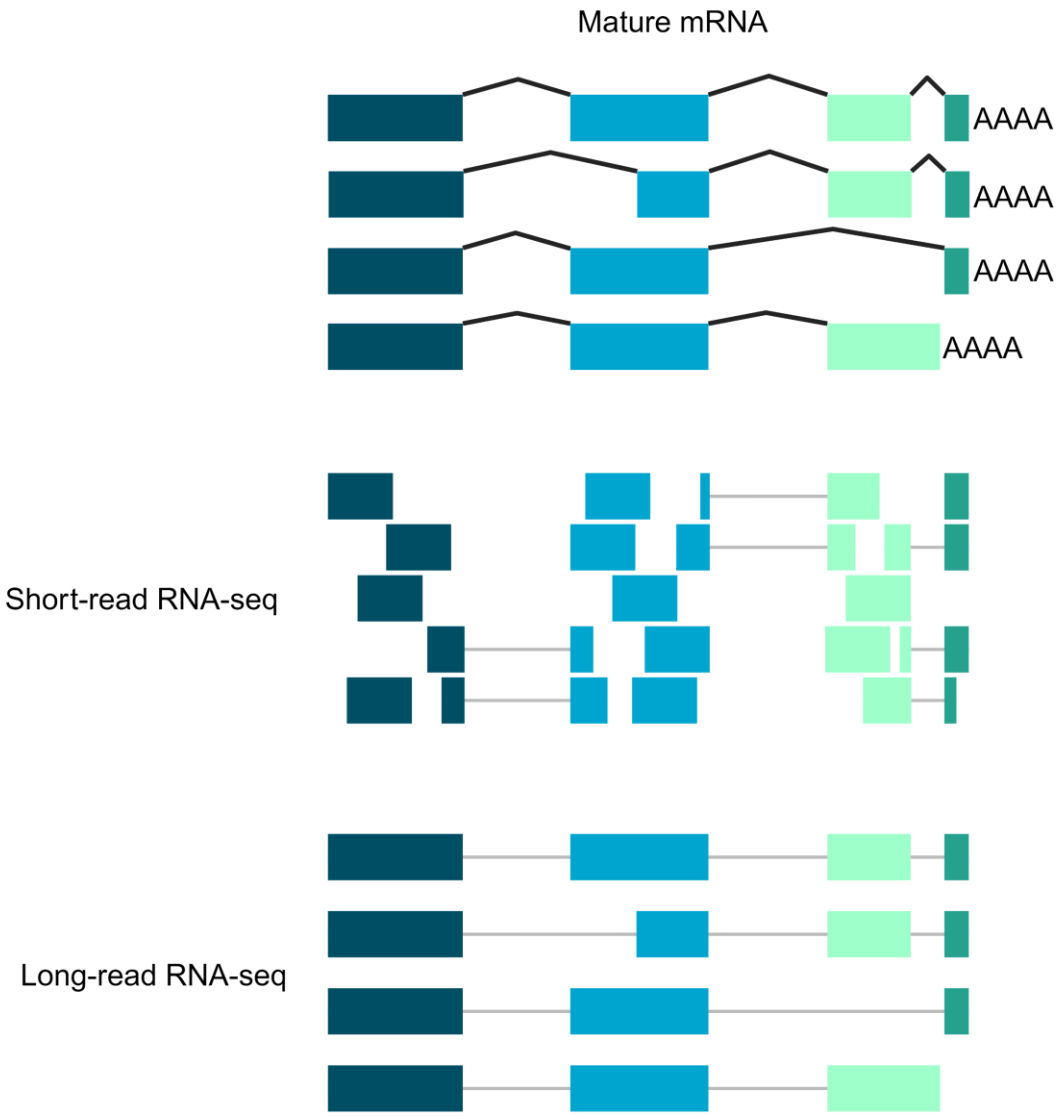


Figure 1.2: Comparison of reads produced by short-read and long-read RNA-seq. Short-read technology produces short reads of set length. In some cases, these short reads span introns and link exons together, however, no single read will encapsulate the entire length of most mRNA transcripts. Long-read sequencing captures the full-length of mRNA transcripts in a single sequencing pass. Created with BioRender.com.

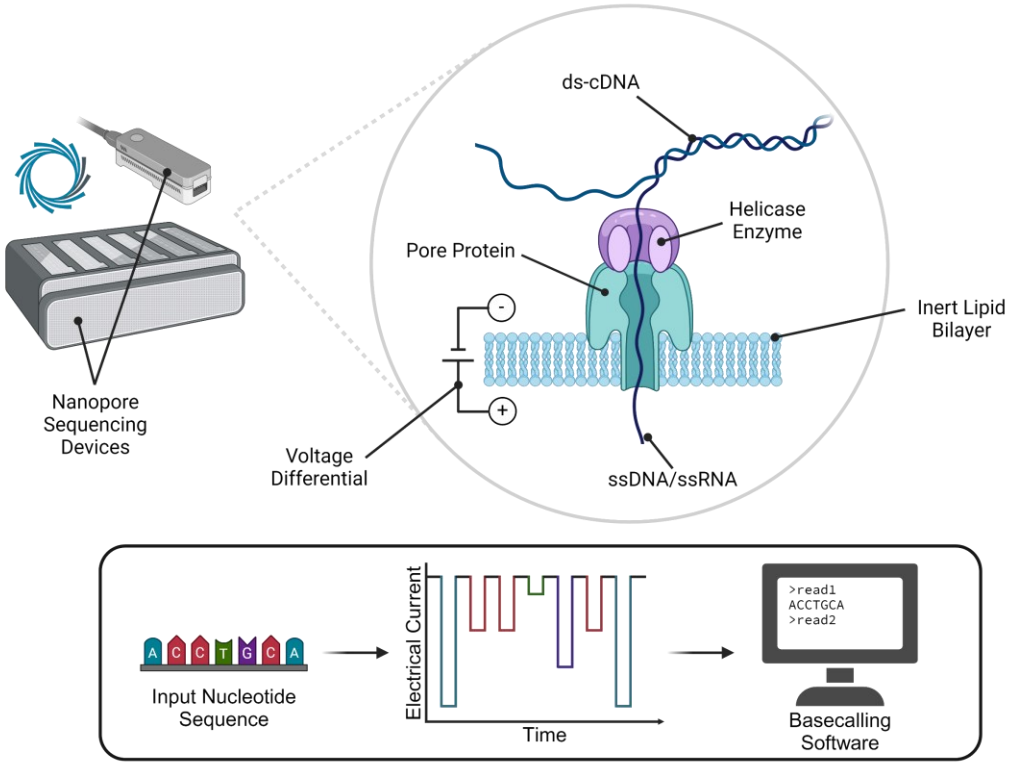


Figure 1.3: Schematic diagram depicting sequencing on ONT platforms. Double-stranded nucleotide molecules (genomic DNA, cDNA or an RNA-cDNA hybrid), are loaded into a flowcell consisting of two chambers separated by an inert lipid bilayer, each containing ionic solution. Embedded within the bilayer are pore proteins that form channels through which the nucleotides can pass. To start sequencing, an electrical current differential is created between the two chambers, which causes the nucleotides to pass through the pores. Nucleotides are unwound by a helicase before entering the pore, meaning a single strand is sequenced in any pass. The electrical current is measured across each pore throughout sequencing. Each nucleotide bases causes a distinct disruption to the electrical signal which is recognised by basecalling software. Created with BioRender.com.

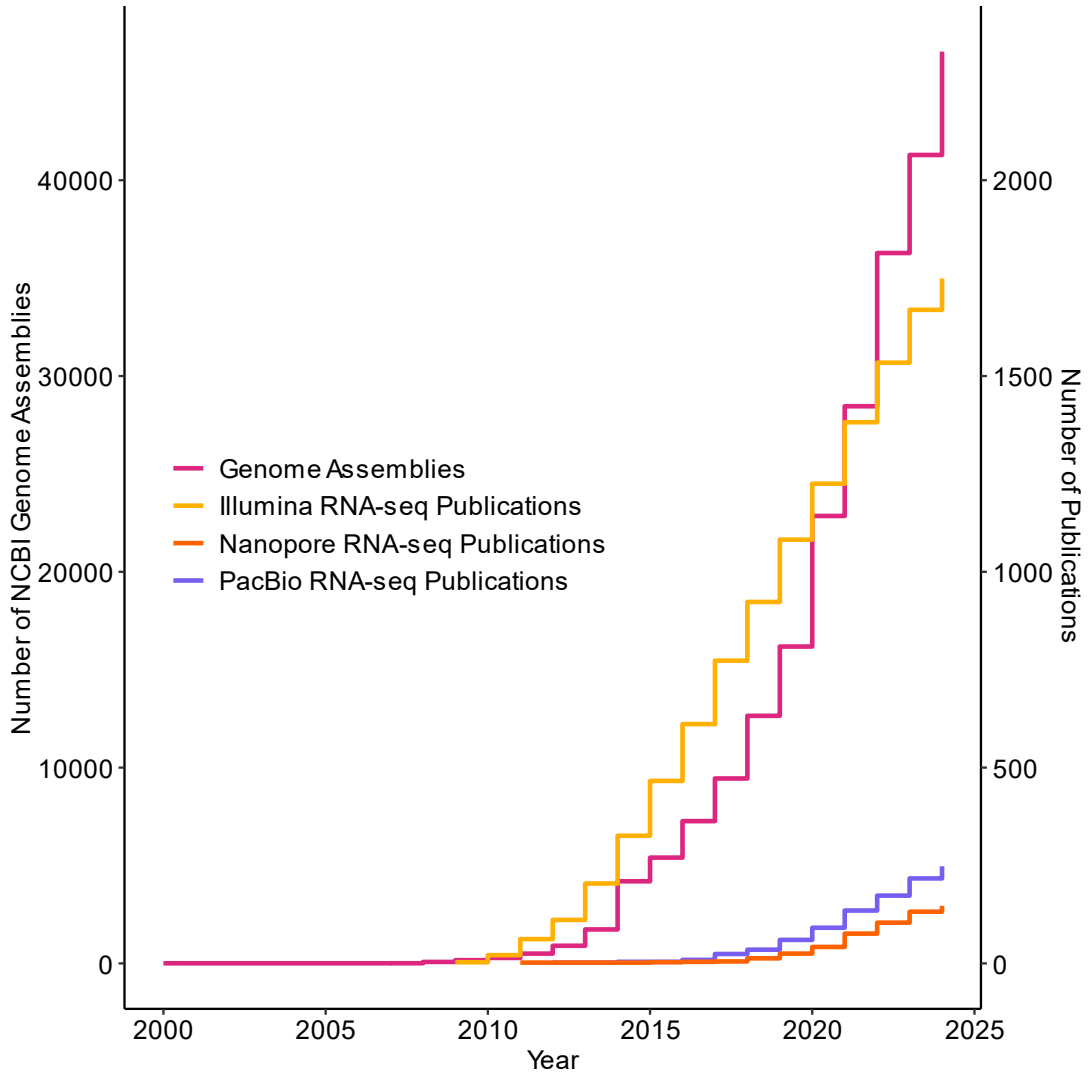


Figure 1.4: Line plot showing cumulative increase in eukaryotic reference genome assemblies through time in pink. Yellow, purple and orange lines represent the number of publications matching search criteria for Illumina, Nanopore and Pacbio RNA-seq. Web of Science was used to search abstracts for these exact terms (08/09/2024): "Illumina + RNA-seq", "Nanopore + RNA-seq", "PacBio + RNA-seq".

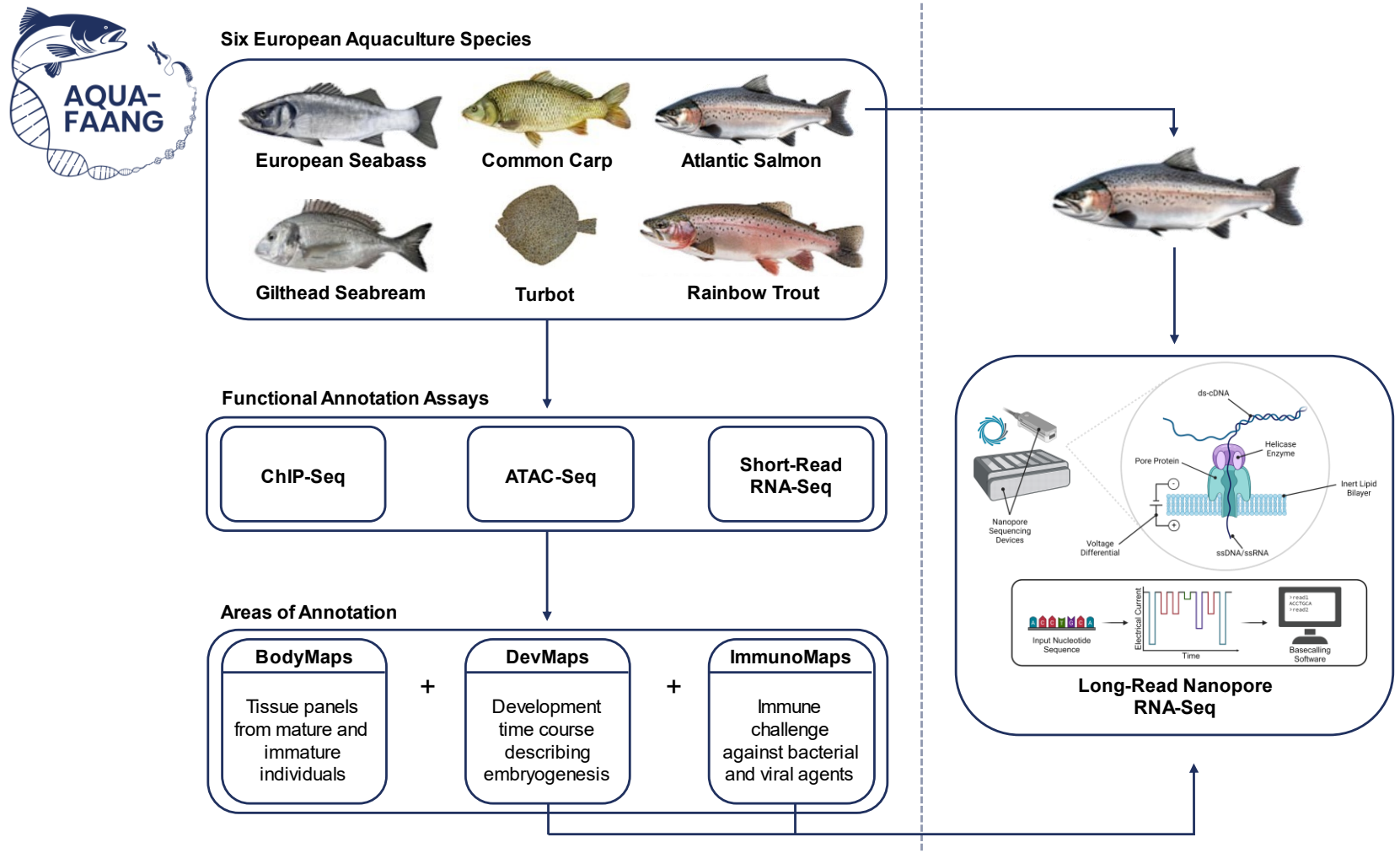


Figure 1.5: Schematic of the relationship between my PhD project and the AQUA-FAANG consortium. Adapted from Johnston et al. (2024).

Chapter 2: Long-Read Nanopore Transcriptome Assembly

Summary

There are currently no published Nanopore long-read transcriptome assemblies for Atlantic salmon. This chapter describes the development of a full-length RNA-seq method and long-read transcriptome assembly pipeline using the Oxford Nanopore Technologies (ONT) platform. Atlantic salmon embryo and head kidney samples from the AQUA-FAANG project were used to generate multiplexed cDNA sequencing libraries using the direct cDNA sequencing kit from ONT. I describe the creation of a bespoke bioinformatic pipeline to process the long-read data and construct a high-confidence long-read transcriptome consisting of 35,480 unique genes and 266,222 transcripts. This transcriptome is used as the basis for subsequent chapters in this work.

2.1 Introduction

Genomic advancements over the last 20 years have resulted in the publication of high-quality reference genome sequences for many model and non-model organisms (Pareek et al., 2011; Mardis, 2011; da Fonseca et al., 2016). Recent efforts have aimed to decipher this information by identifying which regions of the genome have biological relevance, including in non-model farmed animal species (Giuffra et al., 2019; Johnston et al., 2024) via functional annotation.

RNA-seq on SGS platforms is the primary current technique for studying the transcriptome, the sum of all transcripts produced by a cell, tissue or organism at a given time. Defining and characterising the transcriptome allows us to identify which regions of the genome are expressed, what genomic products they produce (e.g. non-coding RNAs, proteins, etc), and subsequently better understand the function of these expressed regions. With this information, we can dissect the fundamental pathways and mechanisms associated with organism function, including growth and development, cell fate, and disease progression, allowing us to make more valid links between genotype and phenotype.

As discussed in Chapter 1, long-read RNA-seq technologies such as those offered by PacBio or ONT are powerful methods for discovering alternative transcripts (Kuo et al., 2020) and identifying alternative TSSs and TTSs (Byrne et al., 2017), leading to improved transcriptome characterisation (Weirather et al., 2017). For example, long-read RNA-seq has revealed previously undescribed transcript diversity in many terrestrial agriculture species such as cow (Halstead et al., 2021 [ONT]; Ren et al., 2023 [PacBio]), pig (Beiki et al., 2019 [PacBio]; Müller et al., 2021 [ONT]) and chicken (Kuo et al., 2017 [PacBio]; Guan et al., 2022 [ONT]), leading to a significant increase in our understanding of transcriptome diversity in farm animals.

However, long-read transcriptomics is yet to be widely applied in non-model fish, including farmed species, with a handful of studies currently published. Nonetheless, the limited work done to date highlights the power of this approach. Using the PacBio Iso-Seq method, Huang et al. (2022) described 1,684 new genes and 60,476 transcripts, of which 71% had evidence of a novel splice site, expressed in turbot (*Scophthalmus maximus*) following bacterial infection. PacBio was used to describe 10,640 unannotated transcripts in rainbow trout (*Oncorhynchus mykiss*) including a novel exon-skipping event in the *gvin1* gene, with individuals expressing this isoform showing increased susceptibility to bacterial cold-water disease (Ali et al., 2021). In Atlantic salmon, >40,000 transcripts containing a novel splice event were identified using PacBio sequencing of gill, liver and head kidney tissues sampled during smoltification (Ramberg et al., 2021). Ramberg and colleagues (2021) focussed on immune organs during smoltification which leaves other important life cycle stages and other tissue types of interest such as brain and skin unstudied by long-read technologies. In all cases, long-read RNA-seq identified novel transcript information, highlighting its potential to improve transcriptome annotations and advance genomic resources for aquaculture species.

There are only a handful of ONT-based RNA-seq studies in finfish, including an analysis of polyA tail length in zebrafish (Begik et al., 2023) and transcriptome assemblies of freshwater species *Danionella translucida*, a transparent cyprinid (Kadobianskyi et al., 2019), and angelfish *Pterophyllum*

scalare (Madireddy, 2024). There are currently no published studies using ONT-based RNA-seq in salmonids. In Atlantic salmon, all annotated gene models in the current Ensembl reference annotation (Ssal_v3.1; Lien et al., 2016) are supported by short-read RNA-seq data with support from long-read data where available. Additionally, the genome of Atlantic salmon possesses its own challenges for RNA-seq; a high proportion of repeats (Lien et al., 2016), and numerous retained duplicated genes following a recent whole genome duplication event (Macqueen & Johnston, 2014) present challenges for short-read RNA-seq (Houston & Macqueen, 2019). Long-read RNA-seq may be required to resolve these complex regions to capture more information on transcript diversity in Atlantic salmon.

This chapter describes the generation of ONT-based RNA-seq data using Atlantic salmon samples representing; 1) six key stages of embryonic development, and 2) head kidney following *in vivo* immune stimulation with polyinosinic:polycytidylic acid (poly I:C) or inactivated *Vibrio*, respectively (Figure 2.2). Using these datasets, I report a novel bioinformatic pipeline to assemble a full-length transcriptome for Atlantic salmon, used as the reference for analyses detailed later in the thesis.

2.2 Materials and Methods

2.2.1 Head Kidney Samples

Atlantic salmon head kidney samples were collected from immature parr that had been immunologically challenged with one of a PBS control, poly I:C or inactivated *Vibrio* (n=6) via intraperitoneal injection (IP). Poly I:C is an immunostimulant that reacts with toll-like receptor 3 and is commonly used as a viral PAMP that elicits a strong antiviral response (Fortier et al., 2004; Andresen et al., 2020). An inactivated strain of *Vibrio*, the causative agent of vibriosis in salmonid fishes (Frans et al., 2011) was used to stimulate the immune system with bacterial PAMPs. The methods used for the poly I:C challenge group were described in Clark et al. (2023), whilst *Vibrio* challenges are described in a PhD thesis (Naseer, 2023). I outline these approaches below for ease of reading.

Fish were kept at the University of Aberdeen (zoology building) and all procedures described were carried out in compliance with the Animals (Scientific Procedures) Act 1986 under UK Home Office licence PPL number 70/8071 (to Professor Sam Martin) and approved by the ethics committee at the University of Aberdeen. All fish were healthy and monitored throughout the study. Fish challenges were administered by Dr Shahmir Naseer, Dr Thomas Clark and Prof. Samuel Martin.

Poly I:C (Sigma #P1530) diluted to 5mg/ml in PBS was heated to 55°C for 15 minutes and cooled to room temperature before use. A *Vibrio* strain P0382 was cultured in tryptic soy broth medium until it reached an OD600 of 1.5. A bacterial pellet from 10mL of culture was resuspended in NaCl (9g/L) and washed a further four times before being resuspended in 1mL of the same NaCl solution. The final resuspension was incubated at 100°C for 30 seconds to kill the bacteria, cooled to room temperature and diluted 1:400 before storage at -80°C. Aliquots of killed *Vibrio* were thawed and allowed to reach room temperature before use. Fish were challenged with 100µL of 1x PBS, poly I:C (500µg/fish) or killed *Vibrio* via intraperitoneal injection. All individuals were then kept in a single 400L freshwater tank maintained at 12°C, with a flow rate of 1000L/hour and a 12:12 light:dark photoperiod for 24 hours. After this incubation period, the fish were killed using a schedule 1 killing protocol. In brief, fish were given a lethal overdose of 2-phenoxyethanol (Merck, Sigma-Aldrich 77,699) at a concentration of 1.25mL/L, followed by destruction of the brain with a scalpel. Head kidney tissue was harvested from each individual immediately, then stored in 1.5mL RNA later at 4°C for 24 hours before storage at -80°C until RNA extraction.

2.2.2 Embryogenesis Samples

Samples for the embryonic timecourse dataset were collected by Dr Perojil-Morata as detailed in a PhD thesis (Perojil-Morata, 2024). Briefly, Atlantic salmon eggs were collected from a single female individual; milt was harvested from 3 males. Eggs and milt were transported to the Roslin Institute, University of Edinburgh on ice in oxygenated containers. The milt from each male was activated under a microscope - the individual with the

most motile sperm was used to inseminate all the egg samples in dechlorinated, oxygenated water. Eggs and milt were kept at 8°C in this way until the required harvest time post-fertilisation. n=3 replicates were taken from 6 key stages of development used by AQUA-FAANG: late blastulation (165 hours post-fertilisation [hpf]), mid gastrulation (214hpf), early somitogenesis (309hpf), mid somitogenesis (418hpf), late somitogenesis (552hpf), and late-eyed or pharyngula (792hpf).

Embryos were harvested according to an SOP: "Total RNA preservation from salmonid embryos using TRIzol" (accessible here: https://www.aqua-faang.eu/protocols.html). In brief, homogenizing beads were placed in 1.5mL Eppendorf tubes before 1mL TRIzol (Invitrogen #15596026) was added to the tubes. Embryos were punctured and dechorionated with watchmaking forceps before being added to the tubes. All samples were stored at -80°C until RNA extraction.

2.2.3 Total RNA Extraction – Head Kidney

Total RNA was extracted by Dr Shahmir Naseer from n=6 head kidney samples per each of the three treatment groups using a standard TRIzol method (Naseer, 2023). 20mg of tissue was homogenised in 1mL of TRIzol by ceramic bead beating using an MP-Bio FastPrep-24 5G tissuelyser. Total RNA was precipitated with 0.5mL of 2-propanol and pelleted before being washed with 75% ethanol and finally eluted in nuclease-free water. After receiving these RNA samples, I carried out several quality control steps before proceeding. Total RNA concentration was measured using a Qubit3 Broad Range (BR) RNA Assay (Invitrogen Q10210), RNA purity was evaluated using a Nanodrop device (Thermo Scientific, model ND-1000), and a TapeStation 4200 (Agilent Technologies) was used to determine RNA integrity. Total RNA with RIN >8.5 was used for ONT library preparation (mean RIN was 9.5 for samples used) (Table 2.1).

2.2.4 Total RNA Extraction - Embryos

I carried out total RNA extracted from *S. salar* embryos using a modified phenol-chloroform method. 3 embryos were pooled for each biological replicate and homogenised in 1mL TRIzol using the Qiagen TissueLyzer II

set to 30Hz for 5 minutes. The samples were incubated for 5 minutes at room temperature before 100µL 1-bromo-3-chloropropane was added. Tubes were shaken vigorously for 15 seconds and then incubated at room temperature for 30 minutes. Next, the samples were centrifuged for 15 minutes, at 4°C, at 20,000g, and the aqueous phase was carefully pipetted off. 200µL of precipitation solution (100µL isopropanol + 100µL solution of 1.2M NaCl, 0.8M sodium citrate sesquihydrate) was added to each sample and left to precipitate overnight at -20°C. Samples were pelleted by centrifuging at 20,000g, 4°C for 10 minutes, before being washed twice with 1mL freshly prepared 70% ethanol. Finally, the pellet was eluted in 20-50µL nuclease-free water depending on the size of the pellet.

Total RNA was assessed for purity, concentration and integrity as for the head kidney samples (Table 2.1).

2.2.5 mRNA Isolation

A Dynabeads mRNA Purification Kit (Invitrogen 61006) was used to isolate mRNA from total RNA for both the head kidney and embryo samples. 6µg of total RNA for each sample was input into the protocol. The manufacturer's instructions were followed except that the volume of Dynabeads magnetic beads was halved, whilst the rest of the reagents were used as per the following protocol. 50µL of Dynabeads per sample were transferred to a clean 0.2mL PCR tube and placed on a magnetic rack to pellet. The supernatant was removed and the beads were resuspended in 50µL of Binding Buffer from the Dynabeads kit. The beads were pelleted once again, and the supernatant was removed and resuspended in 50µL of Binding Buffer for a second time. 6µg of total RNA was diluted in nuclease-free water to a final volume of 50µL in a clean 0.2ml PCR tube, heated to 65°C for 3 minutes and placed on ice immediately after incubation. Each total RNA sample was added to the beads, flicked well to mix, centrifuged briefly and incubated at room temperature for 5 minutes. Next the tubes were flicked and spun once more to resuspend the beads before pelleting on a magnetic rack for at least 2 minutes. The supernatant was removed and beads resuspended in 50µL of Washing Buffer B, mixing thoroughly by flicking.

After a short centrifuge the tubes were returned to the magnetic rack and allowed to pellet for at least 2 minutes. The supernatant was removed and another round of Washing Buffer B was applied to the beads. After final aspiration of Washing Buffer B, the tubes were removed from the magnetic rack and the beads eluted in 11µL of Tris-HCl. Tubes were incubated at 72ºC for 2 minutes and then immediately placed on a magnetic rack for another 2 minutes. The purified mRNA was collected by transferring the supernatant to a clean tube and its quantity assessed using the Qubit3 with a HS RNA assay kit (Invitrogen Q32852).

2.2.6 Nanopore Library Preparation and Sequencing

A sequencing library was prepared for each set of samples using the cDNA Sequencing Kit SQK-DCS109 with the Native Barcoding 96 Kit EXP-NBD196 (Oxford Nanopore Technologies). Sequencing libraries were prepared in independent runs using the same approach (described below) for n=18 embryo samples (June 2021) and n=18 head kidney samples (July 2022).

2.2.6.1 Reverse Transcription and Strand-Switching

Reverse transcription and strand-switching was carried out using RNaseOUT (Invitrogen 10777019), Maxima H Minus Reverse Transcriptase (ThermoScientific 15259496), VNP and SSP primers (ONT) and 10mM dNTPs. For each sample, 2.5µL of VNP primer and 1µL of 10mM dNTPs was added to 100ng of mRNA eluted in 8µL of nuclease-free water. The reaction was mixed gently by flicking and spinning down with a brief centrifuge. The reaction was incubated at 65ºC for 5 minutes and then snap-cooled on a pre-chilled freezer block. During the incubation, a master mix was prepared consisting of 4µL 5x RT buffer, 1µL RNaseOUT, 1µL nuclease-free water and 2µL SSP primer. This 8µL master mix was added to the snap-cooled mRNA and mixed, before the reaction was incubated at 42ºC for 2 minutes. After this, 1µL of Maxima H Minus Reverse Transcriptase was added to the reaction and mixed, then incubated as follows; 42ºC for 90 minutes, 85ºC for 5 minutes, 4ºC hold.

2.2.6.2 RNA Degradation and Second Strand Synthesis

Remaining mRNA was degraded using RNase Cocktail Enzyme Mix (Invitrogen AM2286) and the cDNA purified using AMPure XP beads (Beckman Coulter A63881). 1µL of RNase Cocktail was added to the reverse transcription reaction in Section 2.2.6.2 and incubated for 10 minutes at 37ºC. This reaction was transferred to a clean Eppendorf LoBind tube and 17µL AMPure XP beads were added for a ratio of 1:1.2 beads to sample volume. The solution was incubated for 5 minutes on a rotary mixer at room temperature. Following the incubation, the tubes were placed on a magnetic rack and beads pelleted before the supernatant was aspirated without disturbing the beads. The beads were washed twice with 200µL of 70% ethanol and eluted in 20µL nuclease-free water for 10 minutes at room temperature on a rotary mixer. Beads were pelleted once again on the magnet and the elution transferred to a clean 1.5mL LoBind tube and placed on ice.

The second cDNA strand was synthesised using LongAmp Taq Master Mix (New England Biolabs M0287) and PR2 primer (ONT). For each sample, the following reaction was prepared in a 0.2mL PCR tube: 25µL 2x LongAmp Taq Master Mix, 2µL PR2 primer, 20µL reverse-transcribed sample above, and 3µL of nuclease-free water. The reaction was incubated at 94ºC for 1 minute, 50ºC for 1 minute, 65ºC for 15 minutes, then held at 4ºC. In a clean LoBind tube, 40µL of AMPure XP beads were resuspended by vortexing and the sample was added to the beads. This mix was incubated for 5 minutes at room temperature on rotation. Following this, the beads were pelleted on a magnetic rack, the supernatant carefully removed, and washed twice with 200µL 70% ethanol. To elute, the beads were resuspended in 21µL nuclease-free water and incubated on rotation for 10 minutes at room temperature. The beads were allowed to pellet on a magnetic rack and 21µL elute was transferred to a fresh 1.5mL LoBind tube. 1µL of strand-switched cDNA was quantified using the Qubit3 HS dsDNA kit (Invitrogen Q32851).

2.2.6.3 End-Preparation and Barcode Ligation

The ends of the resultant cDNA were repaired and dA-tailed using the NEBNext Ultra II End Repair/dA-Tailing Module (New England Biolabs E7546). 20µL of the strand-switched cDNA sample was diluted further by adding 30µL of nuclease-free water before a reaction mix was formed by adding 7µL Ultra II End Prep Reaction Buffer and 3µL Ultra II End Prep Enzyme Mix to the cDNA. The mix was homogenised by gentle flicking, spun down and then incubated at 20°C for 5 minutes followed by another 5 minutes at 65°C. Then, the sample was transferred to a clean 1.5mL LoBind tube and 60µL of AMPure XP beads were added to the reaction. The reaction was incubated on rotation for 5 minutes at room temperature. After this, beads were pelleted on a magnetic rack and washed twice with 200µL ethanol without disturbing the pellet. The tubes were removed from the rack and the beads resuspended in 23µL nuclease-free water and incubated at room temperature for 2 minutes. The beads were then pelleted on a magnet and 22.5µL of elute was transferred to a fresh 1.5mL LoBind tube for barcode ligation.

The native barcodes (ONT EXP-NBD196) were ligated onto the end-prepared cDNA using Blunt/TA Ligase Master Mix (New England Biolabs M0367). A different barcode was used for each sample and no barcode was used for both an embryo sample and a head kidney sample. In brief, 2.5µL of native barcode was added to the 22.5µL of end-prepped cDNA and 25µL of Blunt/TA Ligase Master Mix was added to the reaction, which was incubated for 10 minutes at room temperature. Next, 50µL of AMPure XP beads were added to the reaction and incubated at room temperature for 5 minutes on rotation. The tubes were placed on a magnetic rack and the beads allowed to pellet. The supernatant was removed and the beads were washed twice with 200µL of 70% ethanol without disturbing the beads. After the second wash, the tube was removed from the rack and the beads resuspended in 26µL of nuclease-free water and incubated for a further 2 minutes at room temperature. The beads were placed back on the magnetic rack and 26µL elute was transferred to a clean 1.5mL LoBind tube. 1µL of barcoded cDNA was taken for quantification with Qubit 3 using the HS dsDNA assay kit.

2.2.6.4 Library Pooling and Sequencing on PromethION

For each library, the samples were equimolar pooled to a total of 325ng cDNA (18ng/sample). Then, a 2.5x AMPure XP bead purification was carried out, washing twice with 70% ethanol. The final pooled library was eluted in 100µL nuclease-free water and quantified once again on Qubit3 with the HS dsDNA assay kit. Each pooled, barcoded libraries were sent to Edinburgh Genomics who carried out AMXII (ONT) sequencing adapter ligation. Each library was sequenced for 72 hours on the PromethION device with flowcell chemistry v9.4.1 (ONT).

2.2.7 Data Processing – Basecalling, Full-Length Filtering and Mapping

A full bioinformatic pipeline is visualised in Figure 2.1.

Basecalling of PromethION reads was performed with Guppy v5.0.11 (ONT) for the embryo dataset, and Guppy v6.3.7 for head kidney samples. The super-accuracy CRF basecalling model "`dna_r9.4.1_450bps_sup_prom.cfg`" was used with option "`--disable_qscore_filtering`" in both cases. Low quality reads (q-score <7) were filtered out using NanoFilt v2.7.1 (De Coster et al., 2018). Demultiplexing was carried out with the respective Guppy versions in default settings. Full-length reads were identified, oriented and trimmed of ONT barcodes and sequencing adapters with Pychopper v2.5.0 (ONT). CutAdapt v4.1 (Martin, 2011) was used to remove polyA tails with option "`-a "A{50}"`", then Minimap2 v2.22r1 (Li, 2021) was used to map full-length reads to the Ssal_v3.1 genome assembly with options "`-ax splice -uf -k 14`". It was found that a large number of secondary alignments were captured by Minimap2, which was investigated (see sections 2.2.8 and 2.3.2). Following this investigation, it was decided that only primary alignments would be taken forward into transcriptome construction. Non-primary alignments were filtered out using SAMtools v1.13 (Danecek et al., 2021) with "`samtools view -b -F 2308`". Sam files were sorted using "`samtools sort`" and bam file statistics collected with NanoStat v1.6.0 (De Coster et al., 2018) and the flagstat function in the SAMtools suite. Further

summary statistics were collected at each stage using pycoQC v2.5.2 (Leger et al., 2019), NanoStat v1.6.0 (De Coster et al., 2018) and samtools flagstat.

2.2.8 Resolving High Secondary Mapping Rates

A large proportion of ONT reads showed secondary mapping to alternative loci with high confidence after Minimap2 processing. It was suspected that these reads originated from regions of the genome containing high similarity ohnologue genes retained from the Ss4R WGD event (see section 1.4.2). Different regions of the salmon genome underwent the process of rediploidisation at distinct rates, meaning ohnologue pairs from different genomic locations can show very different levels of sequence divergence. (Robertson et al., 2017; Gundappa et al. 2022). In genomic regions where rediploidisation occurred quickly, in the ancestor of extant salmonids (AORe regions), the ohnologue pairs have had the maximal possible evolutionary time to diverge since Ss4R (~100 Myrs; Gundappa et al. 2022), and hence are less similar than ohnologue pairs from genomic regions where rediploidization was highly delayed (LORe regions). To compare the rates of primary and secondary mapping between AORe and LORe regions, the coordinates of duplicated Ss4R blocks belonging to each group was extracted and the number of primary and secondary alignments in each region counted. The counts were imported into RStudio (R4.3.3) and a boxplot was generated with ggplot2 v3.5.0 (Wickham, 2016). A Mann-Whitney U test was conducted in base R with the "`wilcox.test()`" function to compare secondary mapping rates between AORe and LORe regions.

To assess whether Minimap2 was able to properly differentiate between similar ohnologues and assign reads correctly, 5 ohnologue pairs with minimum 95% similarity in their CDS were selected from chromosomes 11 and 26 (i.e. a pair of LORe regions retained from Ss4R) for analysis;

ENSSSAT00000011981 - ENSSSAG00000107398 (chr 11) &
ENSSSAT00000221549 - ENSSSAG00000085143 (chr 26),

ENSSSAT00000012132 - ENSSSAG00000005641 (chr 11) &
ENSSSAT00000217538 - ENSSSAG00000088263 (chr 26),

ENSSSAT00000056862 - ENSSSAG00000040259 (chr 11) &
ENSSSAT00000216562 - ENSSSAG00000096373 (chr 26),

ENSSSAT00000056563 - ENSSSAG00000040107 (chr 11) &
ENSSSAT00000021980 - ENSSSAG00000010004 (chr 26),

ENSSSAT00000121309 - ENSSSAG00000069713 (chr 11) &
ENSSSAT00000082105 - ENSSSAG00000052238 (chr 26).

10 reads with primary alignments to each ohnologue pair were randomly selected, aligned to the CDS of the ohnologues (extracted from Ensembl with the online Biomart tool; Durinck et al., 2005) with the MAFFT version 7 (Katoh et al., 2019) with command "`mafft --thread 8 --threadtb 5 --threadit 0 --reorder --adjustdirection --auto input > output`" (total 22 sequences per alignment) and UTRs manually trimmed from each alignment using BioEdit (Hall, 1999) to retain only the CDS. Substitutions in the CDS that differentiated each ohnologue pair were identified from the alignment. Then, for each ohnologue, I counted the number of times the primary aligned read possessed the correct distinguishing base of the ohnologue it was aligned to, or the distinguishing base belonging to its ohnologue pair. The proportions of 'correct' bases in the 10 reads for each ohnologue pair was calculated and data imported into RStudio for plotting with ggplot2.

2.2.9 Collapsing Redundant Transcript Models and Transcriptome Assembly

To generate the long-read transcriptome, primary alignments for the embryo and head kidney datasets were collapsed separately into a set of consensus models based on genomic location using TAMA Collapse from the TAMA suite (Kuo et al., 2020) with options "`-x no_cap -a 25 -z 25 -sjt 10`". Collapsing transcripts with these parameters will collapse reads with exon start and end sites within 25bp of each other, a common error in long-read RNA-seq (Kuo et al., 2020). The capping mode of TAMA was set to "`no_cap`" to ensure that any remnants of 3' degradation were collapsed into longer transcripts. Read support, i.e. how many individual reads support

each collapsed transcript mode, was generated using TAMA GO. Then, the transcriptome for each dataset was filtered such that all transcript models possessing read support < 3 were removed, thus retaining a robust set of well-supported transcript models.

To compare the collapsed transcriptomes with the existing Ensembl reference annotation (Ssal_v3.1), the embryo and head kidney transcriptome files were converted from .bed format to .gtf with TAMA GO format converter. Then, SQANTI3 v5.1.1 (Tardaguila et al., 2018; Pardo-Palacios et al., 2024a) was used in default mode to characterise the collapsed, filtered transcriptomes and compare them with the existing Ssal_v3.1 Ensembl annotation. SQANTI3 was run on both the unfiltered and filtered assemblies. Finally, both datasets were merged using TAMA Merge with default settings to produce a final, combined, filtered transcriptome. Once again, SQANTI3 was used to compare the combined transcriptome to the reference assembly.

2.3 Results

Over 100 million raw reads constituting approximately 110Gb of sequence data were produced, with 37,623,452 and 39,708,613 reads passing qscore filtering for embryos and head kidney datasets, with a median read length of 977 and 965bp, respectively (Table 2.2 and Figure 2.2B & C). 27.6% of the embryo and 35.5% of the head kidney high-quality reads were classified as full-length by Pychopper. In the embryo dataset, 9,462,770 primary alignments with an N50 of 1,389bp were taken forward to transcriptome construction. A comparable number of primary alignments were retained in the head kidney dataset (10,250,416); however, these aligned reads were shorter on average, with an N50 of 963bp.

2.3.1 Long-Read RNA-Seq Successfully Distinguishes Between Ohnologue Pairs

Examining rates of primary and secondary alignments in AORe and LORe regions confirmed significantly higher rates of secondary mapping in regions of high similarity (Figure 2.3; $p<0.0001$).

My investigation into base pair substitutions that distinguish the CDS of 5 of highly similar ohnologue pairs revealed that Minimap2 was able to correctly assign reads to their respective ohnologues. Most reads possessed the same distinguishing bases as the ohnologues to which they mapped in primary alignments (Figure 2.4). The next most common instance was that no base was present in the reads, instead the alignment was gapped at that locus. In the case of ohnologue pair 2, missing bases were prevalent, being observed in over 80% of the distinguishing bases (Figure 2.4). It was rare that a primary alignment to one of the ohnologue genes had a distinguishing base that was found in its pair.

As a consequence of these results, it was determined that primary alignments were sufficient at assigning reads to the appropriate duplicated regions of the genome and non-primary alignments were removed before the transcript model collapsing step of the assembly pipeline (section 2.2.9).

2.3.2 Comparison of Full-Length Transcriptome with Ensembl Reference

After collapsing redundant reads into consensus transcript models, filtering models based on read support, and merging the two datasets, the final transcriptome consisted of 35,480 unique genes comprising 266,222 transcripts, with an average transcript-to-gene ratio of 7.50. This is markedly higher than the average transcript-to-gene ratio of 2.65 observed in the current reference Ensembl Ssal_v3.1 annotation, which contains 69,389 genes and 184,209 transcripts.

Of the 35,480 unique genes, 27,674 (78%) were deemed protein-coding, compared with 210,892 out of 266,222 (79%) for transcripts. Despite the Ensembl annotation containing a similar proportion of protein-coding transcripts (79%), there are significantly fewer protein-coding transcripts annotated in the Ensembl suggesting that the long-read RNA-seq has captured additional protein-coding diversity. 31,271 of the genes captured in the long-read transcriptome overlapped the existing Ensembl reference annotation, with 4,209 being deemed novel genes. 2,977 of these novel genes were associated with a single transcript and, in total, the novel genes

were associated with 10,752 transcripts, with 6,246 predicted to be non-coding.

34% of the genes in the long-read transcriptome showed a single transcript model (Figure 2.5) compared with 50% in the Ensembl annotation. At the other end of the scale, around a third of the long-read transcriptome genes showed ≥6 transcripts, compared to 12% for the Ensembl reference. The majority of dataset-specific genes, i.e. genes only supported by reads from a single dataset, have a single transcript model whilst in contrast, the vast majority of genes supported by both datasets have 6 or more different transcript models (Figure 2.5). Outside of Atlantic salmon, in the current Ensembl v113 annotation for zebrafish (GRCz11), 56% of genes possess only a single transcript model, whilst only 3% of genes have 6 or more transcript models associated with them. This highlights the complexity of the Atlantic salmon genome when compared with other well-studied teleosts.

2.3.3 Classification of Transcript Diversity

Figure 2.6 depicts the broad structural categories SQANTI3 assigned to each transcript model in my long-read transcriptome. Whilst 79% of long-read gene models overlapped with reference genes, only 41,371 of the transcripts overlapped Ensembl transcripts with identical splice junctions (16% of total), with around 15,000 classified as full-splice-matches (FSM) (Table 2.3). The rest were classified as incomplete-splice-matches (ISM), which indicates that the splice junctions of the long-read transcript corresponded with those of a reference Ensembl transcript, but with 5’ or 3’ exons missing (e.g. shown in Figure 2.7).

The most prevalent structural category of the novel transcripts ‘novel not-in-catalogue’ (NNC) accounted for 60% of the long-read transcript models, indicating that many transcripts contained splice sites not predicted in the Ensembl annotation. 12% of transcripts were deemed ‘novel-in-catalogue’ (NIC), possessing a new combination of known exons, constituting the second most common structural category behind NNC (e.g. shown in Figure 2.8).

Finally, the total number of ‘Genic Genomic’, ‘Antisense’, ‘Fusion’ and ‘Intergenic’ transcript models combined was 31,854, constituting 14% of all novel transcripts. Over half of these transcripts were predicted to be non-coding (55%). The 8,195 antisense transcripts make up 3% of total transcripts and derive from 3,382 unique gene models, 69% coding for a single transcript (2,330/3,382). Of the 8,195 antisense transcripts, 6,259 are multi-exonic. Figure 2.9 depicts a subset of 112 transcripts that overlapped with reference transcripts of *rpl13a,* as well as a novel gene that produces 8 transcripts antisense to *rpl13a.*

2.3.4 Mono-Exonic Transcripts Captured by Long-Read Transcriptome

While most transcript models captured in the long-read transcriptome were multi-exonic, a significant proportion (29.8%) deriving from novel genes were mono-exonic (e.g. Figure 2.10). In contrast, long-read transcripts deriving from genes associated with an annotated Ensembl gene were mainly multi-exonic (3.8%; Figure 2.10). As a proportion of total transcripts, mono-exonic models constituted only 5.7% - comparable to the Ensembl annotation. The prevalence of mono-exonic transcripts deriving from novel genes captured by long-read RNA-seq is further explored in detail in Chapter 5.

2.3.5 Transcript and Gene Model Support Originates from Both Datasets

51.7% of genes in the long-read transcriptome had support from reads originating from both the embryo and head kidney datasets, revealing that the majority of expressed genes are common to both embryogenesis and immune response (Figure. 2.11). However, only 16.7% of the transcripts were common to both datasets, with most transcripts therefore unique to either embryo or head kidney samples. Furthermore, the embryo dataset contained 39% more unique transcripts than the head kidney dataset (129,053 vs 92,799).

2.4 Discussion

In this chapter, I constructed the first full-length transcriptome assembled with ONT RNA-seq technology for Atlantic salmon. This long-read transcriptome is the first of its kind in any farmed finfish species. The

transcriptome contains 266,222 putative full-length transcripts produced from 35,480 unique genes, greatly expanding the number and functional diversity of Atlantic salmon genes and transcripts available for future studies.

2.4.1 Long-Read Nanopore Sequencing Yields High-Quality Reads

Long-read transcriptome construction has been shown to be strongly impacted by RNA degradation (Prawer et al., 2023) and chimeric reads (White et al., 2017b). To offset this possibility, only RNA of high integrity was used in this study, while robust bioinformatic filtering was employed to remove potential artifacts. Despite this, only a quarter of sequenced reads were deemed full-length, with the majority filtered out as either RNA degradation products, internal polyA priming artifacts, or potential chimeric reads resulting from erroneous ligation events during library preparation. Halstead et al. (2021) obtained a similar quantity of raw data using the PCR cDNA Nanopore sequencing method in *Bos taurus*, yet retained 70% of reads post-filtering. This may suggest that the native cDNA method introduces bias towards truncated or reads that are artificially fused together during sequencing or basecalling (Sessegolo et al., 2019). Nonetheless, the 24 million reads that passed filtering still represented a robust set of high-quality long-read transcripts, indicating that the filtering strategies employed in this chapter were successful at removing confounding sequencing artifacts.

2.4.2 Novel Transcript Diversity Revealed by Long-Read RNA-Seq

The majority of publicly available transcriptome resources have been constructed using evidence from short-read RNA-seq data (Stark et al., 2019). This may lead to poor or partial genome annotations due to challenges such as resolving complex genomic regions and consolidating intron-exon boundaries (Byrne et al., 2019). Long-read RNA-seq has been widely adopted to characterise novel transcripts (Kuo et al., 2020; Halstead et al., 2021; Glinos et al., 2022; Reese et al., 2023). The transcriptome assembled in this chapter revealed many new transcripts compared to the Ensembl Atlantic salmon annotation, most containing novel splice junctions. This highlights the value of long-read RNA-seq for the identification of novel

transcript diversity. The current Ssal_v3.1 Ensembl Atlantic salmon annotation has not been updated since 2021 and has yet to incorporate new long-read data. As such, the data in this thesis will be made publicly available and can be used by Ensembl for future gene-builds.

A large proportion of novel genes in the long-read transcriptome were deemed non-coding, consistent with findings in humans (Kuo et al., 2020), chicken (Kuo et al., 2017) and pig (Beiki et al., 2019). Interest in non-coding RNAs (ncRNA) has increased in recent years (Micheel et al., 2021) due to the myriad roles of ncRNAs in transcriptional and post-transcriptional regulation (Statello et al., 2021; Sartorelli & Lauberth, 2020). In salmonids, significant regulation of ncRNAs has been observed in response to viral infection in Atlantic salmon (Boltaña et al., 2016; Xia et al., 2022) whilst tissue-specific expression patterns of miRNA and lncRNA were revealed in the immune organs of healthy coho salmon (*Oncorhynchus kisutch*; Leiva et al., 2020). These findings demonstrate that ncRNA is an important mechanism for the genomic and transcriptomic regulation in salmonids. High-quality transcriptome assemblies, such as those constructed here with long-read RNA-seq, will be an important resource for future ncRNA research (Uszczynska-Ratajczak et al., 2018).

2.4.3 Long-Read RNA-Seq Resolves Complex Transcriptomic Regions

Salmonid genomes comprise large duplicated regions retained from the Ss4R WGD, with 50-60% of genes existing in duplicated ohnologue pairs (Berthelot et al., 2014; Lien et al., 2016). This poses a challenge to distinguishing duplicated genomic exons with short-read RNA-seq (Houston & Macqueen, 2019; Deschamps-Francoeuret al., 2020). My long-read RNA-seq approach was successful in distinguishing between highly similar ohnologue pairs and has the potential to aid in research aiming to unpick the function of gene duplicates and resolve their expression patterns (see General Discussion).

2.4.4 Mono-Exonic Transcripts Underrepresented in Reference Annotation

Mono-exonic transcripts are an underexplored class of transcript diversity that have historically been discarded as transcriptional noise (Su et al., 2024)

but are readily captured by long-read RNA-seq (Kuo et al., 2020; Zhang et al., 2022). As with previous long-read studies, a significant proportion of the transcript diversity originating from novel Atlantic salmon genes was attributed to mono-exonic transcripts. As mono-exonic transcripts are yet to be explored in salmonids, Chapter 5 addresses this knowledge gap.

2.4.5 Potential Improvements to Transcriptome Assembly Pipeline

There are some areas where the transcriptome assembly pipeline detailed in this chapter may benefit from additional analysis and polishing. For instance, long-read Nanopore RNA-seq has been shown to struggle when sequencing stretches of homopolymers (Wang et al., 2021). Consequently, polishing strategies have been developed to correct these types of systematic errors. Polishing tools such as Pilon (Walker et al., 2014) and NextPolish (Hu et al., 2020) leverage the high accuracy of NGS short-read RNA-seq methods to correct errors in long-read sequences, thereby reducing raw error rates ten-fold (Dohm et al., 2020), whilst other software such as Homopolish (Huang et al., 2021) can correct errors *in-situ*. Whilst mainly used in long-read genomic DNA sequencing, recent advances in error correction for long-read RNA-seq have been shown to be effective for transcriptome profiling with tools like IsONcorrect (Sahlin & Medvedev, 2021) reducing error rates to 1%. Due to the relation of this work with the AQUA-FAANG consortium, matched short-read RNA-seq data exists for each sample used in the long-read transcriptome assembly pipline. Thus, both short-read error correction or self-error correction could be employed in this instance to potentially improve the Nanopore transcriptome characterisation.

Long-read RNA-seq can also suffer from other error types including RNA degradation which can produce reads with 5’ truncation (Prawer et al., 2023), or priming of poly(dT) to an internal stretch of adenine bases during cDNA synthesis, which can cause fragmentation of full mRNA sequences (Sessegolo et al., 2019). The use of TAMA in the transcriptome assembly pipeline reduces the impact of RNA degradation through its “no-cap” collapsing mode which collapses reads into a longer model where reads possess the same splice sites (Kuo et al., 2020). However, employing an

additional clustering strategy such as CD-HIT (Fu et al., 2012) may aid in producing a set of non-redundant reads for input into the collapsing stage of transcriptome construction (Ramberg et al., 2021) thus reducing the impact of RNA degradation and internal poly(dT) priming.

Finally, SQANTI3 was used to determine if a transcript model had coding potential via its use of GeneMarkS-T which employs self-trained Markov models to predict coding potential (Tang et al., 2015). Using an additional ORF predictor such as Transdecoder (Haas, 2023) would allow verification of coding potential for each transcript and thus increase confidence in coding prediction.

2.4.6 Concluding Words

Overall, the work presented in this chapter demonstrates the utility of long-read RNA-seq, describing a wealth of transcript diversity that isn't currently annotated by Ensembl in the Atlantic salmon reference genome. I have used the long-read transcriptome constructed in this chapter for novel investigations of transcript expression dynamics employed during embryonic development (Chapter 3) and immune responses (Chapter 4). Finally, the long-read data produced in this project will be submitted to Ensembl and NCBI to contribute to improved future genome annotations.

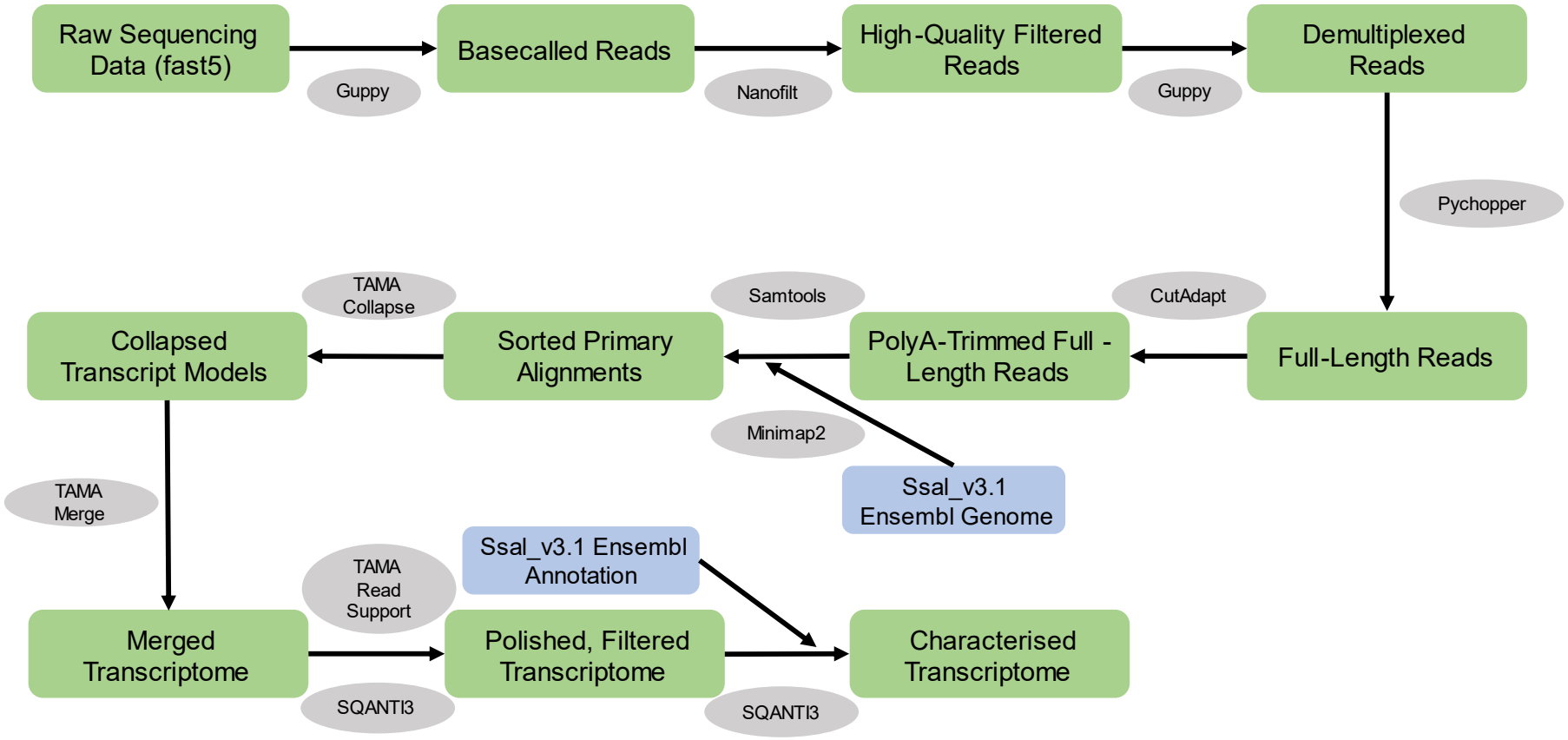


Figure 2.1: Schematic of the bioinformatic pipeline used to process raw data, construct and polish the long-read transcriptome and compare it with the Ensembl reference annotation.

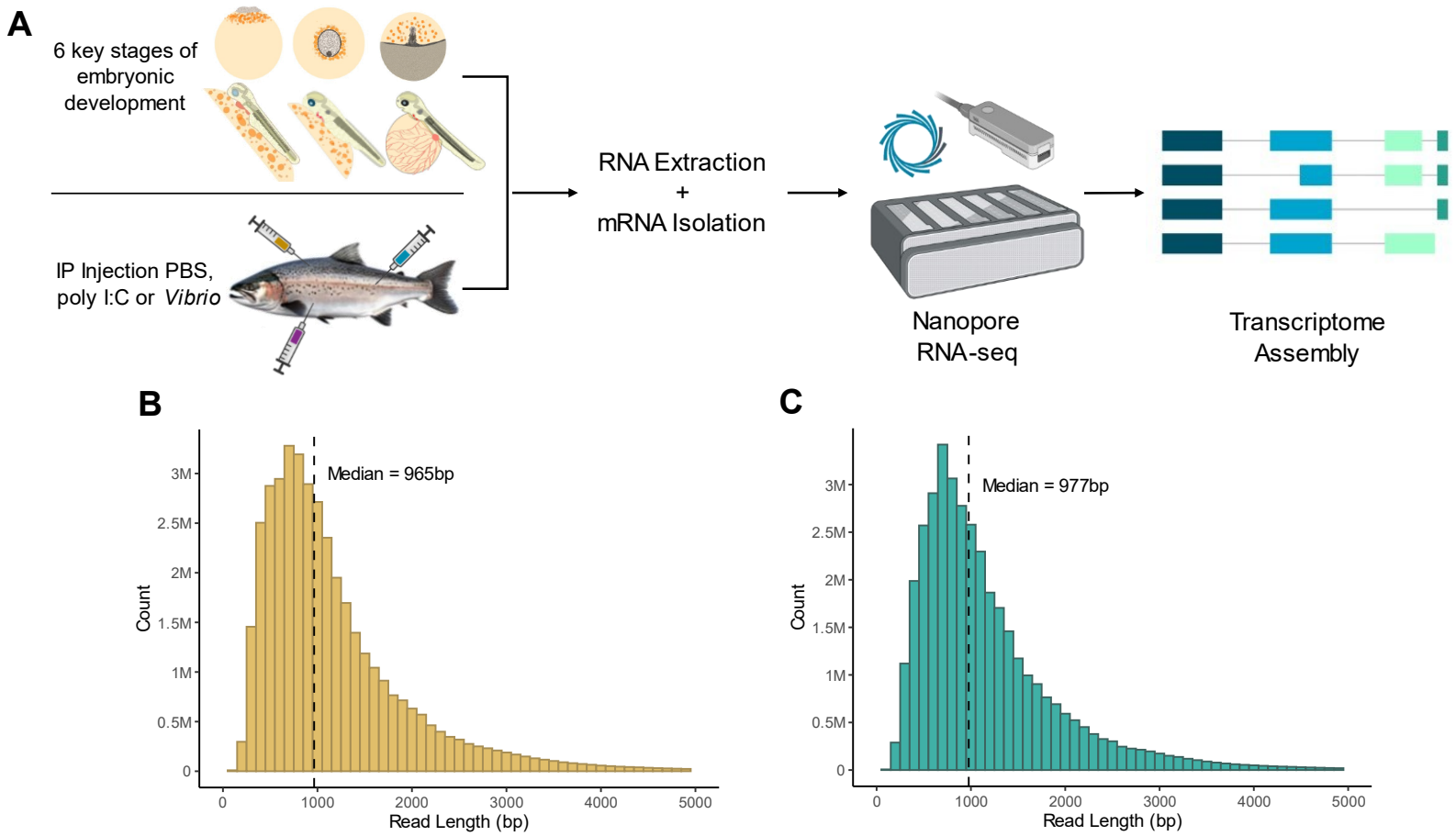


Figure 2.2: (A) Schematic of experimental design and read-length distribution of reads passing q-score filtering (q>7) for (B) head kidney and (C) embryo datasets. Embryo diagrams provided by Perojil-Morata (2024); salmon image from Johnston et al. (2024); nanopore and transcript diagram created with BioRender.com.

Table 2.1: Total RNA extraction quality control for all embryo and head kidney samples.

	Sample	Development Stage/Treatment Group	Nanodrop Concentration (ng/µL)	A260/280	A260/230	Concentration (ng/µL)	RIN Score	Final Volume (µL)	Total RNA (ng)
Embryos	165R1	Late Blastulation	70.5	2.03	1.80	55.2	8.8	60	3312
	165R2		63.2	2.00	1.48	49.5	9.0	60	2970
	165R3		49.1	2.10	1.50	40.0	8.8	45	1800
	214R1	Mid Gastrulation	59.7	2.17	1.10	52.4	9.3	50	2620
	214R2		43.7	2.12	0.52	40.8	9.1	50	2040
	214R3		59.2	1.94	1.89	49.4	8.9	50	2470
	309R1	Early Somitogenesis	107.6	1.96	2.23	119	8.1	50	5950
	309R2		84.2	1.96	2.38	96.2	9.8	50	4810
	309R3		107.1	1.92	2.32	115	9.6	50	5750
	408R1	Mid Somitogenesis	90.8	2.07	1.91	111	9.9	50	5550
	408R2		152.9	2.12	1.83	125	10.0	50	6250
	408R3		198.8	2.14	1.82	204	9.4	50	10200
	552R1	Late Somitogenesis	126.4	1.97	2.32	156	10.0	100	15600
	552R2		238.9	2.01	2.41	278	10.0	100	27800
	552R3		234.3	2.03	2.15	298	10.0	100	29800
	792R1	Late Pharyngula	488.1	2.13	2.24	452	9.5	100	45200
	792R2		415.1	1.98	2.44	394	9.8	100	39400
	792R3		501.7	2.03	2.44	468	10.0	100	46800

	Sample	Development Stage/Treatment Group	Nanodrop Concentration (ng/μL)	A260/280	A260/230	Concentration (ng/μL)	RIN Score	Final Volume (μL)	Total RNA (ng)
Head Kidney	PBS1	PBS	365.3	1.96	1.87	372	9.5	40	14880
	PBS2		181.2	1.88	1.68	169	9.7	40	6760
	PBS3		133.0	1.83	1.84	118	9.8	40	4720
	PBS4		138.4	1.92	1.10	124	9.9	40	4960
	PBS5		182.1	1.90	1.54	174	10.0	40	6960
	PBS6		337.9	1.94	1.86	356	9.9	40	14240
	PolyIC1	Poly I:C	460.8	1.89	2.03	508	9.6	40	20320
	PolyIC2		310.7	1.91	1.65	326	9.5	40	13040
	PolyIC3		616.9	1.93	2.21	662	8.5	40	26480
	PolyIC4		360.1	1.83	2.30	390	9.5	40	15600
	PolyIC5		604.1	1.94	2.42	660	9.3	40	26400
	PolyIC6		589.0	1.97	2.27	626	9.6	40	25040
	Vib1	Inactivated *Vibrio anguillarum*	298.8	1.98	1.44	310	9.7	40	12400
	Vib2		259.3	1.92	1.38	258	9.4	40	10320
	Vib3		475.0	1.97	1.92	512	9.2	40	20480
	Vib4		352.4	1.91	1.47	368	9.7	40	14720
	Vib5		326.4	1.90	1.94	324	9.5	40	12960
	Vib6		267.3	1.87	1.50	274	9.4	40	10960

Table 2.2: Sequencing metrics for embryo and head kidney datasets

		Number of Reads	Mean Length (bp)	Median Length (bp)	N50	Mean Read Quality (q-score)
Embryos	**Raw Reads**	50,653,062 (55.6Gb)	1098	877	1371	9.4
	Filtered Reads (q>7)	37,623,452	1203	977	1422	11.1
	Full-Length Reads	10,370,701	947	684	1366	13.3
	Mapped Reads	9,462,770	999	717	1389	13.4
Head Kidney	**Raw Reads**	53,258,607 (57.2Gb)	1075	850	1374	9.5
	Filtered Reads (q>7)	39,708,613	1204	965	1446	11.3
	Full-Length Reads	14,077,488	538	334	900	12.8
	Mapped Reads	10,250,416	671	471	963	13.3

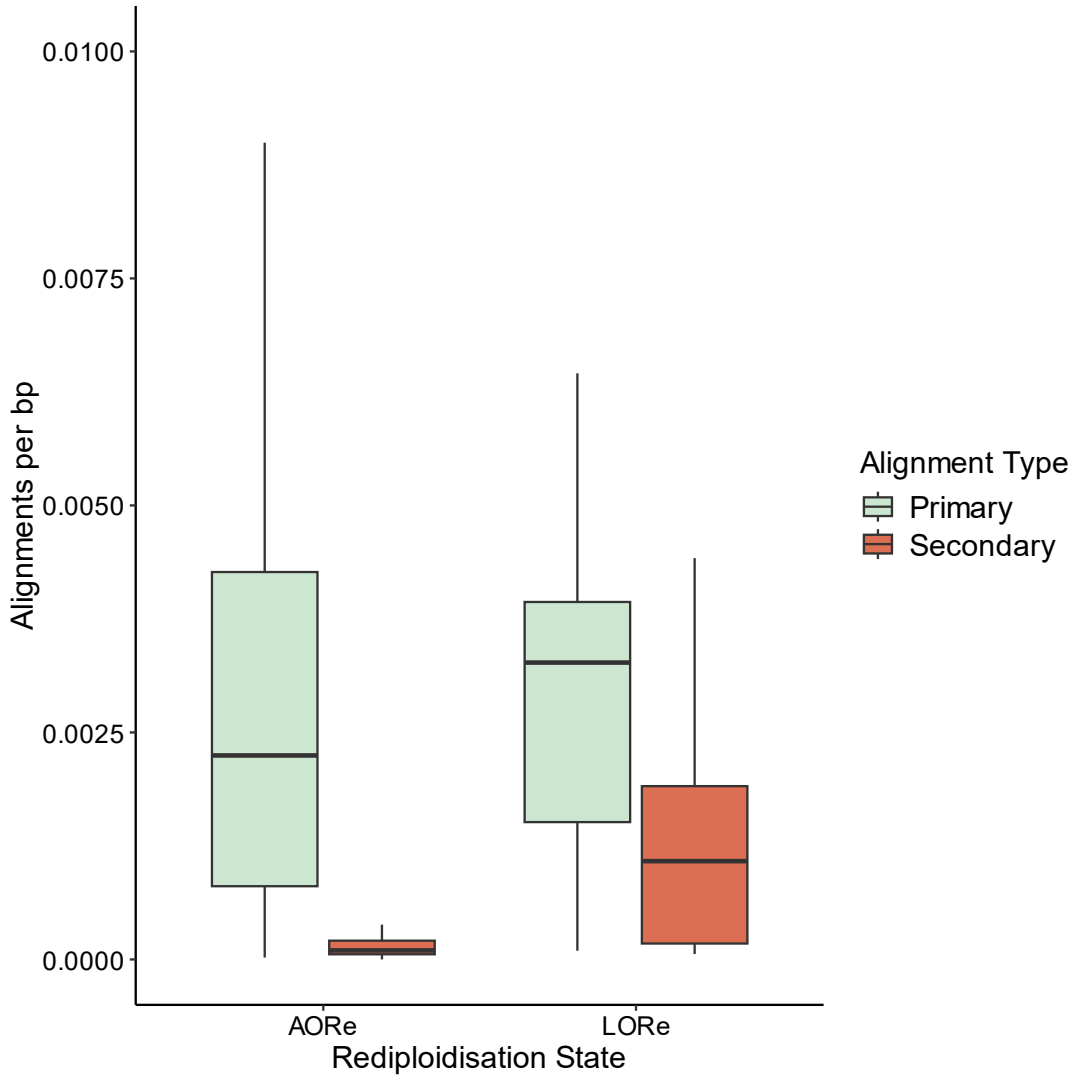


Figure 2.3: Boxplot showing rates of primary and secondary mapping in distinct duplicated genomic regions of the Atlantic salmon genome, where ohnologues share relatively lower (AORe) or higher (LORe) sequence similarity, owing to rediploidisation history following the Ss4R WGD (after Gundappa et al. 2022)

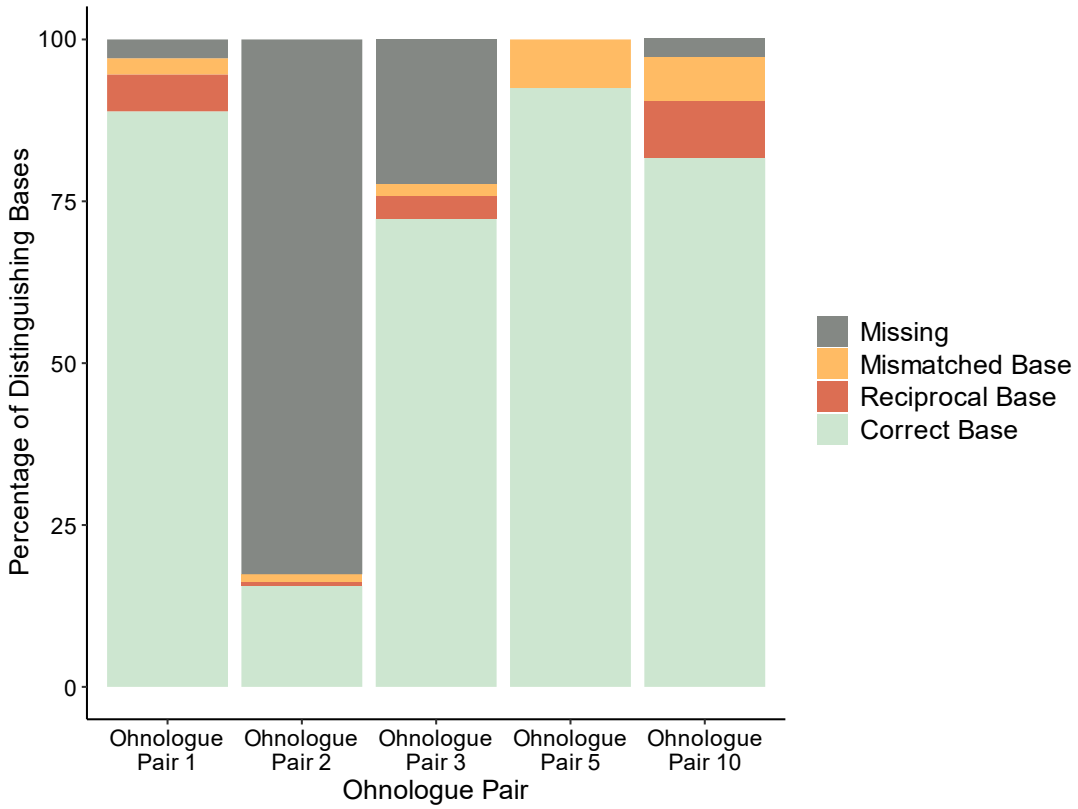


Figure 2.4: Ability of Minimap2 primary alignments to distinguish high-similarity ohnologue pairs sampled from duplicated chromosome pairs 11 and 26 in the current Ensembl Atlantic salmon reference genome (Ensembl v112; Ssal_v3.1). These chromosomes are LORe regions that have diverged minimally following Ss4R. Green shading indicates the proportion of reads that match the ohnologue they mapped to at the position of bases distinguishing the two ohnologues in each pair. Orange shading indicates the proportion of reads matching a distinguishing base found in the other ohnologue (note: this could represent allelic variation compared to the reference genome). Yellow shading shows the proportion of bases different from both ohnologues in the reference genome, which is likely to represent sequencing error. Grey shading represents reads that had no bases overlapping distinguishing bases in the reference ohnologue pairs, which could result from distinct gene models between the long-read and Ensembl annotations.

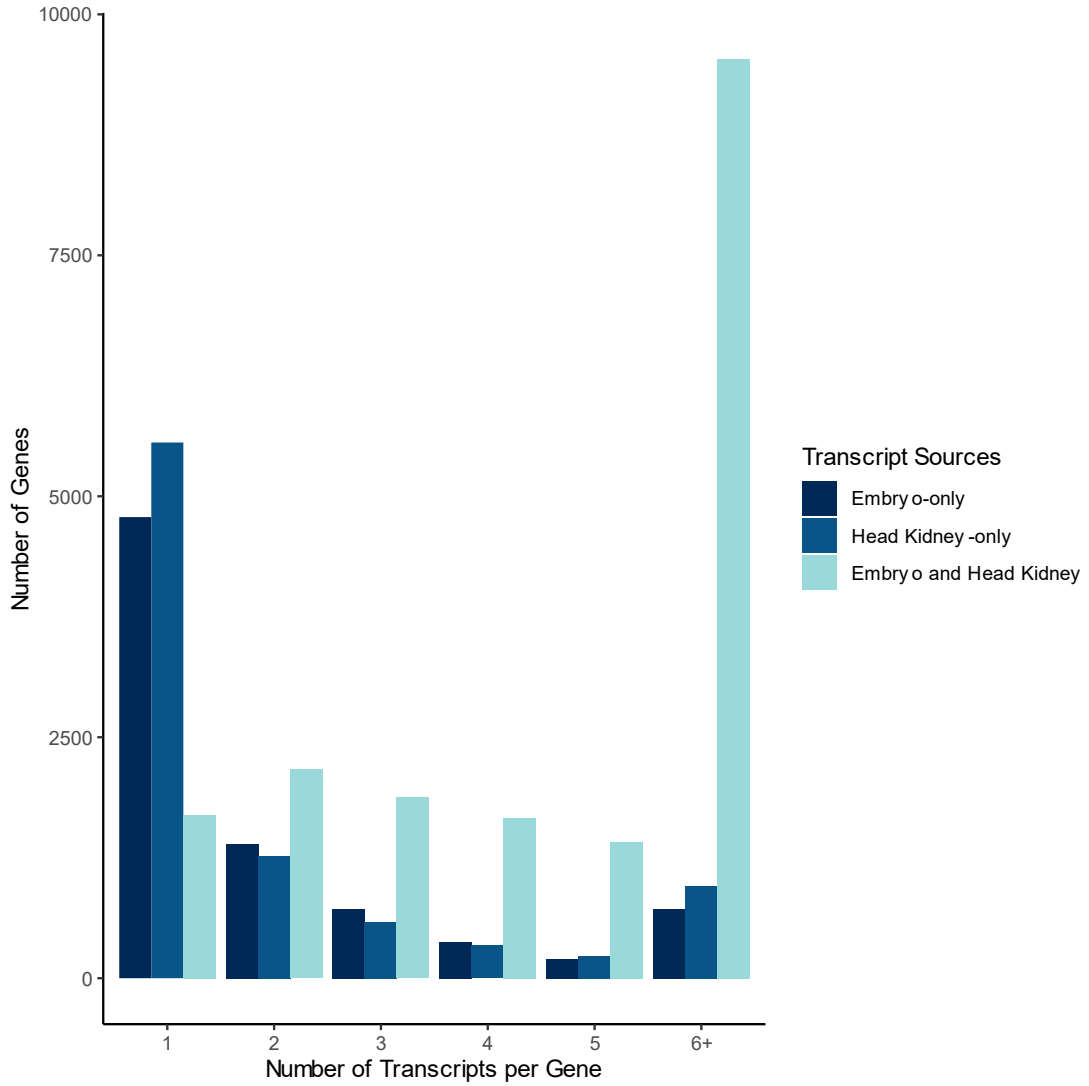


Figure 2.5: Number of transcripts per gene for the Atlantic salmon long-read transcriptome. Bars are coloured according to which dataset supports that gene. For example, if a gene is only supported by reads from the embryo dataset, it falls in the "Embryo-only" category.

Table 2.3: SQANTI3 transcript model structural categories for the Atlantic salmon long-read transcriptome

Structural Category	Count	%
FSM	15,563	5.85
ISM	25,808	9.69
NIC	32,217	12.10
NNC	160,780	60.39
Genic Genomic	4,569	1.72
Antisense	8,195	3.08
Fusion	7,077	2.66
Intergenic	12,013	4.51

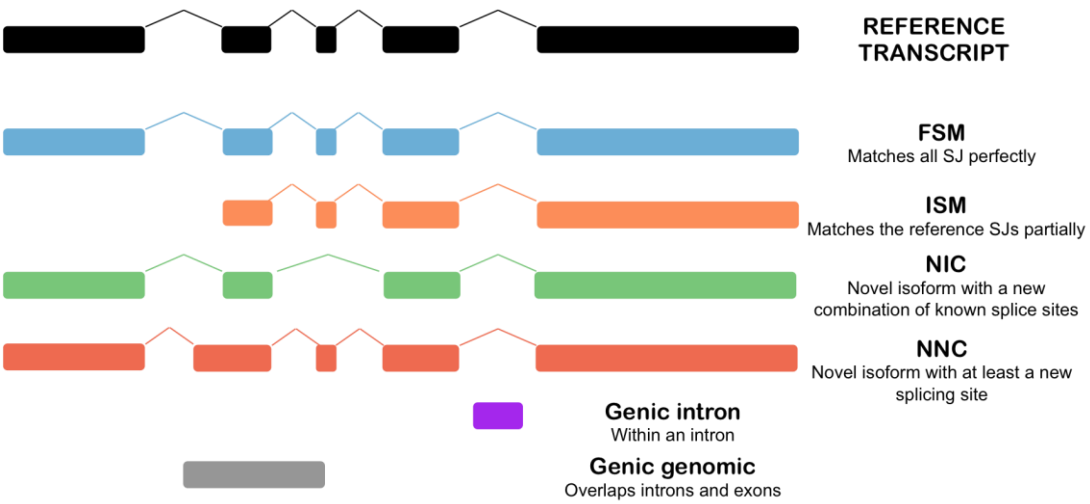


Figure 2.6: SQANTI3 structural categories (taken from Pardo-Palacios et al. 2024). FSM = full splice match; transcript matches exon structure of annotated transcript, ISM = incomplete splice match; transcript has fewer exons at 5' or 3' end than reference model, NIC = novel in catalogue; new combination of exons using known splice junctions, NNC = novel not in catalogue; new combination of exons with at least one novel splice site, Genic Genomic = overlaps existing introns and exons, Antisense = transcript is antisense to an annotated gene, Fusion = fusion between two annotated loci, Intergenic = transcript maps to locus currently undescribed in reference annotation.

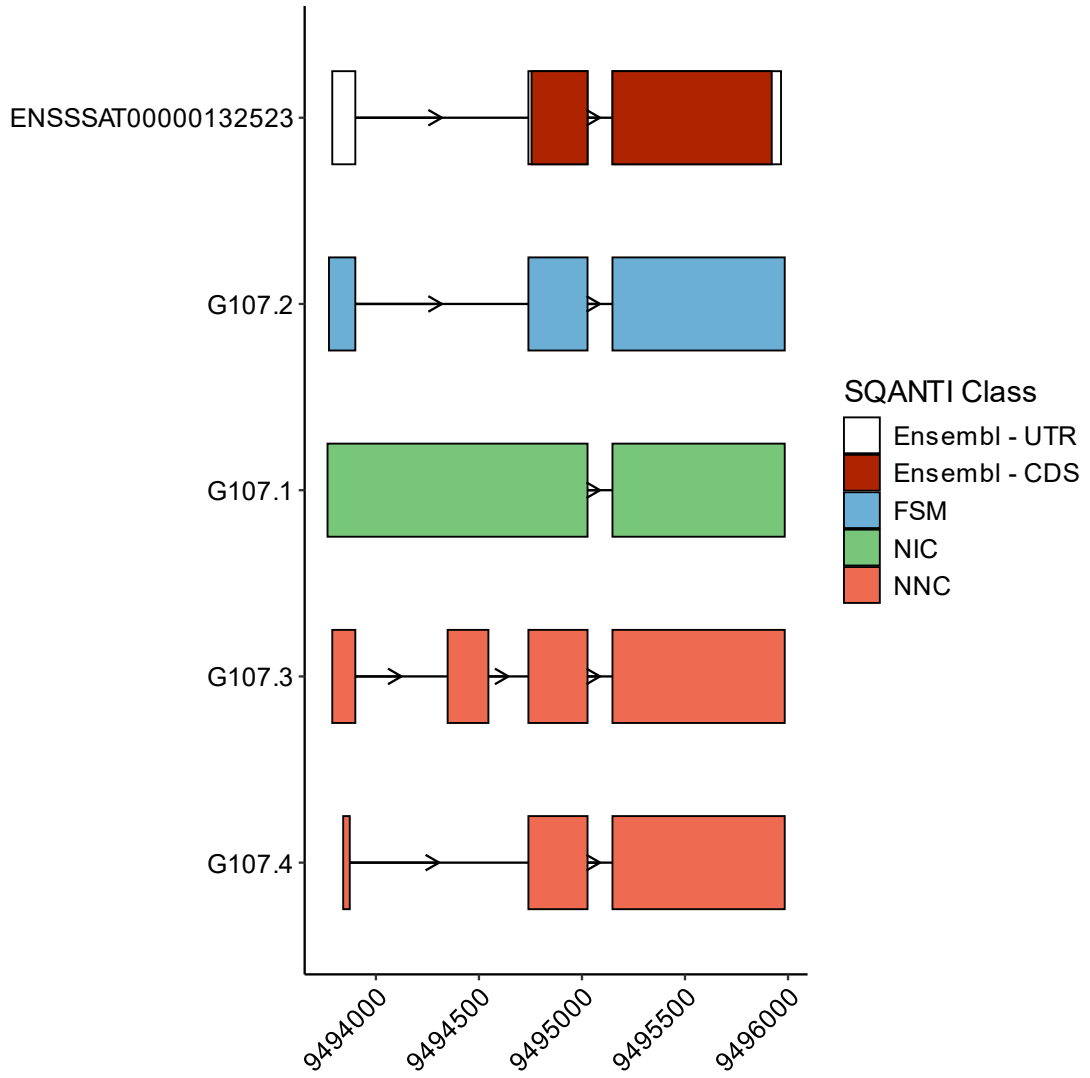


Figure 2.7: Structure of transcripts derived from long-read gene G107. The associated Ensembl gene is protein O-mannose kinase (pomk: ENSSSAG00000073686). Legend shows the SQANTI3 structural category for each long-read transcript: FSM = full-splice-match, NIC = novel-in-catalogue, NNC = novel-not-in-catalogue. Transcript G107.2 is classified as FSM sharing all splice sites identical with the reference transcript. G107.1 is considered an NIC transcript, as the splice sites are shared with the reference, but the exon combination differs as a result of intron retention. G107.3 and G107.4 are NNC. In the case of G107.3 this is due to a novel exon between reference exons 1 and 2 and a novel splice site at the 5' end, whilst for G107.4, a novel splice site is present at the 5' end.

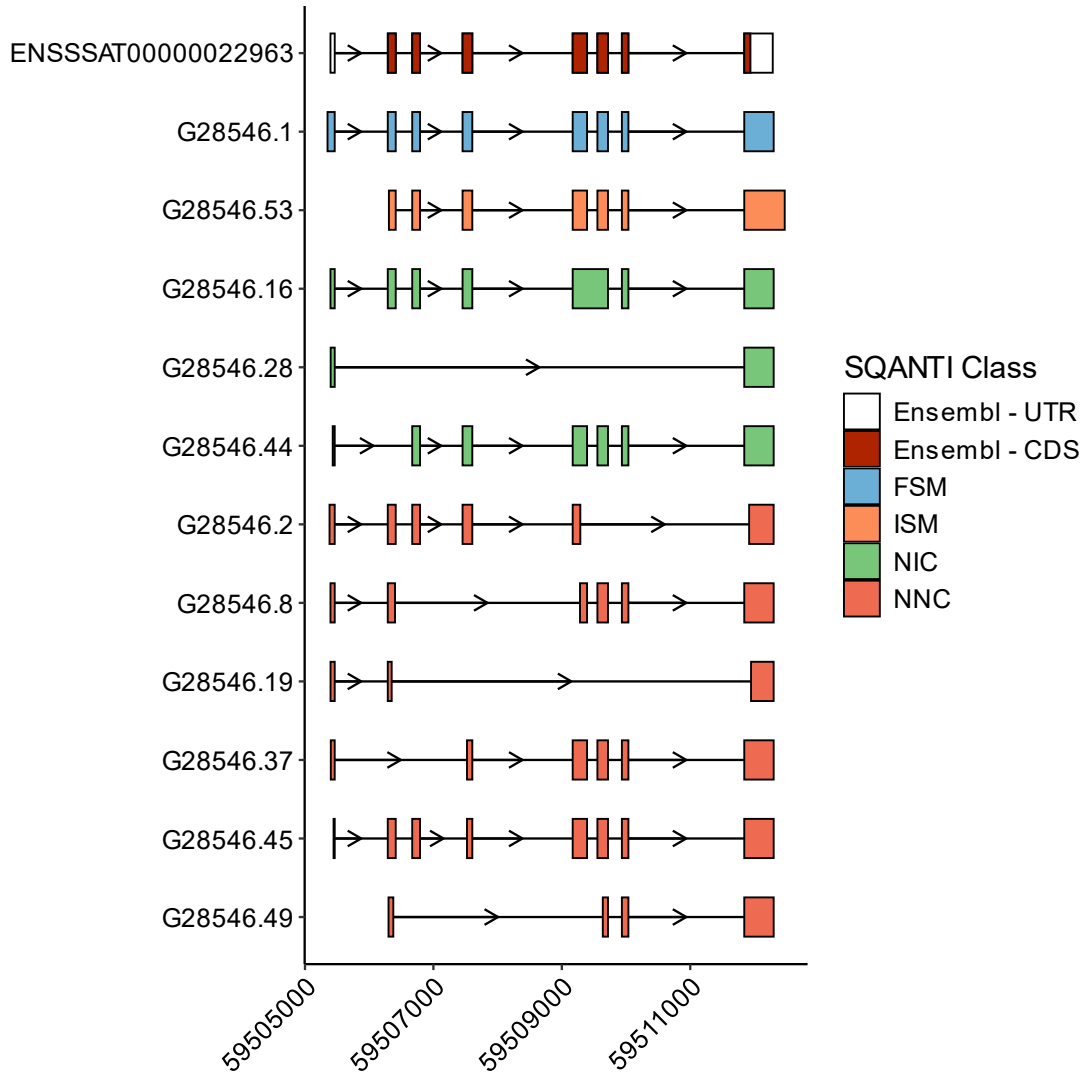


Figure 2.8: Structure of a subset of 11 transcripts derived from long-read gene G28546 – total number of long-read transcripts for this gene is 59. The associated Ensembl gene is cathepsin S (cats: ENSSSAG00000010327). Legend shows the SQANTI3 structural category for each long-read transcript: FSM = full-splice-match, ISM = incomplete-splice-match, NIC = novel-in-catalogue, NNC = novel-not-in-catalogue. Different structural categories are represented by transcripts of G28546 including G28546 which is a perfect splice match to the single reference transcript. G28546.53 is classified as an ISM because it shares the same splice junctions as the reference but is missing an exon at the 5' end. Transcripts .16, .28 and .44 are NIC displaying different combinations of known splice sites, whilst the rest of the transcripts were classified as NNC containing novel splice sites.

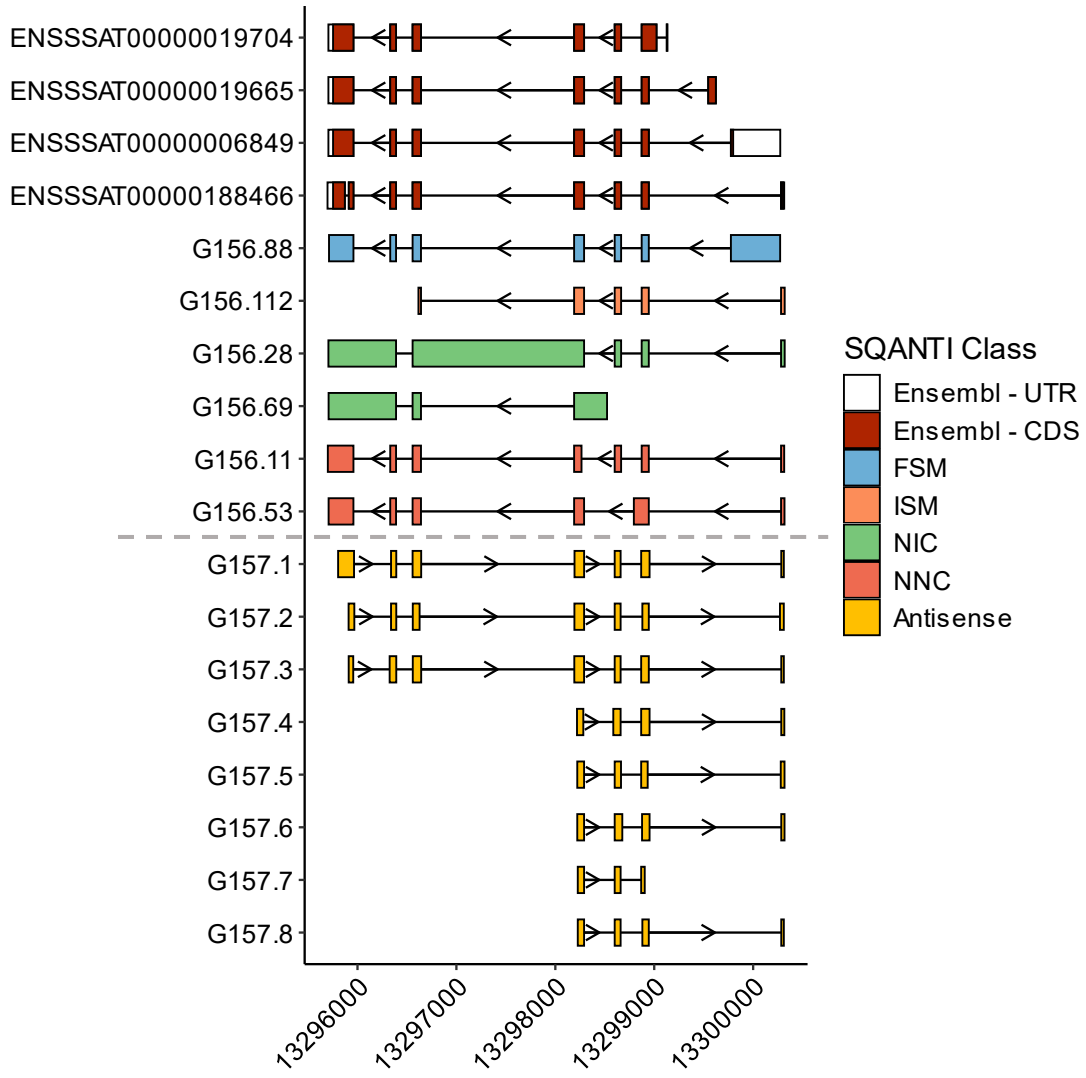


Figure 2.9: Structure of two genes in the long-read transcriptome; G156 & G157. G156 has 112 transcripts, 6 of which have been displayed in the plot. G156 is associated with rpl13a (ENSSSAG00000008822). G157 codes for 8 transcripts which are antisense to rpl13a.

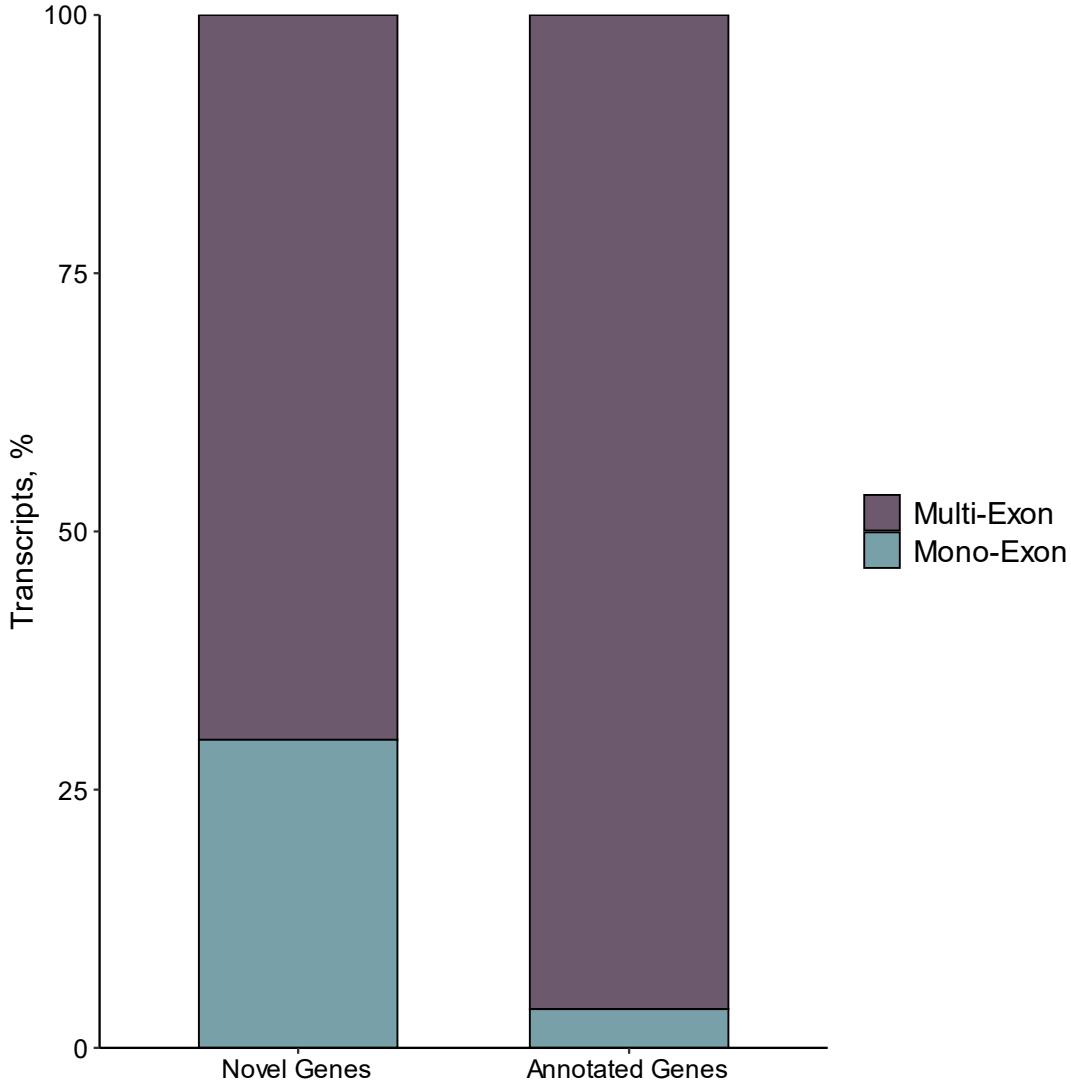


Figure 2.10: Distribution of mono- vs multi-exon transcript models in gene models considered novel or those matching a pre-existing gene annotation in the Ensembl reference annotation.

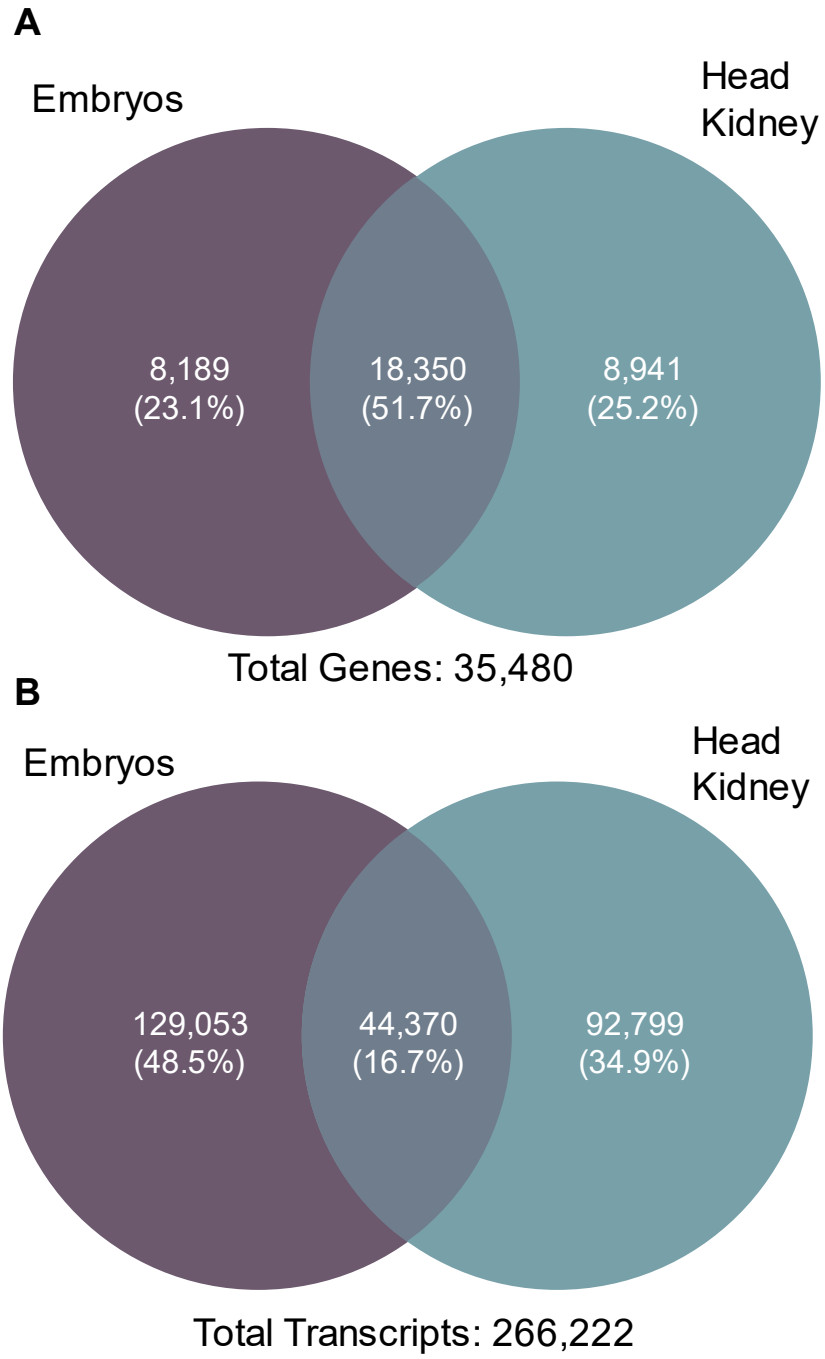


Figure 2.11: Venn diagrams showing which dataset supports each gene model (A) and transcript model (B) in the long-read transcriptome.

Chapter 3: Transcript Resolved Expression in Atlantic Salmon Head Kidney Following Viral and Bacterial Challenge

Summary

In this chapter, I examined transcript expression and regulation during separate immune responses mounted against both viral and bacterial challenges. Nanopore reads obtained from Atlantic salmon head kidney samples were mapped to the transcriptome assembly generated in Chapter 2, then quantified and analysed using a custom pipeline. Differential transcript expression analysis was carried out and within-gene alternative transcript usage explored. 1,096 and 627 transcripts were differentially expressed during viral and bacterial responses, respectively, with many novel transcripts and some novel gene models described for the first time.

3.1 Introduction

RNA transcript diversity plays a key documented role in the regulation of the immune system through alternative splicing (Schaub & Glasmacher, 2017; Su & Huang, 2021; Tao et al., 2024; Inamo et al., 2024), intron retention (Ni et al., 2016) and alternative transcription start sites (Carpenter et al., 2014; Mola et al., 2023). For example, an exon-retention transcript variant of the well-characterised immune system modulator *cd47* was upregulated in human acute myeloid leukaemia (van Der Werf et al., 2023) whilst alternative transcripts of several genes impacted the activation of the interferon pathway, an important antiviral response mechanism (Liao & Garcia-Blanco, 2021). While transcript diversity is recognised as a driver of immune system regulation, much of our current knowledge comes from studies conducted in a small number of mammal species and the role of transcriptional variation in immune regulation remains underexplored in most non-model species.

Aquatic diseases represent a significant challenge to the aquaculture sector, causing stock mortality and raising welfare concerns, with important economic and social ramifications (Stentiford et al., 2015). As such, there is a drive to improve our understanding of the immune system in aquaculture species to reveal mechanisms that influence the outcomes of vaccination

and underpin disease resistance traits. Consequently, one of the main aims of the European consortium AQUA-FAANG (Chapter 1, section 1.4.4) was to annotate genome functional elements associated with early immune responses for six commercial fish species including Atlantic salmon.

Short-read RNA-seq is a gold-standard method for studying gene expression, widely used to understand salmonid immune responses to bacterial and viral stimulation. For example, Clark et al. (2023) described a divergent arm of the type-I IFN response in Atlantic salmon and rainbow trout which employs both a core set of phylogenetically conserved ISGs and a second set of salmonid-specific ISGs, contributing to our knowledge of antiviral immune repertoire in salmonids. Short-read RNA-seq has been used to understand salmonid immune system responses to bacterial and viral pathogens (Gervais et al., 2021; Krasnov et al., 2021; Taylor et al., 2022), parasitic infection such as sea lice (Robledo et al., 2018; Casuso et al., 2022), heat stress (Shi et al., 2019), dietary alterations like functional feeds (Tawfik et al., 2024), and vaccination (Fu et al., 2022). However, as noted (Chapter 1, section 1.1.6), using short reads makes it challenging to distinguish alternative transcripts (Steijger et al., 2013; Kuo et al., 2017).

Long-read RNA-seq is a valuable approach for capturing alternative transcript diversity (Seki et al., 2019), but is yet to be widely applied to study functional immune responses in non-model fish species. Much of the use of long-read RNA-seq in fish species is limited to transcriptome construction with limited examples of quantification analysis. For example, PacBio RNA-seq identified a host of novel genes and transcripts from immune tissues in black rockfish *Sebastes schlegelii* (Cao et al., 2020) whilst over 60,000 immune-related transcripts were identified with PacBio technology in turbot *Scophthalmus maximus* following bacterial infection (Huang et al., 2022). In both studies, the focus was on characterising transcriptome diversity and no quantification was conducted. Similarly, Yu and colleagues (2024) used PacBio RNA-seq to construct a long-read transcriptome for the Xinjiang Yarrowfish *Leuciscus merzbacheri*, but applied short-read RNA-seq to quantify gene expression against the long-read transcriptome reference. As

such, the use of long-read RNA-seq for understanding fish immune responses remains in its infancy.

This chapter aims to use a long-read Nanopore RNA-seq approach to both identify novel transcript variants and quantify transcript expression in response to viral and bacterial immunostimulation. Two immune challenges were conducted, one with the PAMP poly I:C to mimic viral infection, and the other with an inactivated strain of *Vibrio,* a common bacterial pathogen affecting salmonid aquaculture. I describe the development of a custom bioinformatic pipeline for differential transcript expression analysis and identify a wealth of novel transcript diversity regulated by both immune stimuli uncaptured by short-read RNA-seq in the same samples.

3.2 Materials and Methods

3.2.1 Data Overview/Sample Collection

Methods for sample collection, nanopore sequencing and bioinformatic processing are provided in Chapter 2, section 2.2. This subsection overviews sample collection and sequencing methods specific to this chapter.

Head kidney tissue was harvested from Atlantic salmon parr 24 hours after intraperitoneal injection with either PBS (control, n=6) poly I:C (n=6) or inactivated *Vibrio* (n=6). Poly I:C is a viral PAMP and was used to mimic viral infection, representing a synthetic double-stranded RNA detected by viral pattern recognition receptors, which potently activates inflammatory responses and the interferon pathway in teleost fish (Wang et al., 2024). An inactivated strain of *Vibrio,* the causative agent of vibriosis (Lages et al., 2019), was used to stimulate an antibacterial, proinflammatory response as a source of bacterial PAMPs. Total RNA was extracted via standard TRIzol total RNA extraction. As detailed in Chapter 2, section 2.2.1, sample collection and total RNA extraction was conducted by Dr Shahmir Naseer, Dr Thomas Clark and Professor Samuel Martin at the University of Aberdeen. The methods for poly I:C stimulation and sample collection are also published in Clark et al. (2023).

mRNA was isolated from the total RNA samples using the Dynabeads mRNA Purification Kit (Invitrogen 61006) before a pooled, barcoded sequencing library was generated using the Direct cDNA-Sequencing Kit (SQK-DCS109) with Native Barcoding (EXP-NBD196) from Oxford Nanopore Technologies. The library was sequenced on the PromethION device for 72 hours.

To obtain full-length reads, samples were basecalled with Guppy v6.3.7 (ONT), low-quality reads (q<7) were filtered out with NanoFilt v2.7.1 (De Coster et al., 2018) before demultiplexing with Guppy v6.3.7. Full-length reads were identified, oriented and trimmed of barcodes and sequencing adapters with Pychopper v2.5.0 (ONT).

To visualise the relationship between common transcript and gene models expressed by samples in the three treatments, the UpSetR v1.4.0 (Conway et al., 2017; Gehlenborg, 2019) R package in RStudio (R4.3.3) was used to generate an UpSet plot at the gene-level and transcript-level. An UpSet plot is an alternative to Venn diagrams, which allows the intersections between sets of data to be plotted in a matrix layout (Lex et al., 2014). This makes it suitable for displaying the number of transcript or gene models in the final transcriptome that were expressed in each treatment group, or different combinations of treatment groups.

3.2.2 Transcript Quantification

Full-length reads from all biological replicates in the three treatment groups were mapped against the long-read transcriptome generated in Chapter 2 using Minimap2 v2.24 (Li, 2021) with options "`-ax map-ont -N100 -t 8`". Reads were mapped back to the long-read transcriptome rather than an existing short-read annotation to quantify the expression of novel transcripts identified in Chapter 2. After .sam files were converted to .bam files using SAMtools v1.13 (Danecek et al., 2021), non-primary alignments (secondary, supplementary and non-mapped) were removed using "`samtools view -b -F 2308`", before sorting with the sort function of the SAMtools suite.

Salmon v1.8.0 (Patro et al., 2017) in 'alignment-based' mode with options "`-p 8, -l A, --ont`" was used for transcript-level quantification due to its

ability to accurately quantify transcript expression from long-read RNA-seq data rapidly with low computational resources (Corley et al., 2019; Dong et al., 2023). Salmon outputted a single quant.sf file for each biological replicate detailing the counts and transcripts-per-million (TPM) for each transcript. A custom bash script was used to restructure the data into two files: 1) counts for all biological replicates, and 2) TPMs for all biological replicates. These output files were imported into R for differential expression analysis.

3.2.3 Data Exploration and Quality Check

Counts from Salmon were imported into R and filtered with a custom R script to retain only transcript models which had five or more counts in at least 4/6 replicates for any single treatment group. I elected to use strict filtering criteria to reduce noise and keep only transcripts showing robust expression, decreasing false positives. The filtered dataset was transformed using the variance stabilising transformation (Anders & Huber, 2010) function, "`vst(, blind = FALSE`)" from the DESeq2 R package v1.42.1 (Love et al., 2014). The "`plotPCA()`" function from DESeq2 was used to generate a principal component analysis (PCA) matrix. PCA produces a series of principal components which attempt to explain variation in the dataset and estimate the contribution of each principal component to that variation. Plotting the a PCA in two dimensions allows for visualisation of similarity in transcript expression between treatments, including whether biological replicates group together within treatments. A PCA plot of the first two principal components was generated using ggplot2 v3.5.0 (Wickham, 2016).

In addition to PCA, sample-similarity matrices allow for an alternative visualisation of sample/treatment similarity through a series of pair-wise comparisons between all samples. The DESeq2 R package was used to calculate sample-to-sample Euclidean distances and then hierarchical clustering based on these distances was applied. The R package pheatmap v1.0.12 (Kolde, 2019) was used to plot the distance matrix heatmap.

3.2.4 Differential Transcript Expression

DTE analysis was carried out on the filtered count data with DESeq2. The standard DESeq2 pipeline was used, treating individual transcript count data as if they were genes, an established approach for short-read DTE analysis (Soneson et al., 2015). Short-read DTE pipelines conduct a transcript-length correction to account for short-read alignments rarely covering the entire length of a transcript. Thus, multiple reads are required to make up a single full-length transcript. In comparison, in long-read RNA-seq one read does represent a full transcript and therefore, no transcript-length correction was conducted for the long-read count data. The minimum count threshold for removing lowly expressed transcript models was based on short-read DGE thresholds in order to increase robustness of the DTE analysis. To explain this, imagine a hypothetical transcript 1kb in length. This transcript would require 7 short-read transcripts (assuming 150bp length) to capture the full-length of the RNA. However, 7 full-length reads mapping to this hypothetical model would yield ~7kb of sequencing data. Thus, using the same read count thresholds for long-read DTE analysis means that significantly more per-base-level data is required to keep a transcript in the analysis compared with short-read analyses.

The dispersions of the DESeq2 model were plotted using a custom R script in ggplot2. Two iterations of the workflow were carried out, one to compare poly I:C vs. PBS, and a second to compare *Vibrio* vs. PBS. The Wald test in DESeq2 was used to identify differentially expressed transcripts (DETs). Transcripts showing adjusted (Benjamini-Hochberg i.e. FDR) p-values < 0.05 and a log2 fold change threshold of ±1 (a doubling or halving of expression) were considered DETs between control and treatment groups. To reduce noise from lowly expressed transcripts and aid visualisation, log2 fold changes were shrunk with the "`lfcShrink()`" function in DESeq2, with the default "`apeglm`" method. MA and volcano plots for each comparison were plotted with ggplot2 to visualise the relationship between i) read counts and fold change, and ii) fold change and significance.

For each comparison, DETs from the long-read transcriptome were ordered by adjusted p-value and the top twenty unique genes possessing at least one DET extracted into a table. The link between these genes and Ensembl annotated genes was done using the outputs of SQANTI3 (Tardaguila et al, 2018; Pardo-Palacios et al., 2024a) from Chapter 2. Using the biomaRt v2.58.2 R package (Durinck et al., 2009), gene symbols for the genes with associated Ensembl IDs were added to the table if present in the Ssal_v3.1 reference annotation. If the associated Ensembl gene ID had no gene symbol, a custom loop script employing biomaRt was used to assign a gene symbol, if present, representing the predicted orthologue in rainbow trout, zebrafish, mouse (*Mus musculus*) and human (*Homo sapiens*) in that order of priority. In other words, if no gene symbol was present in the Ensembl salmon annotation, the orthologous gene name from trout was used. If no orthologous gene name existed in trout, the zebrafish one was used, and so on.

To assign symbols/annotations to genes that lacked: i) an Ensembl gene symbol/name after the above steps, ii) any Ensembl gene ID (i.e. novel genes in the long-read transcriptome), open-reading frames (ORFs) and coding sequences (CDSs) were extracted from the transcript model sequences in the long-read transcriptome via the TAMA GO ORF-seeker tool included in the TAMA suite (Kuo et al., 2017). A custom protein database was generated by combining the UniProt protein databases for Atlantic salmon (UP000087266), rainbow trout (UP000193380), zebrafish (UP000000437), mouse (UP000000589) and human (UP000005640). DIAMOND (Buchfink et al., 2021) with options "`--evalue 1e-10 --max-target-seqs 10`" and used to BLAST the translated protein sequences from the transcript CDSs against the combined database. The top hit for each gene was used to assign a gene symbol to genes not present in the Ensembl annotations. Finally, for each gene, the number of transcripts passing filtering and the number of DETs was calculated using a custom R script and added to the table. A dotplot to visualise the number of annotated transcripts per gene vs. the number of DETs was plotted using ggplot2.

The samples used in this study were previously used for differential gene expression analysis using short-read RNA-seq (Clark et al. 2023), contrasting the poly I:C vs control groups. To compare the results from long-read and short-read RNA-seq, I overlapped DETs with Ensembl gene IDs from the poly I:C vs. control long-read comparison with the equivalent list of short-read DEGs from Clark et al. (2023) using RStudio.

3.2.5 Gene Ontology Analysis

Gene Ontology (GO) enrichment analysis was performed separately for the poly I:C vs. control and *Vibrio* vs. control contrasts, to determine biological processes overrepresented among DETs. Only transcripts with Ensembl transcript IDs annotated by SQANTI3 were used. The R package AnnotationForge v1.44.0 (Carlson & Pagès, 2024) was used to create a custom R .db object from the Ssal_v3.1 reference transcriptome annotation extracted from Ensembl using biomaRt. The "`enrichGO()`" function in the clusterProfiler R package v4.10.1 (Yu et al., 2012; Wu et al., 2021) was used to perform GO enrichment tests. The full set of filtered transcripts possessing associated Ensembl transcript IDs was set as the background and the analysis was restricted to GO terms in the category 'Biological Processes'. The "`dotplot()`" function, in conjunction with ggplot2, was used to plot all the enriched GO terms overrepresented in each treatment group.

3.2.6 Differential Transcript Usage

DTU analysis was conducted in R using the DRIMSeq v1.30.0 package (Nowicka & Robinson, 2016). DRIMSeq aims to determine whether the proportions of transcript expression differ between treatments within a gene, and which transcripts contribute to those changes. As with the DTE analysis (section 3.2.4), two iterations of DRIMSeq were performed to compare each immune stimulation group with the control group. The set of transcripts filtered for DTE analysis were further processed to remove singlets, i.e. transcripts representing the only transcript produced by a gene, using the "`dmFilter()`" function. A histogram showing the distribution of the number of transcripts per gene input into DRIMSeq was plotted using the "`plotData()`" function and customised with ggplot2. In order, the DRIMSeq

functions "`dmPrecision()`", "`dmFit()`" and "`dmTest()`" were used with default parameters to carry out DTU analysis. To improve FDR control, a post-hoc filtering step was carried out as per Love et al. (2018).

Next, the R package stageR v1.24.0 (Van den Berge et al., 2017) was used to screen the genes identified as showing evidence of DTU, and confirm the transcripts contributing to that DTU event. An overall FDR threshold of adjusted $p < 0.05$ was set for the stageR confirmation process. The "`stageRTx()`" and "`stageWiseAdjustment()`" functions in the stageR package were used to carry out the analysis using the approach detailed in Love et al. (2018). The DRIMSeq "`plotProportions()`" function in conjunction with ggplot2 and a custom colour palette was used to generate a ribbon plot for gene models displaying significant evidence of DTU.

3.2.7 Exploration of Alternative Transcript Expression

Following DTU analysis with DRIMSeq and stageR, an alternative approach was used to identify gene models displaying alternative transcript expression in response to each immune stimulus. I focussed on identifying genes possessing: 1) DETs unique to either poly I:C or *Vibrio* responses (vs. controls), 2) DETs derived from the same gene expressed in both treatment groups, and 3) DETs originating from genes deemed to be novel by SQANTI3. This included antisense transcripts or transcripts belonging to gene models in intergenic or intronic space not currently annotated by Ensembl.

For candidate genes of interest, the expression level of each DET and non-DET passing initial filtering was plotted using ggplot2. TPMs for each biological replicate were extracted from the TPM output file from Salmon (section 3.2.2), plotted in a dotplot, and coloured with the same custom colour palette used in the PCA and UpSet plots. A custom R script was used to add a symbol to each transcript's dotplot to denote if it was a DET, and if that transcript was identified as a DET in only one treatment group or both.

The R package ggtranscript v0.99.9 (Gustavsson et al., 2022) was used to visualise transcript structures for long-read defined genes of interest, it's

associated gene in the Ensembl Ssal_v3.1 annotation, and to display predicted promoter sequences from the Ensembl Ssal_v3.1 regulatory build, a comprehensive annotation of regulatory elements such as promoters and enhancers. Gtf files for the long-read transcriptome annotation and the Ssal_v3.1 annotation, and a gff3 file for the Ssal_v3.1 regulatory build were imported into RStudio using the rtracklayer v1.62.0 R package (Lawrence et al., 2009). Coordinates of each feature were extracted from the gtf and gff3 files using a custom R script. UTRs were partitioned in the Ensembl annotations to show the coding-sequences in the reference models, and the canonical transcript marked with an asterisk. Both the dotplots and transcript model plots were exported as svg files and imported into InkScape to combine into a single figure.

3.3 Results

3.3.1 Data Overview and Quality Assessment

Of the 266,222 transcripts and 35,480 gene models described in the long-read transcriptome (Chapter 2), 137,169 transcripts and 27,291 gene models were supported by at least three full-length reads originating from the combined immune challenge dataset. 22,730 of the gene models (83.3%) were supported by reads originating from all treatment groups, with proportionately few models specific to a single group (Figure 3.1A). In contrast, at the transcript level, only 31.0% of the transcript models were supported by evidence from all treatments. 64,094 transcript models were supported by reads originating from a single treatment group (Figure 3.1B).

Each sample in the head kidney dataset possessed more than 600,000 full-length reads (Table 3.1) which satisfies the ENCODE consortium long-read RNA-seq sequencing depth threshold for “Good” data (ENCODE Project Consortium, 2025). 20,739 transcripts derived from 4,530 unique genes passed the count filtering stage, thus representing a robust set of transcripts for DTE analysis. 1,050 (23.2%) and 1,064 (23.5%) of the genes in poly I:C and *Vibrio* groups had their most highly expressed transcript (based on mean TPM value across all 6 replicates) match the Ensembl canonical transcript model. PCA of this data revealed clear segregation of biological replicates

into their respective treatments (Figure 3.2). Most variation was explained by the first two PCs; 60.6% for PC1, which largely separated poly I:C from control and *Vibrio* samples, and 14.5% for PC2, which clearly segregated control from *Vibrio* samples, while also capturing variation within the poly I:C samples.

Following hierarchical clustering of samples, the sample-similarity matrix showed that biological replicates cluster within their respective treatment groups except for one poly I:C replicate, which grouped with the PBS control group (Figure 3.3). It is possible that this replicate did not mount a strong response to the poly I:C stimulation. This sample was not removed from the analyses due to its separation from the PBS group in the PCA. Additionally, the sample-similarity matrix supports the PCA analysis with PBS and *Vibrio* replicates grouping together more closely than with the poly I:C group.

The dispersion plot generated after running DESeq2 shows a decrease in dispersion as the mean expression increases with a tight spread of data following the fitted line (Figure 3.4). This indicates that data filtering was successful and provided a robust set of transcripts for DTE analysis.

3.3.2 Response to Viral Mimic Challenge

A strong transcriptomic response was observed following challenge with poly I:C, with 1,096 DETs derived from 462 unique genes. Of these, 1,046 were upregulated > 2-fold, whilst 50 were downregulated > 2-fold. A large proportion of DETs (189/1,096; 17.2%) were upregulated beyond an estimated log2 fold change of 5, translating to >32-fold increase in expression (Figure 3.5). Strong upregulation of transcripts originating from known antiviral genes including *isg15, dhx58, aste1a,* and *cmpk2,* as well as genes associated with the interferon and JAK/STAT pathway including *ncoa7* and *tasl* was observed (Table 3.2). Indeed, one transcript isoform of *isg15* showed a log2 fold change of approximately 8, representing a 256-fold increase in expression. DETs with the greatest fold changes included transcripts derived from *ubil, isg15, scyb7 and ifit9* genes. All of the unique genes possessing the most significant DETs were upregulated in response to poly I:C. The 50 downregulated DETs included transcripts derived from

genes including *gp182, cd79a, id3, hspa8, nr2f5* and *aplnrb.* The number of DETs per gene initially increased as the number of transcripts passing filtering increased, however, genes with greater than 30 transcripts tended to represent a small proportion of all possible transcripts classified as DETs (Figure 3.6).

DETs were split into upregulated and downregulated subsets to identify biological processes enriched in both sets. For the upregulated DETs, GO enrichment analysis revealed strong over-representation of many biological processes (Figure 3.7) including “immune response” (GO:0006955), explained by transcripts derived from genes *ccl19, cxl10* and *m17,* “defense response” (GO:0006952) and “response to biotic stimulus” (GO:0009607) explained by isoforms of *rsad2* (aka: viperin), a key ISG with many antiviral functions (Ghosh & Marsh, 2020). Other enriched processes included “response to stress” (GO:0006950), “regulation of apoptotic processes” (GO:0042981), and “protein ubiquitination” (GO:0016567). 5 transcripts from 5 unique genes (ENSSSAG00000009861 – ENSSSAT00000030432; ENSSSAG00000063668 – ENSSSAT00000170540; ENSSSAG00000092333 – ENSSSAT00000188457; ENSSSAG00000054169 – ENSSSAT00000086241; ENSSSAG00000241361 – ENSSSAG00000048880) explained the term “protein ubiquitination”, all without gene names in the Ensembl reference annotation. A BLAST search of the reference cDNA sequences (extracted from BioMart online; Durinck et al., 2009) showed that these genes were *herc3, nuclear factor 7, a ring2A-like gene,* and an *ankyrin repeat and IBR domain-containing protein 1.* One gene was uncharacterised in both Ensembl and RefSeq annotations. Interestingly, no biological processes were enriched for the downregulated DETs.

370 unique Ensembl gene IDs were associated with the 1,096 DETs. Among these genes, 265 (72%) overlapped with 2,466 DEGs (poly I:C vs. controls) identified by short-read RNA-seq in Clark et al. (2023). Investigating the genes represented in my long-read data revealed some interesting differences with the short-read data. For example, Atlantic salmon possess two ohnologues of the gene *trim25,* located on collinear blocks 2q-12qa (Lien

et al., 2016), which is involved in ubiquitination during early the immune response (Martín-Vicente et al. 2017). Only a single ohnologue of *trim25* is significantly expressed in the short-read dataset (ENSSSAG00000054152, chr2), whilst my long-read method identified significant upregulation of transcripts derived from both ohnologues (also including ENSSSAG00000046838, chr12). Another example is *rsad2*, which has 4 copies in the Atlantic salmon genome, 2 ohnologues on chromosomes 1 and 9 (collinear blocks 1p-9qa; Lien et al., 2016), as well as an additional two copies located in tandem on chromosome 9. Clark et al. (2023) identified upregulation of both ohnologues and a single tandem gene in response to poly I:C (chr1: ENSSSAG00000048046, chr9: ENSSSAG00000108937 & ENSSSAG00000108840) yet transcripts from all 4 copies of *rsad2* were upregulated according to my long-read approach.

3.3.3 Response to Bacterial Challenge

Fewer transcripts were significantly regulated by the *Vibrio* treatment than for poly I:C, specifically 563 upregulated and 64 downregulated DETs. Among the most significant DETs were transcripts derived from genes associated with the inflammatory response including*, saa, hamp, steap4*, *acod1, socs3a, il-1rii, c209e and igfbp6* (Table 3.3). With the exception of *ccl19, c209e, sat1* and *gabbr-2,* all transcripts deriving from the genes listed in Table 3.2 were identified as DETs showing a strong response to the *Vibrio* stimulation. As with the poly I:C challenge, the most significant DETs were all upregulated, however, a less pronounced response was observed in response to *Vibrio* with only 23 DETs (3.7%) possessing a log2 fold change ≥ 5 (Figure 3.8). Despite fewer DETs, more downregulated transcripts were identified in the *Vibrio* dataset, dominated by transcripts deriving from myeloperoxidase gene *mpx* (ENSSSAG00000048994), which is secreted by neutrophils and plays a role in host defence to microbial pathogens (Kettle et al., 1993; Khan et al., 2018). Interestingly one transcript derived from *rtp2* was highly downregulated following *Vibrio* challenge (log2FC = -3.7) but upregulated by poly I:C. A similar trend was seen in the ratio of DETs to transcripts per gene in the *Vibrio* group as for poly I:C (Figure 3.9).

GO enrichment analysis revealed only two enriched biological processes in the *Vibrio* upregulated DETs: GO:0006955 “immune response” and GO:0002376 “immune system processes”, explained by transcripts originating from genes involved in innate immunity including *ccl19, c7b, il10, il1fma, ck-2.1,* and *pglyrp5*. 9 biological processes were enriched in the downregulated DETs including “regulation of gene expression” (GO:0010468), “regulation of macromolecule (GO:0010556) and cellular (GO:0031326) biosynthetic processes”, and terms associated with metabolism including “regulation of primary (GO:0080090), cellular (GO:0031323) and macromolecule (GO:0060255) metabolic processes”. All 9 enriched processes in the downregulated DETs were explained by the same 3 transcripts, 1 each from the glucocorticoid receptor encoding gene *nr3c1,* the translation repressor gene *4ebp*, and *spi1a*, encoding a key transcription factor for immune cell development.

3.3.4 Differential Transcript Usage

DTU analysis was conducted to identify significant changes in the proportion of transcript expression within individual genes. After filtering out genes with only one expressed transcript, 18,278/18,233 transcripts from 2,070/2,068 genes were retained in the poly I:C/*Vibrio* groups. In both sets, just over half of the genes had only 2 or 3 transcripts (1,076/2,070 for poly I:C, 1,077/2,068 for *Vibrio*), whilst only 22 genes had greater than 75 transcripts in both treatment groups (Figure 3.10).

Despite striking evidence for a strong response to both viral and bacterial stimuli, only 9 genes showed significant evidence of DTU by DRIMSeq and stageR, with 11 unique transcripts showing significant changes in their within-gene expression proportion (Table 3.4). Of these, only 2 were identified as DETs. G13151 was the only gene identified as possessing evidence of DTU in both poly I:C and *Vibrio* treatment groups (Figure 3.11). Indeed, G13151.3 was also identified as a DET displaying significant upregulation in response to both immune treatment groups (Figure 3.12A).

SQANTI3 classified G13151.3 as a fusion transcript between *cdk17* (ENSSSAG00000082189) and an Ensembl reference gene with no

orthologous gene name (ENSSSAG00000103330). Examining the structure of all transcripts originating from G13151 reveals that only transcript G13151.3 shares significant overlap with *cdk17* (Figure 3.12B), with the remainder of the transcripts possessing a TSS further downstream. Both alternative start sites are supported by promoters from the Ensembl regulatory build. Examining the same region in the NCBI RefSeq annotation showed significant overlap between G13151.3 and a different gene, *pctk2* (RefSeq gene ID: 100380451), however, there was limited support in either reference annotation for a gene located at the downstream TSS. Furthermore, a BLAST search aligned G13151.3 with the CDS of *pctaire2*, another name for *pctk2*, which is a FIP-2 like gene described by Collins et al. (2007) upregulated in response to amoebic infection. The high expression level of G13151.3, its supporting promoter region upstream of the TSS, and the potential immune function of the overlapping gene *pctk2* suggests that 1) this transcript variant has a significant role in the innate immune response across distinct pathogen classes, including parasitic infection (Collins et al., 2007) and in response to viral and bacterial stimuli, and 2) that the reference annotations are incomplete for this gene.

3.3.5 Altered Expression of Novel Transcript Variants in Response to Viral and Bacterial Stimulation

DTE analysis revealed significant expression of novel alternative transcript forms in response to both poly I:C and *Vibrio* treatments. The expression of several classes of alternative transcript not currently annotated by Ensembl were significantly altered in each treatment group; as elaborated by examples that follow.

Intron retention was a common source of variation among alternative transcripts in both treatment groups. For example, DETs possessing intron-retention events between exons 2-3 and 4-5 of *lgals17*: ENSSSAG00000078726 (G23085.2, .4, .6, .7 and .8, Figure 3.13), a gene recently identified as belonging to a major QTL associated with resistance to tilapia lake virus in Nile tilapia (Barría et al., 2021), were upregulated in response to poly I:C. Intron retention was observed in other poly I:C-

upregulated DETs, including between exons 2-3 and 4-5 in the ISG *cd9:* ENSSSAG00000059637 (G29720.2, Figure 3.14) and exons 1-2 in transcripts G29007.2 and G29007.4 derived from another ISG, *aste1:* ENSSSAG00000048960 (Figure 3.15; Levraud et al., 2019). Many of the intron-retention events occur in the coding sequences of the Ensembl reference annotations, potentially leading to changes in peptide sequence and subsequent protein function. However, a good example of an intron retention event in a UTR is the *Vibrio*-specific DET G21850.3, a novel transcript derived from *il-rii* (ENSSSAG00000069905), a gene encoding interferon receptor 1, which was upregulated by *Vibrio* (Figure 3.16).

Among the DETs were numerous transcripts defined by putative alternative TSS. A single TSS producing 3 transcripts of *cd9:* ENSSSAG00000059637 is annotated by Ensembl, supported by a promoter annotation. In contrast, 22 unique transcripts originating from at least 5 distinct putative TSSs, including the Ensembl-annotated promoter region, were significantly upregulated in response to poly I:C in my long-read data (Figure 3.14). Compared to the Ensembl annotation, a putative novel alternative TSS was observed in *il-rii* (Figure 3.16), whilst 2 putative novel alternative TSSs were among the poly I:C downregulated DETs from *tmem106a* (Figure 3.17), a gene shown to limit the release of enveloped viruses from cell surfaces (Mao et al., 2022).

An interesting finding was the identification of immune-regulated intronless transcripts in my long-read data that are not annotated by Ensembl. Examples can be seen in transcript G26969.3 derived from *sat1* ENSSSAG00000000816 (*Vibrio* upregulated) (Figure 3.18), G29007.5 from *aste1* ENSSSAG00000048960 (*Vibrio* upregulated) (Figure 3.15), and G23085.9 from *lgals17* ENSSSAG00000078726 (poly I:C upregulated) (Figure 3.13). All of these example mono-exonic transcripts span multiple exons of the gene, suggesting that they are not exonic fragments or sequencing artifacts, and have biological relevance. SinEx DB 2.0 is a database containing single-exon coding genes for 10 mammalian species (Jorquera et al., 2021). The data and methods presented in this chapter have

the capacity to extend this type of database beyond mammals to additional non-model species.

Novel alternative exonic chaining events were also associated with immune DETs. For example, *Vibrio* upregulated transcript G4497.1 has 8 exons, which aligns closely with the two Ensembl reference transcripts for *hspa5* ENSSSAG00000054661, encoding a heat shock protein 70 involved in clearance of unfolded proteins and reduction of inflammation (Corrigall et al., 2004; Fu et al., 2023). In contrast, *Vibrio* upregulated transcript G4497.7, which consists of 5 exons with similar exon boundaries to G4497.1, lacks exons 4-6 of G4497.1 (Figure 3.19) which fall in the Ensembl coding sequence, potentially leading to altered protein structure and function. Additionally, both exon-skipping and a novel exon splice site are evident in *Vibrio* upregulated transcript G4497.47. Interestingly, both G4497.7 and G4497.47 show similar TPM values to upregulated *hspa5* transcripts including a full repertoire of coding exons, hinting at biological activity.

Novel exon-skipping and alternative exon chaining events can also be seen in transcripts from *cd9* (ENSSSAG00000059637 Figure 3.14) and its ohnologue (ENSSSAG00000079939) found in collinear blocks 3q-6p (Lien et al., 2016) which were both upregulated in response to poly I:C (Figure 3.20). 6 paralogues of *cd9*, a member of the tetraspanin family, exist in salmonid genomes, split into three groups of 2 ohnologue pairs each (Dehler et al., 2023). The third group, *cd9c,* has been shown to have significant roles in the interferon response to viral infection in teleosts (Briolat et al., 2014; Dehler et al., 2019) and Dehler et al. (2023) showed that the *cd9c* ohnologue pair was highly upregulated in response to viral haemorrhagic septicaemia virus (VHSV), whilst *cd9a* was downregulated and *cd9b* had no change. The two long-read *cd9* ohnologues upregulated in response to poly I:C stimulation (G29720, Figure 3.14; G25771, Figure 3.20), overlap the same *cd9c* ohnologue pair upregulated in response to VHSV, however, no evidence of *cd9a* downregulation was found.

In all the above examples, long-read RNA-seq has successfully identified a wealth of alternative transcripts that show significant changes in expression

in response to PAMPs that are not currently annotated by the Ensembl Ssal_v3.1 Atlantic salmon annotation.

3.3.6 Common Alternative Transcript Regulation in Response to Viral and Bacterial Stimulation

The innate response to bacterial and viral pathogens shares many of the same pathways and genes, for example the interferon response (Perry et al., 2005; Mertowska et al., 2023). My dataset revealed 75 genes with shared DETs between treatment groups. For example, one transcript variant G10941.4, derived from *igfbp6* ENSSSAG00000117627*,* a member of the insulin growth factor binding proteins (Macqueen et al., 2013) was commonly upregulated in response to both poly I:C and *Vibrio* challenge. However, an additional 4 novel DETs were specific to the *Vibrio* group, including an intron retention event in G10941.2, a novel UTR splice site in G10941.1 and G10941.3, and a novel exon start site in G10941.5 (Figure 3.21). Past work into duplicated copies of *igfbp6* revealed that one paralogue (*igfbp6-a2)* was commonly upregulated in response to both bacterial and viral infection in rainbow trout (Alzaid et al., 2016), however, interrogating the NCBI RefSeq annotation reveals G10941 to be annotated as a different paralogue, *igfbp6-b1* (RefSeq: NM_001123650.1).

As a distinct example, a DET (G11846.6) derived from *rtp2* ENSSSAG00000112886, closely related to the Ensembl canonical transcript, was upregulated by poly I:C, but downregulated by *Vibrio* (Figure 3.22). All other transcripts of *rtp2* were upregulated in poly I:C only, but the greatest upregulation in transcript G11846.6. In contrast, DETs deriving from a paralogue of *rtp2* ENSSSAG00000085797 only showed significant upregulation in response to poly I:C infection (Figure 3.23). Whilst the Ensembl canonical transcript of *pim1* ENSSSAG00000066472*,* a known ISG involved in the interferon response (Ko et al., 2022) was upregulated in both poly I:C and *Vibrio* groups, 4 novel transcript variants were also DETs in response to poly I:C only, with evidence of novel exon-skipping in G10669.2 and G10669.7 and intron retention events in G10669.10 and G10669.11 (Figure 3.24). G10669.2 also shows evidence of a novel 3’ UTR splice site.

3.3.7 Identification of Novel Gene Expression

Long-read RNA-seq also allowed the expression regulation of novel gene models to be examined. G22520 has two DETs, both upregulated in response to *Vibrio* infection, and classified as antisense to *saa* ENSSSAG00000100178*,* itself a highly upregulated antibacterial gene (Figure 3.25). There is no current Ensembl gene model on the antisense strand to *saa.* The long-read antisense *saa* transcripts overlap the 5' end of the sense strand *saa* and have 10-fold lower expression than the sense transcripts. Further examination revealed that transcripts overlapping the sense strand *saa* gene were classed as originating from the same long-read gene, however, there appears to be a distinct split in loci between these transcripts. For example, G25519.1 and G25519.2 overlap the 5' end of *saa*, while G25519.14 and G25519.16 overlap the 3' end. Interestingly, the transcripts located at the 3' end of the reference model were classed as overlapping *saa5* ENSSSAG00000069990 (Figure 3.25) despite deriving from the same long-read gene as those overlapping *saa*.

Some of the most highly upregulated transcripts in the *Vibrio* group belong to novel gene G15937, absent in the Ensembl annotation. Examining this novel gene model revealed 2 transcripts upregulated in both treatment groups and 15 only in response to *Vibrio* (Figure 3.26); all transcripts for this gene were significantly upregulated. The gene resides on chromosome 2 and was predicted to have protein-coding isoforms by SQANTI3. This gene was predicted in the NCBI RefSeq annotation (XM_014157806.2; "Carcinoembryonic Antigen-Related Cell Adhesion Molecule 20" or *ceacam20*) as a single transcript gene compared with the 17 DETs identified by long-read RNA-seq.

3.4 Discussion

This chapter reports a comprehensive transcript-level resolved analysis of the response to viral and bacterial PAMPs in a non-model fish species. By adapting established analysis tools, I developed a bioinformatic approach for conducting DTE with long-read RNA-seq and identified a wealth of previously unannotated transcript diversity regulated by immune system stimulation.

3.4.1 Resolution of Transcript-Level Expression

DGE analysis with short-read RNA-seq is well established for studying immune function in salmonids. However, the role of individual transcripts had not been studied before this work. Many of the identified DETs derived from DGEs identified in matched samples by short-read RNA-seq (Clark et al., 2023). Further, my finding that *cd9* upregulation specific to the poly I:C stimulated group only occurred in one of three *cd9* ohnologues is consistent with past work in rainbow trout (Dehler et al. 2023). The overlap between my results and past work provides confidence in my quantitative analysis pipeline. In contrast, the strong upregulation of transcripts deriving from *igfbp6* paralogue *igfbp6-b1* in response to both viral and bacterial mimic stimulation is not commensurate with the findings of Alzaid and colleagues (2016) who found only paralogue *igfbp6-a2* to be upregulated in the innate immune response to both types of pathogens. Further research is needed to elucidate paralogue-specific transcript expression dynamics of the *igfbp6* gene family.

Another interesting finding was the downregulation of two transcripts of the *cd79a* gene, considered to be a salmon pan-B-cell marker (Peñaranda et al., 2019), in response to poly I:C stimulation. This finding is in direct opposition to a recent study in rainbow trout which found elevated expression of *cd79* in the head kidney following viral, bacterial and parasitic infection (Cheng et al., 2021). It is possible that gene-level expression analysis masks transcript-specific expression patterns, and more research into the function of downregulated transcripts in the innate immune response is warranted.

However, long-read RNA-seq adds novel layers of functional annotation. Past research on the role of transcript diversity in the immune response has focussed on mammalian systems. For example, in humans, Inamo and colleagues (2024) identified unannotated transcripts with links to Alzheimer's and systemic lupus erythematosus, whilst the prevalence of alternative splicing in cancer is well described (Zhang et al., 2021). In fish, rates of alternative splicing were elevated in Nile tilapia *Oreochromis niloticus* exposed to cold (Li et al., 2020a) indicating that transcript-level expression is

intrinsically linked with stress responses. In a similar way, the ability of my long-read approach to delineate transcript expression forms a valuable resource for examining transcript diversity and studying its function in the immune system of Atlantic salmon. Further functional studies of the identified DETs in this chapter will lead to a deeper understanding of the immune response in finfish, a key area for aquaculture research.

3.4.2 Annotation of Transcript Expression in Novel Genes

Aquaculture research increasingly relies on comprehensive genome assemblies with rich annotation for identification of functional elements related to traits of interest (Houston et al., 2020). One of the most common uses of long-read RNA-seq is to identify novel transcripts and genes. My work identified a handful of DETs originating from genes not annotated by Ensembl, including a novel antisense gene overlapping *saa*, upregulated in response to *Vibrio*. Antisense genes are often long non-coding RNAs (lncRNA) with roles in the regulation of gene expression (Pelechano & Steinmetz, 2013; Wight & Werner, 2013; Brophy & Voigt, 2016) and those that overlap a sense gene on the opposite strand are termed Natural Antisense Transcripts (NAT; Khorkova et al., 2014). NATs can regulate expression of the sense gene through a variety of mechanisms including the tethering of repressive polycomb complexes which silence transcription in that region (Katayama et al., 2005), transcription of the NAT causes RNA polymerases to bind to the antisense strand, thus blocking transcription of the sense gene (Wahlestedt et al., 2013), and once transcribed, the NATs can hybridise with sense mRNA forming structures which are then cleaved, thus reducing translation (Werner et al., 2014). Antisense genes are thought to regulate the innate immune system in Atlantic salmon (Tarifeño-Saldivia, 2017). The discovery of the novel *saa* antisense gene shows that my full-length approach can identify novel loci as well as quantify their expression. This could potentially lead to the discovery of further antisense genes, and other lncRNAs, of biological relevance. However, further functional studies to determine how antisense copies of immune-related genes regulate their expression or have additional biological roles are warranted.

I also captured the upregulation of a potential fusion transcript (G13151.3) in response to both viral and bacterial stimulation. Neither the Ensembl nor RefSeq annotations captured this event, with each differing in its classification of gene models in this region. The fact that the fusion transcript overlapped a FIP2-like gene with putative immune function (Collins et al., 2007) indicates this transcript has biological relevance. Much research into novel diagnostic approaches and development of disease control methods rely on comprehensive annotations of functional genomic elements. However, in the case of G13151, the discrepancies between both reference assemblies means that reference-based studies could potentially overlook this transcript. As such, the annotation of novel transcripts and genes using long-read RNA-seq could expand the immune repertoire available for biomarker discovery and therapeutic treatment targets.

3.4.3 Future Versatility of Long-Read Approach

The approaches developed in this chapter have broad applications in future studies of salmonid immunity. Fish have adaptive immunity (Robertsen, 2018) which is key to successful vaccination outcomes (Tammas et al., 2024). My work focussed on the short-term response to viral and bacterial PAMPs, and did not involve actual pathogen infections with a long-enough timecourse to resolve the adaptive phase of immunity. Methods outlined in this chapter can be readily transferred to understand the importance of transcript-level gene regulation in adaptive immune responses, both to aquatic pathogens, and following vaccination (Gunter et al., 2023). Such work would benefit from longer timecourses to resolve the kinetics of transcript expression responses during multiple phases of the immune response, more akin to the experimental design used in my next chapter to explore embryogenesis. In addition, whilst I focussed on the innate immune response in the head kidney, the primary haemopoietic organ in salmonids, applying my methods to other organs with immune functions such as liver (Taylor et al., 2022) and intestine (Kortner et al., 2024) could provide evidence for tissue-specific transcript diversity and its expression. This would further our knowledge of the transcriptional basis underlying immune function in salmonids.

Overall, this chapter offers a valuable resource for annotation of transcripts important to the immune response in salmonids, a key topic for aquaculture research which often relies on comprehensive genome annotations.

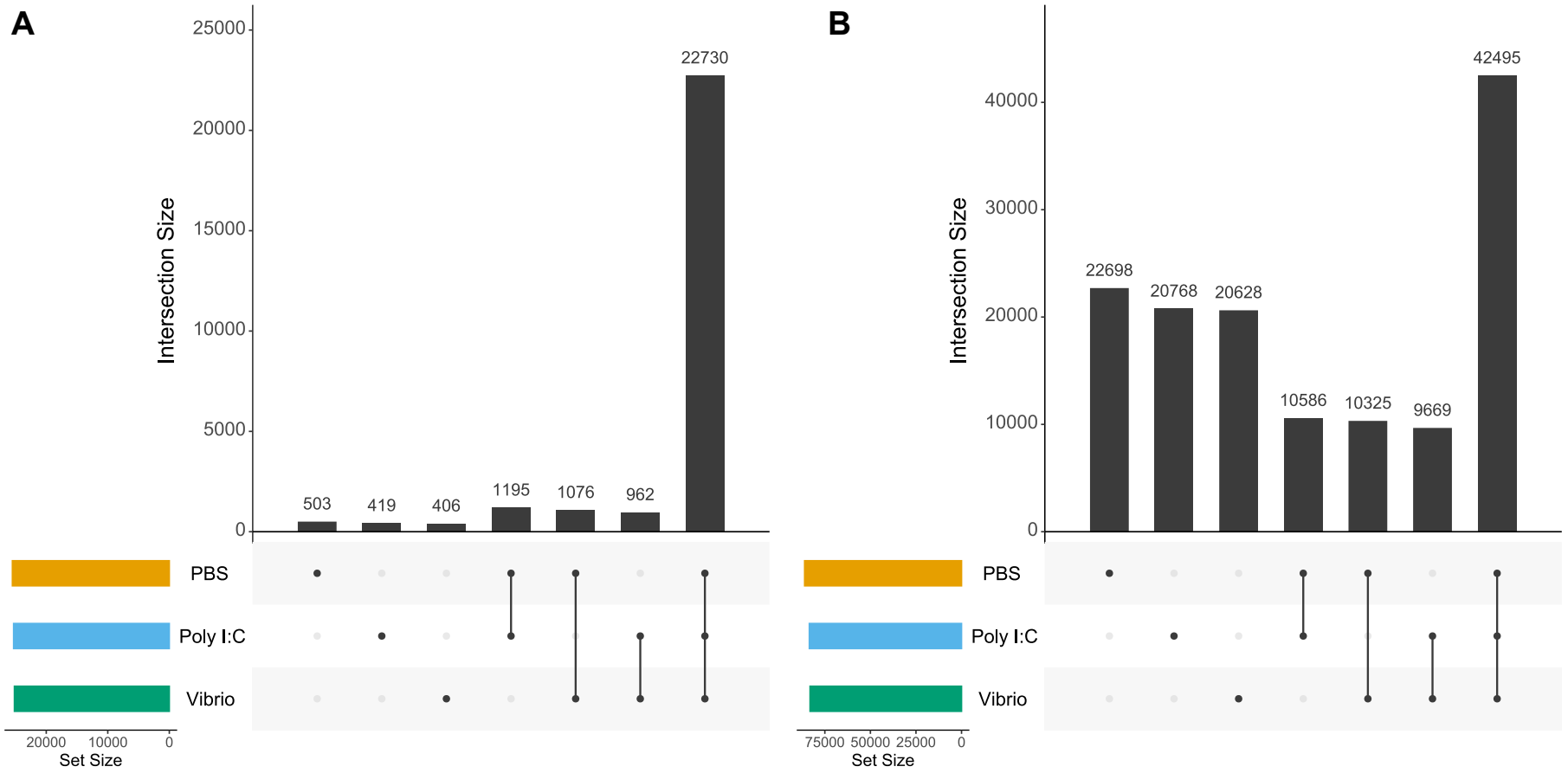


Figure 3.1: UpSet plots showing the number of genes (A) and transcripts (B) supported by full-length nanopore reads from the immune challenge dataset. The sets on the left represent the control group (PBS) and the two treatment groups (poly I:C & Vibrio). The dots at the bottom represent the combinations of treatment groups, whilst the bars show the number of reads in each combination of group.

Table 3.1: Number of reads for each head kidney sample during transcriptome assembly pipeline

Sample	Treatment Group	Number of Filtered Reads (q>7)	Number of Full-Length Reads	Number of Trimmed Full-Length Reads	Number of Primary Alignments
PBS1	PBS	2,075,669	725,511	724,398	551,967
PBS2		2,262,803	918,446	916,529	641,169
PBS3		2,148,130	745,534	744,424	572,017
PBS4		2,142,565	750,874	749,816	582,748
PBS5		2,230,777	764,400	763,332	589,508
PBS6		2,208,416	831,777	830,354	612,567
PolyIC1	Poly I:C	2,162,794	782,245	780,810	576,241
PolyIC2		2,245,582	790,591	789,141	577,148
PolyIC3		2,470,250	925,274	922,937	637,633
PolyIC4		1,972,589	770,722	769,214	548,772
PolyIC5		2,238,209	823,733	822,382	624,419
PolyIC6		1,831,583	717,757	715,565	448,593
Vib1	Inactivated *Vibrio anguillarum*	2,179,914	717,164	716,233	566,080
Vib2		2,198,447	769,824	768,383	570,758
Vib3		2,047,480	758,464	757,068	547,772
Vib4		2,179,068	869,828	867,730	584,605
Vib5		2,018,907	781,413	779,775	555,705
Vib6		1,720,663	633,931	632,706	462,714

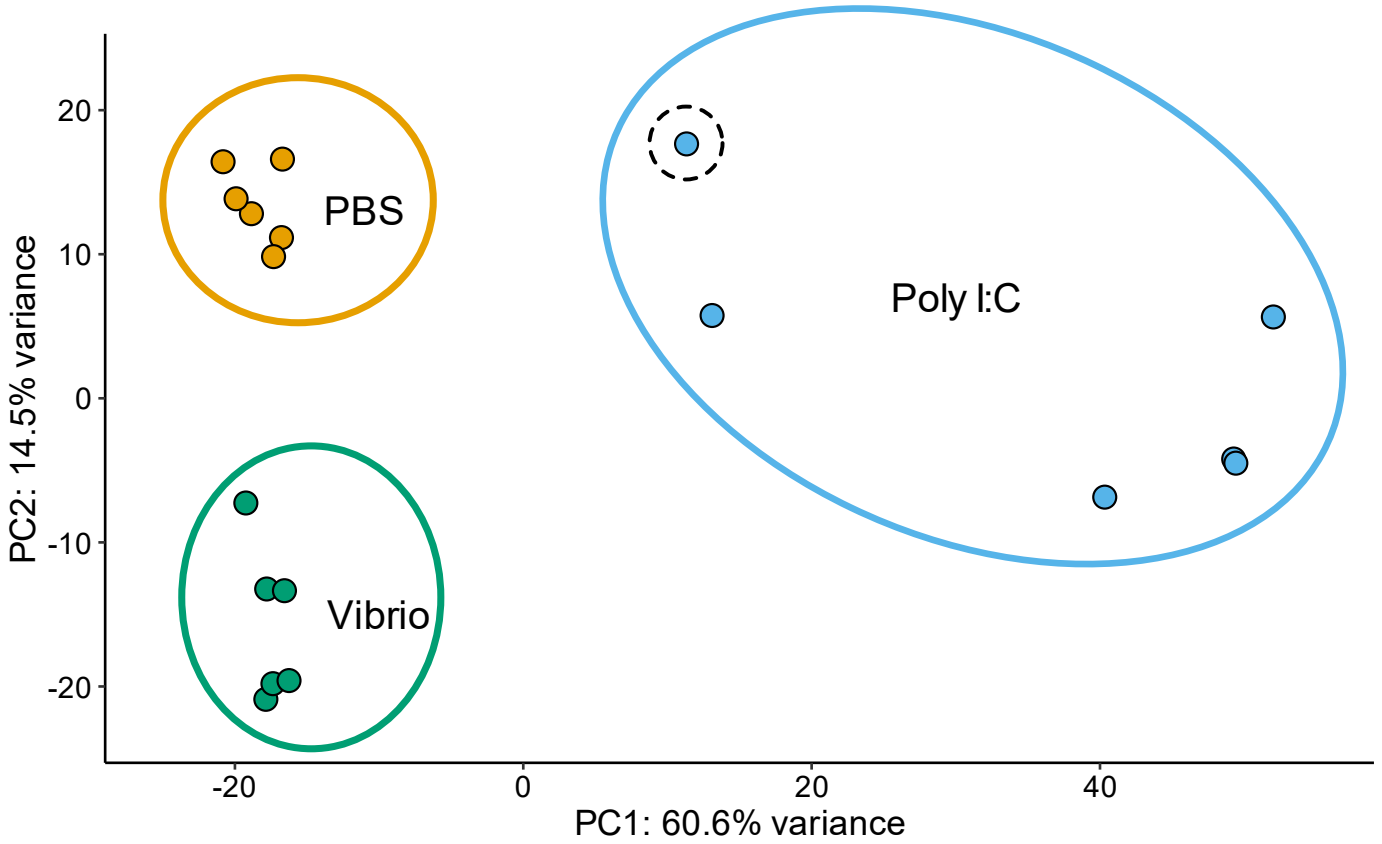


Figure 3.2: PCA for immune challenge dataset with three treatment groups: PBS control (orange), poly I:C (blue) and Vibrio (green). The poly I:C replicate ringed with a black dotted line groups with PBS samples in a sample similarity matrix plot (see Figure 3.3).

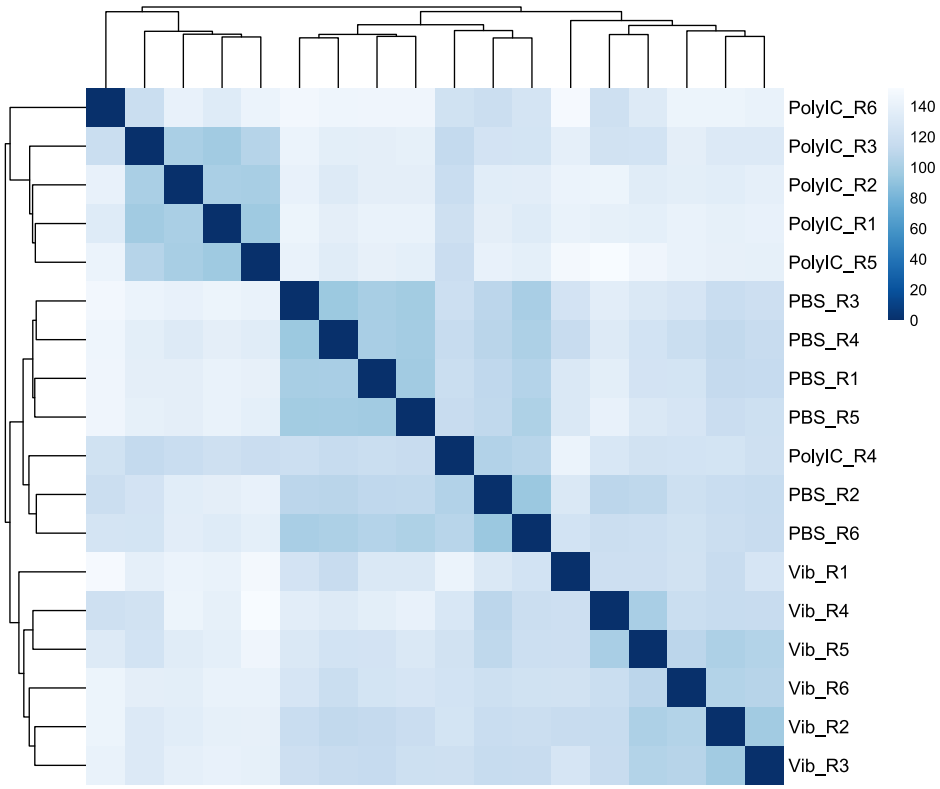


Figure 3.3: Sample similarity matrix plot of all samples for the three treatment groups (n=6): PBS control, poly I:C, and Vibrio. Data displayed is Euclidean distance and dendrogram based on hierarchical clustering. All samples group together within their respective treatments except polyIC_R4, which groups with the PBS controls.

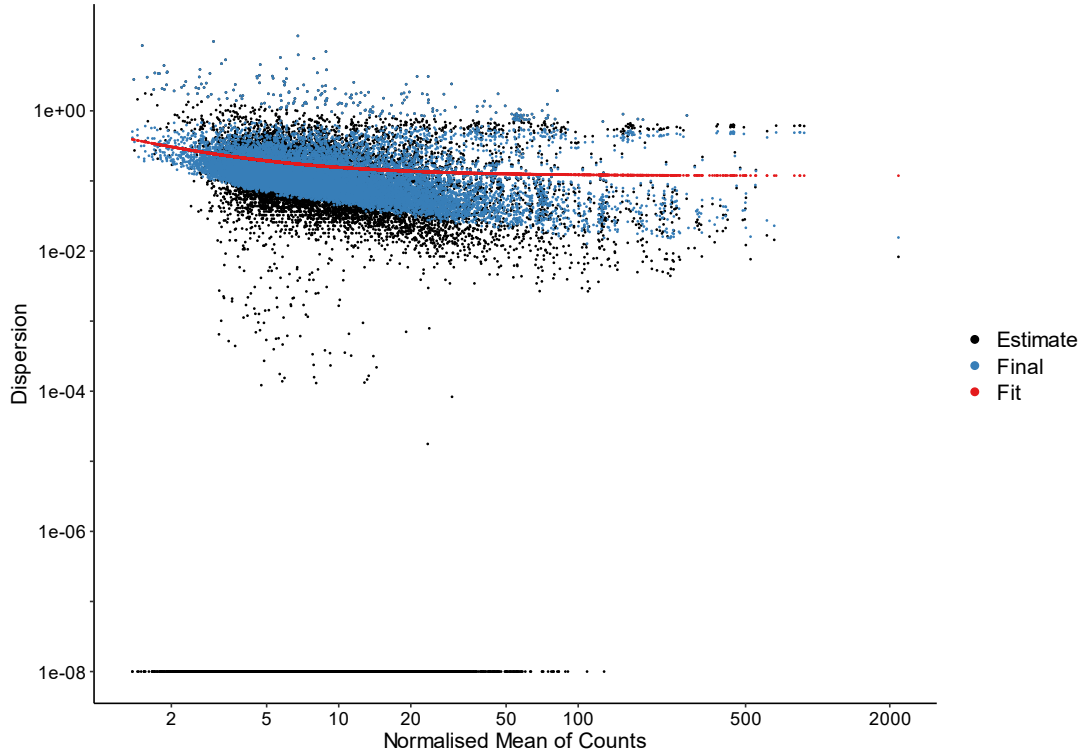


Figure 3.4: Plot of dispersion estimates based on within-group mean read counts. Each black dot represents a transcript and its initial maximum likelihood dispersion estimate. The red line is a curve fitted to the initial dispersion estimates, whilst blue dot are transcripts shrunk towards the fitted curve.

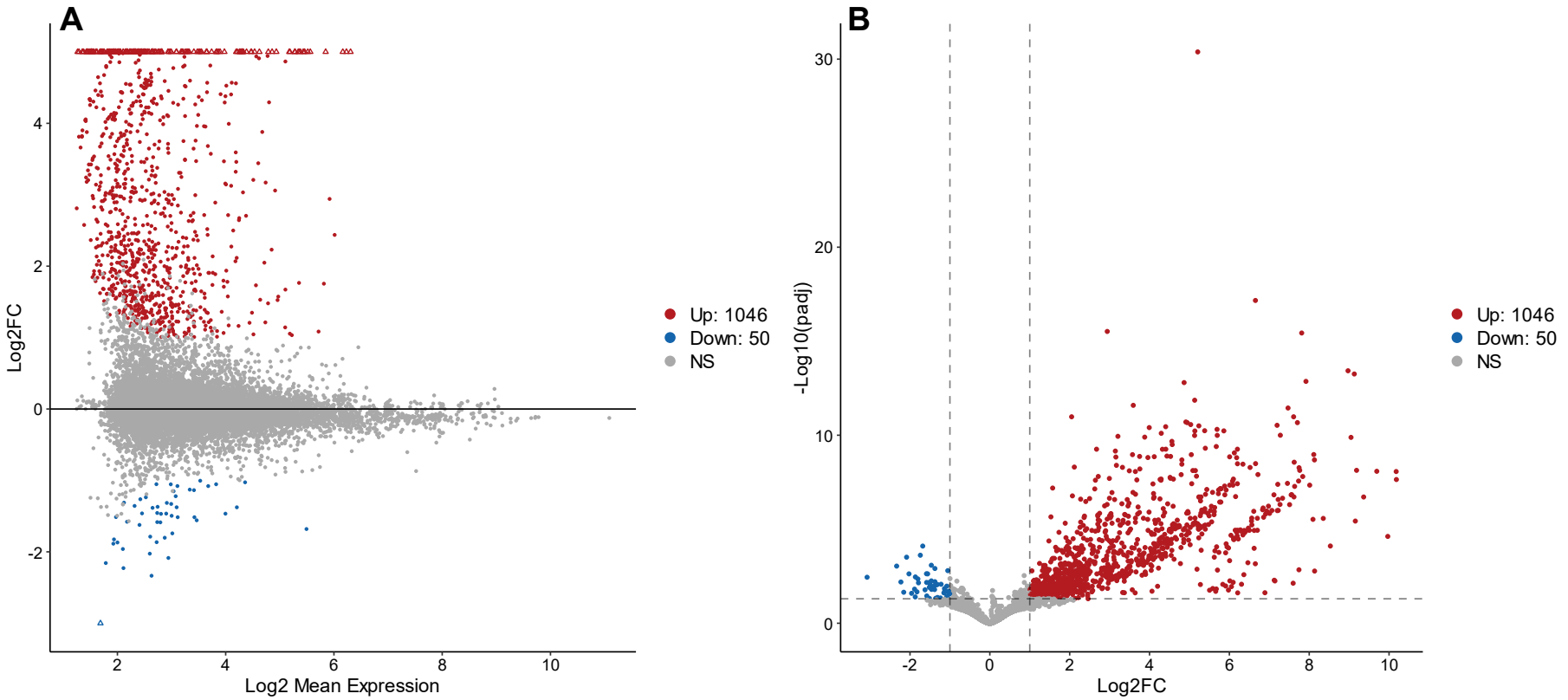


Figure 3.5: MA and volcano plots for the poly I:C group. (A) MA plot showing log2 fold change vs log2 mean expression; red/blue denotes up/down regulated transcripts, and grey shows transcripts that were not significantly differentially expressed. Triangle shapes are used for transcripts with log2 fold changes >5 or < -3. (B) Volcano plot showing -log10 adjusted p-values vs log2 fold change. The colours are the same as in (A), grey dotted lines represent FDR adjusted p-value and log2 fold change thresholds of ≤0.05 and ≥ ±1 respectively.

Table 3.2: Details of the top 20 unique genes with DETs showing the lowest adjusted p-values in response to poly I:C challenge. Gene name source is indicated by the following symbols: [Ss] = Atlantic salmon Ensembl annotation, [Om] = rainbow trout Ensembl orthologue, [Dr] = zebrafish Ensembl orthologue, [Mm] = mouse Ensembl orthologue, [Hs] = human Ensembl orthologue. Average Log2FC represents the mean log2 fold change of all DETs deriving from their respective Gene IDs.

Gene ID	Locus	Associated Ensembl ID	Gene Name	Description	Function	# Transcripts (DETs)	Average Log2FC
G11436	Chr16: 26,681,602 - 26,683,447	ENSSSAG00000051856	*pex11a* [Ss]	Peroxisomal Membrane Protein 11A	Involved in peroxisome elongation. Peroxisomes play key immunometabolic roles in mammals (Di Cara et al., 2023)	6 (6)	4.68
G11847	Chr16: 51,827,384 - 51,832,270	ENSSSAG00000085797	*rtp2* [Mm]	Receptor Transporter Protein 2	Previously unknown immune function in mammals and fish, upregulated in response to bacterial infection in Atlantic salmon (Eslamloo et al., 2020)	9 (9)	5.89
G15720	Chr2: 14,338,782 - 14,343,469	ENSSSAG00000048950	*aste1a* [Dr]	Asteroid Homolog 1	Highly upregulated during infection with nervous necrosis virus (NNV) in Asian seabass *Lates calcarifer* (Liu et al., 2016)	6 (6)	5.35
G16078	Chr2: 29,314,242 - 29,318,031	ENSSSAG00000051388	*ncoa7* [Mm]	Nuclear Receptor Coactivator 7	Regulation of RNA transcription. Conserved ISG between salmonids and humans (Clark et al., 2023)	1 (1)	3.99
G17779	Chr20: 29,236,981 - 29,410,445	ENSSSAG00000064740	*isg15* [Ss]	ISG15 Ubiquitin-Like Modifier	Involved in the interferon pathway, upregulated early during antiviral immune response in Atlantic salmon (Kileng et al., 2007; Andresen et al., 2020)	4 (4)	7.01
G17787	Chr20: 29,409,469 - 29,410,698	ENSSSAG00000121778	*isg15* [Dr]	ISG15 Ubiquitin-Like Modifier	Involved in the interferon pathway, upregulated early during antiviral immune response in Atlantic salmon (Kileng et al., 2007; Andresen et al., 2020)	4 (4)	5.38

Gene ID	Locus	Associated Ensembl ID	Gene Name	Description	Function	# Transcripts (DETs)	Average Log2FC
G17788	Chr20: 29,410,525 - 29,421,041	ENSSSAG00000067410	*pxmp2* [Dr]	Peroxisomal Membrane Protein 2	Pore-forming activity in peroxisomal membrane (Rokka et al., 2009)	1 (1)	4.41
G20738	Chr23: 42,758,091 - 42,766,071	ENSSSAG00000110347	*trex1,* *trex2* [Mm]	Three Prime Repair Exonuclease 1 or 2	Involved in DNA repair. Lack of Trex gene expression associated with inflammatory responses in mammals (Namjou et al., 2011)	22 (21)	5.51
G21070	Chr24: 15,924,576 - 15,925,934	ENSSSAG00000067408 ENSSSAG00000105254	*ubil* [Ss] *ubil* [Ss]	Ubiquitin-Like Protein	Key role in immune response through ubiquitination of pattern recognition receptors (Zinngrebe et al., 2014)	4 (4)	8.48
G21072	Chr24: 15,933,351 - 15,934,678	ENSSSAG00000098221 ENSSSAG00000067408	*ubil* [Ss] *ubil* [Ss]	Ubiquitin-Like Protein	As per G21070	4 (4)	8.47
G21075	Chr24: 15,942,124 - 15,944,078	ENSSSAG00000063700 ENSSSAG00000105519	*svop* [Oo] *ubil* [Ss]	Synaptic Vesicle 2 Related Protein Ubiquitin-Like Protein	SV2 protein family involved in transmembrane transporter activity in neurological activity (Rossi et al., 2022) As per G21070	8 (8)	7.80
G21999	Chr25: 38,243,003 - 38,245,270	ENSSSAG00000048770	*tasl* [Mm]	TLR Adaptor	Involved in innate immune response. In mammals, Tasl interacts with toll-like receptors and influences activation of interferon pathway (Heinz et al., 2020)	3 (3)	2.87
G23681	Chr28: 6,958,286 - 6,982,992	ENSSSAG00000068373	*ch25h* [Ss]	Cholesterol 25-Hydroxylase	Proposed antiviral activity in teleost fish including rainbow trout and common carp (Adamek et al., 2021)	2 (2)	4.62

Gene ID	Locus	Associated Ensembl ID	Gene Name	Description	Function	# Transcripts (DETs)	Average Log2FC
		novelGene_ENSSSAG00000051405_AS	*N/A*	Novel Antisense Gene in LR Transcriptome	Novel antisense gene of ENSSSAG00000051405, orthologous to IL-13 in Northern pike *Esox lucius*		
G25189	Chr3: 37,984,737 - 38,018,439	ENSSSAG00000116694	*N/A*	*C-C Motif Chemokine 2**	Unclear if Ligand or Receptor from BLAST search. C-C motif chemokine 2 suggested to be involved with inflammation in Atlantic salmon (Grimholt et al., 2015)	13 (13)	3.31
		novelGene	*N/A*	Novel Gene in LR Transcriptome	Transcribed intergenic region at 3' UTR of ENSSSAG00000116694 (see rest of row). Predicted to be non-coding by SQANTI3		
G25590	Chr3: 60,694,664 - 60,700,401	ENSSSAG00000003156	*dhx58* [Ss]	DExH-Box Helicase 58	Involved in antiviral innate immune response in teleosts. Regulator of RIG-like receptor pathway promoting interferon production (Zhao et al., 2023)	1 (1)	5.21
G29720	Chr6: 34,792,706 - 34,803,941	ENSSSAG00000059637	*cd9* [Ss]	CD9 (Tetraspanin Family)	Regulator of innate immune response and inflammation (Brosseau et al., 2018). Upregulated in response to viral infection in rainbow trout (Dehler et al., 2023)	37 (22)	1.58
G32049	Chr9: 10,498,454 - 10,510,125	ENSSSAG00000007886	*cmpk2* [Ss]	Cytidine/Uridine Monophosphate Kinase 2	Maintenance of intracellular UTP/CTP, , also acknowledged as an ISG displaying upregulation in response to viral stimuli (Liu et al., 2019) in teleosts	9 (9)	4.67
G32269	Chr9: 24,171,385 - 24,194,408	ENSSSAG00000044215	*N/A*	*Sacsin**	Predicted to encode sacsin by BLAST search. Sacsin upregulated in grass carp *Ctenopharyngodon idella* in response to reovirus infection (Dai et al., 2017)	17 (17)	4.39

Gene ID	Locus	Associated Ensembl ID	Gene Name	Description	Function	# Transcripts (DETs)	Average Log2FC
G8785	Chr14: 37,620,574 - 37,641,911	ENSSSAG00000105208	*N/A*	*Uncharacterised*	Novel gene in Ensembl annotation, no BLAST results. Predicted to have coding potential by SQANTI3	20 (20)	3.40
		novelGene_ENSSSAG00000089678_AS	*N/A*	Novel Antisense Gene in LR Transcriptome	Novel antisense gene. Antisense of ENSSSAG00000089678, classified as lncRNA in Ssal_v3.1 Ensembl annotation		
		novelGene	*N/A*	Novel Gene in LR Transcriptome	Transcribed intergenic region of ENSSSAG00000089678. Some transcripts protein-coding, but product uncharacterised. Others non-coding		
G9741	Chr15: 4,165,138 - 4,167,253	ENSSSAG00000106328	*scyb7*[Ss]	Platelet Basic Protein	Incorrect annotation of gamma IP-encoding gene *γip*. Showed increased expression following vaccination with inactivated Salmon Pancreatic Disease virus (Collins et al., 2021)	2 (2)	7.50

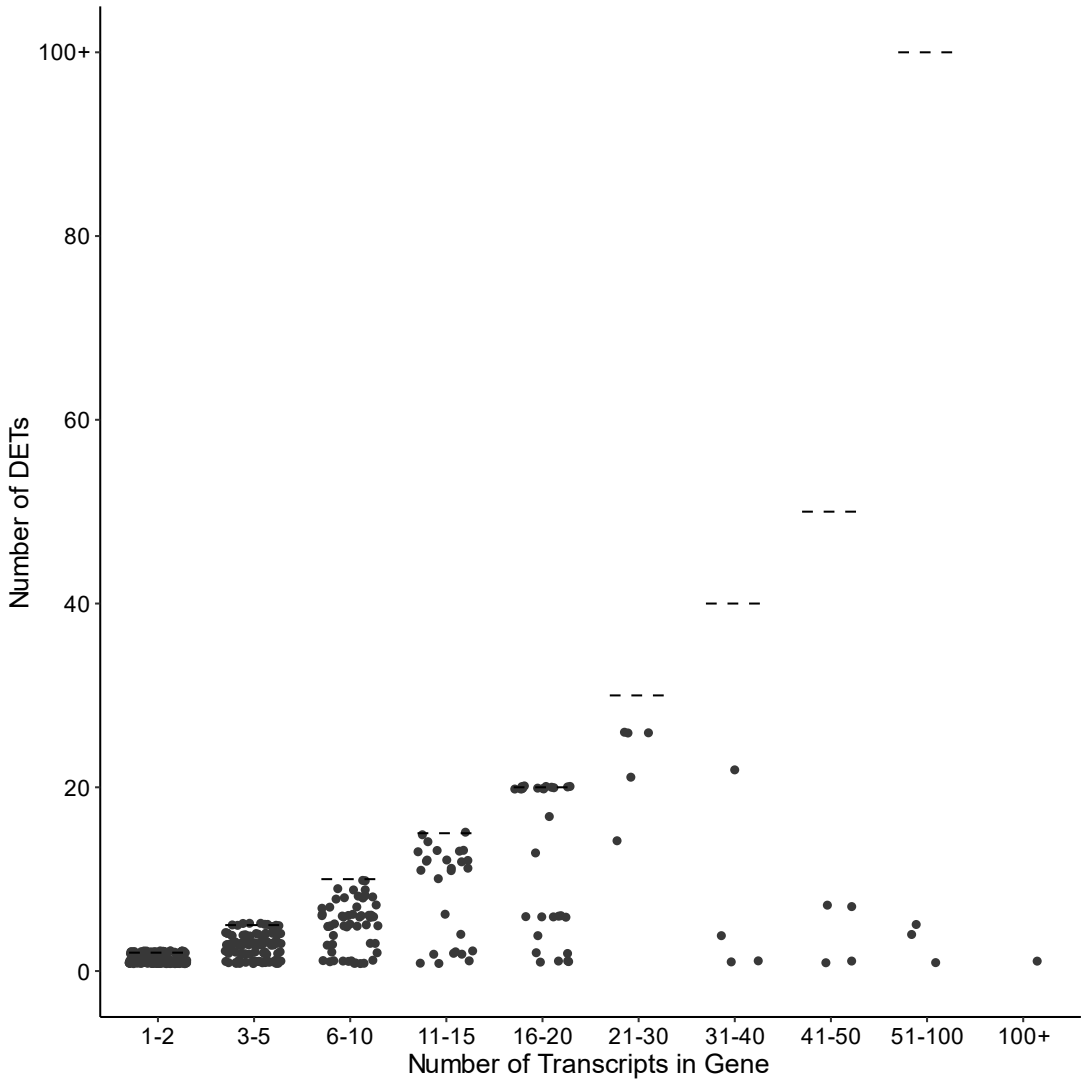


Figure 3.6: Dotplot showing number of DETs versus the number of filtered transcripts (filtering retained only transcripts with ≥5 reads in at least 4 out of 6 biological replicates) per gene in the poly I:C group. Dashed lines indicate the maximum number of possible DETs for each bin, i.e. the point at which all transcripts per gene would be classified as DETs.

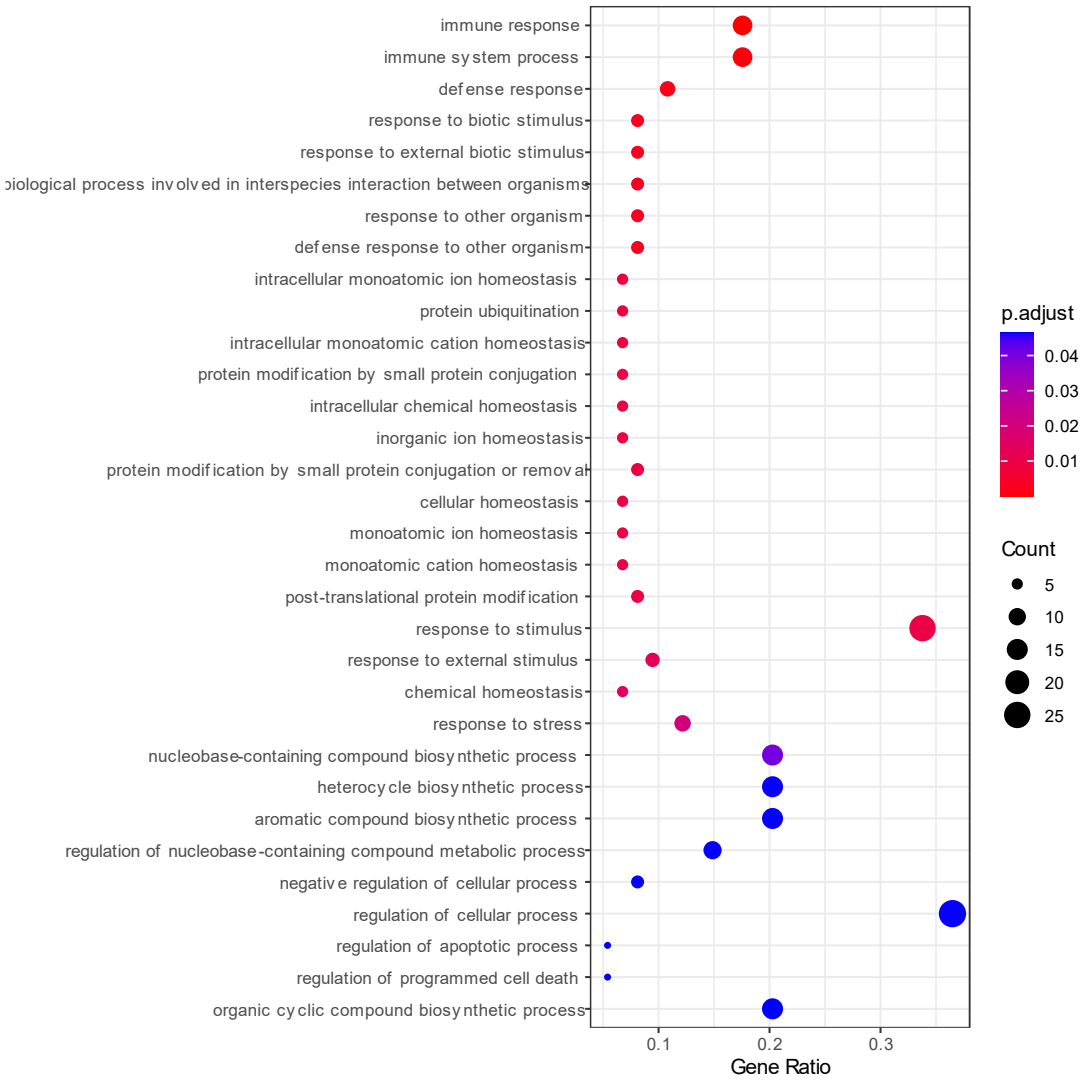


Figure 3.7: Dotplot of enriched GO terms ("Biological Processes") for the upregulated DETs in the poly I:C group. Y-axis denotes terms with the size of each dot indicating the number of DETs supporting each term and colour indicating adjusted p-value, as per the key on the right.

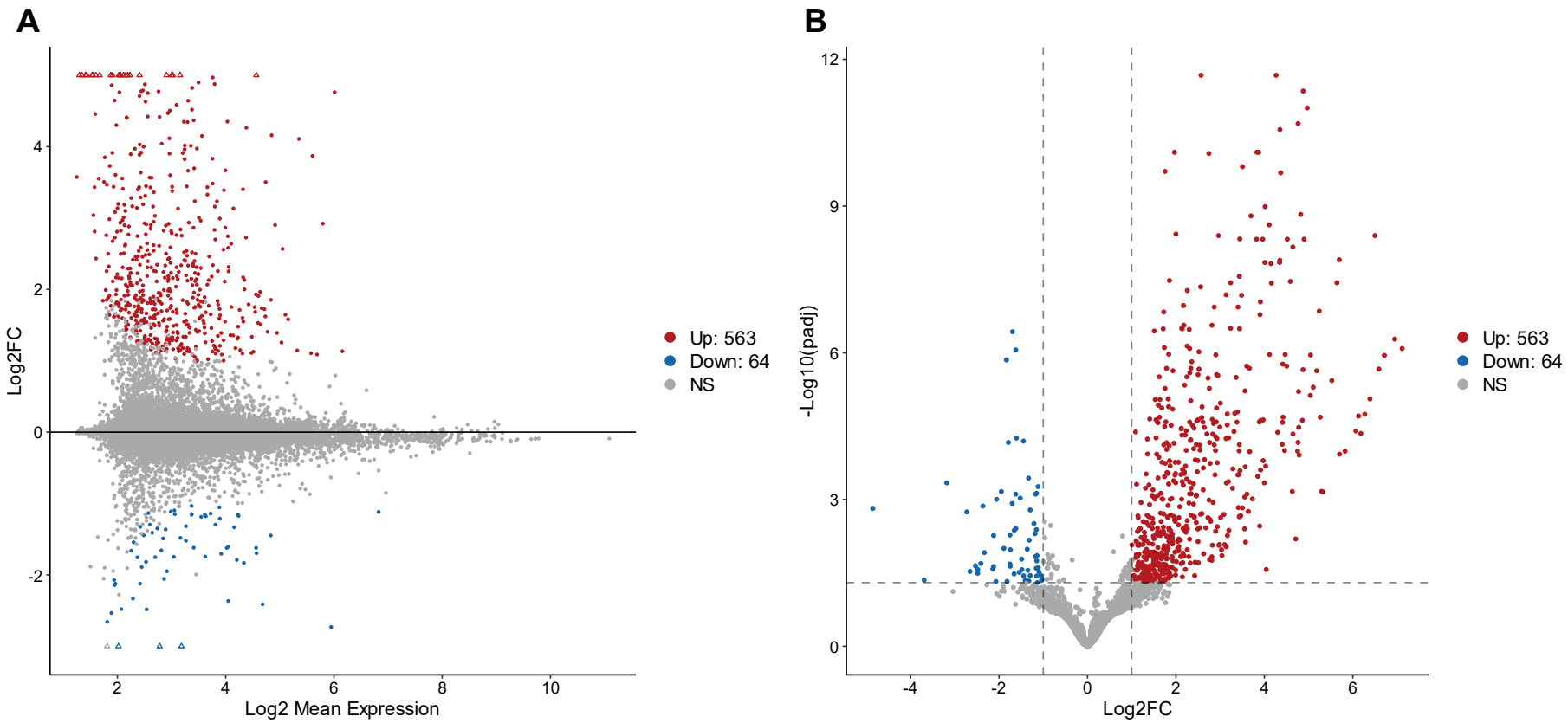


Figure 3.8: MA and volcano plots for the Vibrio group. (A) MA plot showing log2 fold change vs log2 mean expression; red/blue denotes up/down regulated transcripts, and grey shows transcripts that are not significantly differentially expressed. Triangle shapes used for transcripts with log2 fold changes >5 or < -3. (B) Volcano plot showing -log10 adjusted p-values vs log2 fold change. The colours are the same as in (A), grey dotted lines represent FDR adjusted p-value and log2 fold change thresholds of ≤0.05 and ≥ ±1 respectively.

Table 3.3: Details of the top 20 unique genes with DETs showing the lowest adjusted p-values in response to Vibrio challenge. Gene name source is indicated by the following symbols: [Ss] = Atlantic salmon Ensembl annotation, [Om] = rainbow trout Ensembl orthologue, [Dr] = zebrafish Ensembl orthologue, [Mm] = mouse Ensembl orthologue, [Hs] = human Ensembl orthologue. Average Log2FC represents the mean log2 fold change of all DETs deriving from their respective Gene IDs.

Gene ID	Locus	Associated Ensembl ID	Gene Name	Description	Function	# Transcripts (DETs)	Average Log2FC
G10312	Chr15: 52,296,583 - 52,307,490	ENSSSAG00000120636_ ENSSSAG00000063834	N/A	Novel Fusion Gene in LR Transcriptome (Fusion of hsp90aa1.1 & hsp90aa1.2)	*hsp90aa1.1* and *hsp90aa1.2* are tandem duplicates of *hsp90aa1* gene. This suggests a fusion event between the two duplicates	5 (5)	4.17
		ENSSSAG00000063834	*hsp90aa1.2* [Ss]	Heat Shock Protein 90 Alpha Family Class A Member 1	Member of the heat shock protein 90 (HSP90) family. Found to regulate viral replication via JAK/STAT pathway during innate response in mammals (Liu et al., 2022a)		
G10941	Chr15: 92,790,563 - 92,795,420	ENSSSAG00000117627	*igfbp6* [Ss]	Insulin-Like Growth Factor Binding Protein 6	Upregulated in Atlantic salmon muscle cell cultures following inflammation (Pooley et al., 2013). Increase in IGFBP-6 expression acts as a negative regulator of IGF-I and II activity and is marker of inflammation (Alzaid et al., 2016)	5 (5)	2.72
G12552	Chr17: 13,368,090 - 13,370,128	ENSSSAG00000089279	N/A	Phospholipase A2 Inhibitor 31 kDa Subunit-Like*	Subunit of a PLA2 inhibitor protein first identified in cobra sp. *Naja kaouthia* (Ohkura et al., 1994)	4 (4)	3.20
G13153	Chr17: 58,518,654 - 58,525,214	ENSSSAG00000117100	N/A	Hemagglutinin/Amebocyte Aggregation Factor-Like*	Found to be upregulated in Atlantic salmon gill upon reinoculation with *Paramoeba perurans* (McCormack et al., 2021)	3 (3)	2.77

Gene ID	Locus	Associated Ensembl ID	Gene Name	Description	Function	# Transcripts (DETs)	Average Log2FC
G13158	Chr17: 58,592,692 - 58,599,828	ENSSSAG00000096534	N/A	Hemagglutinin/Amebocyte Aggregation Factor-Like*	Found to be upregulated in Atlantic salmon gill upon reinoculation with *Paramoeba perurans* (McCormack et al., 2021)	3 (3)	2.95
G13899	Chr18: 29,198,645 - 29,232,117	ENSSSAG00000113880	*comtd1* [Ss]	Catechol-O-Methyltransferase Domain Containing 1	Conserved gene involved in inactivation of neurotransmitters including dopamine (Vidgren et al., 1994)	5 (5)	3.71
G1577	Chr1: 135,771,301 - 135,772,638	ENSSSAG00000002773	*ccl19* [Ss]	Chemokine (C-C motif) Ligand 19	Expressed in Atlantic salmon head kidney and interbrachial lymphoid tissue (Bjørgen et al., 2019). Upregulated in Atlantic salmon spleen in response to bacterial infection (Sun et al., 2024)	9 (8)	2.82
G15937	Chr2: 23,384,855 - 23,398,482	novelGene	*ceacam20**	Novel Gene in LR Transcriptome – *Carcinoembryonic Antigen-Related Cell Adhesion Molecule 20**	Novel gene absent from Ensembl annotation. BLAST search (NCBI database) suggests gene to encode CEACAM20. Part of immunoglobulin superfamily conserved in vertebrates. Proposed function in mucosal and intestinal immunity in mammals (Kelleher et al., 2019)	17 (17)	2.14
G16368	Chr2: 42,218,127 - 42,222,194	ENSSSAG00000076658	c209e [Ss]	C-Type Lectin Receptor	Transmembrane pathogen recognition. Stimulation of cytokine production and C-type lectin receptors (CLRs) involved in cell necrosis during inflammation in mammals (Drouin et al., 2020)	43 (41)	1.77
G18656	Chr21: 696,131 - 701,658	ENSSSAG00000031095	*acod1* [Ss]	Aconitate Decarboxylase 1	Downregulation of acod1, also known as irg1, shown in Atlantic salmon post amoebic infection (Talbot et al., 2021). A key regulator of immunometabolism during inflammatory response (Wu et al., 2020)	5 (5)	3.07

Gene ID	Locus	Associated Ensembl ID	Gene Name	Description	Function	# Transcripts (DETs)	Average Log2FC
G21850	Chr25: 27,816,391 - 27,823,456	ENSSSAG00000069905	*il-1rii* [Om]	Interleukin 1 Receptor Type 2	Cytokine receptor that binds interleukins. Innate immune response involved in inflammation (Morrison et al., 2012)	9 (9)	5.14
G22519	Chr26: 25,196,314 - 25,217,197	ENSSSAG00000100178	*saa* [Ss]	Serum Amyloid A	Encodes acute phase protein involved in antibacterial innate immune response in teleost fish (Krasnov et al., 2021; Buchmann, 2022)	19 (19)	3.58
		ENSSSAG00000069990	*saa5* [Ss]	Serum Amyloid A5			
G26153	Chr3: 87,153,987 - 87,157,516	ENSSSAG00000115686	*socs3a* [Ss]	Suppressor of Cytokine Signaling 3	Immunosuppression activity, upregulated during anti-inflammatory response in Atlantic salmon (Grayfer et al., 2018)	3 (3)	2.79
G26969	Chr4: 40,469,385 - 40,471,643	ENSSSAG00000000816	*sat1* [Ss]	Spermidine/Spermine N1-acetyltransferase	Encodes polyamine catabolic enzyme with ferroptotic function associated with inflammation response in salmonids (Clark et al., 2019)	9 (6)	1.85
G28437	Chr5: 55,981,529 - 55,983,197	ENSSSAG00000053028	*hamp* [Ss]	Hepcidin Antimicrobial Peptide	Encodes acute phase protein involved in antibacterial immune response in salmonids (e.g. Eslamloo et al., 2020)	4 (4)	4.38
G28633	Chr5: 62,584,203 - 62,742,348	ENSSSAG00000003833	*steap4* [Ss]	Six Transmembrane Epithelial Antigen of Prostate 4	Metalloreductase significantly upregulated in Atlantic salmon in response to bacterial infection (Krasnov et al., 2021). Metalloreductase gene family implicated in inflammatory response (Zhang et al., 2012)	20 (20)	2.93
		novelGene	N/A	Novel Gene in LR Transcriptome	Transcribed intronic region of steap4. Predicted to be non-coding by SQANTI3		
G29010	Chr5: 79,374,085 - 79,377,265	ENSSSAG00000092624	*gabbr2* [Mm]	Gamma-Aminobutyric Acid B Receptor 2	Subunit of GABA-binding protein. GABA-signalling related to autophagy in response to intracellular bacterial infection (Kim et al., 2018)	4 (3)	2.19

Gene ID	Locus	Associated Ensembl ID	Gene Name	Description	Function	# Transcripts (DETs)	Average Log2FC
G3902	Chr11: 24,866,738 - 24,867,462	novelGene	N/A	Novel Gene in LR Transcriptome	Transcribed intronic region of ENSSSAG00000014172, which is homologous to slc25a22b in zebrafish. Predicted to be non-coding by SQANTI3	2 (2)	6.54
G6022	Chr12: 61,404,740 - 61,409,522	ENSSSAG00000064977	*mmp19* [Dr]	Matrix Metallopeptidase 19	Member of the matrix metalloproteinase class of proteins which are involved in enzymatic degradation of extracellular matrix. Upregulated during inflammation in mammals (Manicone & McGuire, 2008)	1 (1)	4.26
G8628	Chr14: 31,538,023 - 31,539,906	ENSSSAG00000049319	*camp*	*Cathelicidin Antimicrobial Peptide*	Member of cathelicidin family which are involved in innate immunity in teleosts (Katzenback, 2015; Brunner et al., 2020)	19 (19)	4.56

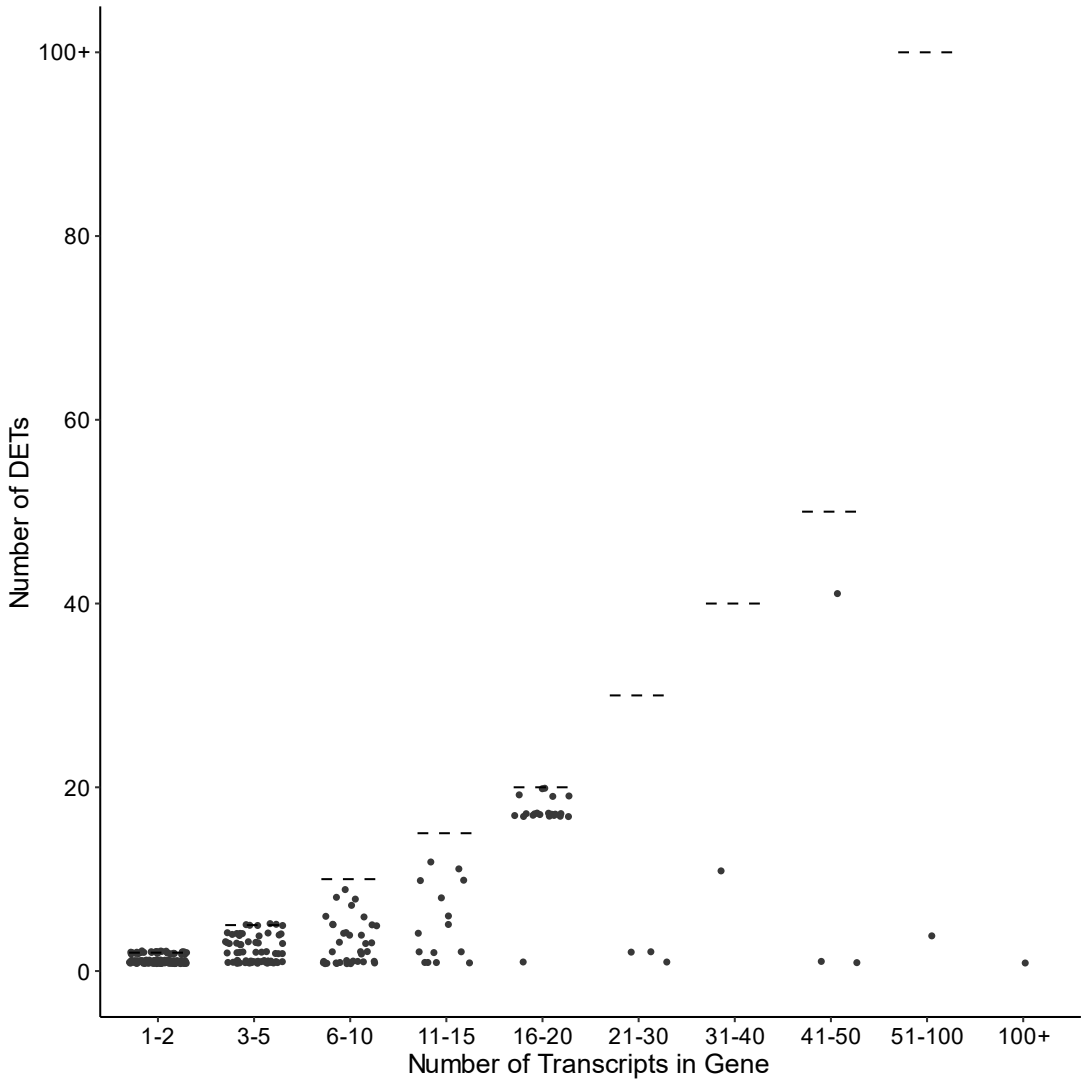


Figure 3.9: Dotplot showing the number of DETs versus the number of filtered transcripts (filtering retained only transcripts with ≥5 reads in at least 4 out of 6 biological replicates) per gene in the Vibrio group. Dashed lines indicate the maximum number of DETs for each bin. i.e. the line at which all transcripts were classified as DETs.

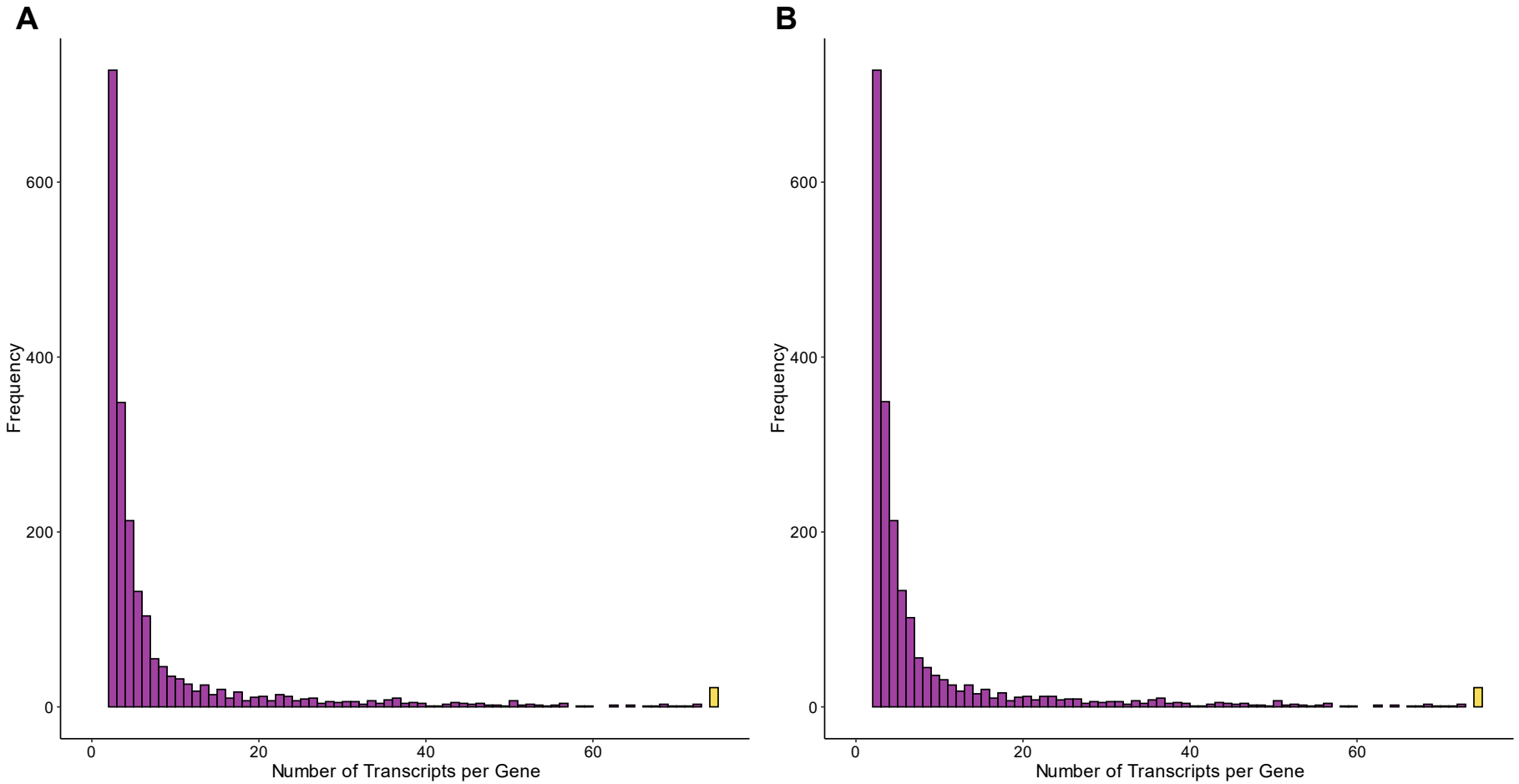


Figure 3.10: Histograms of the number of transcripts per gene input into the DTU analysis with DRIMSeq for (A) poly I:C and (B) Vibrio, versus controls. Yellow bar is a summation of all genes possessing 75 or more transcripts.

Table 3.4: DTU results for both treatment groups. DET column indicates whether each listed transcript was also identified as a DET

Gene ID	Transcript ID	Treatment Group	Gene p-value	Transcript p-value	DET?
G11138	G11138.1	poly I:C	3.51e-4	0.00e+0	Y
	G11138.2			0.00e+0	N
G13151	G13151.3	poly I:C	7.32e-10	1.85e-13	Y
		Vibrio	4.84e-3	4.54e-5	Y
G16725	G16725.12	poly I:C	2.46e-4	3.30e-11	N
G17946	G17946.1	poly I:C	4.06e-2	0.00e+0	N
	G17946.10			0.00e+0	N
G20128	G20128.26	poly I:C	3.96e-2	1.85e-8	N
G23108	G23108.5	poly I:C	4.65e-3	1.82e-2	N
G23563	G23563.2	poly I:C	4.36e-3	1.04e-9	N
G26010	G26010.61	poly I:C	6.23e-3	5.90e-4	N
G35151	G35151.17	poly I:C	6.12e-5	1.03e-12	N

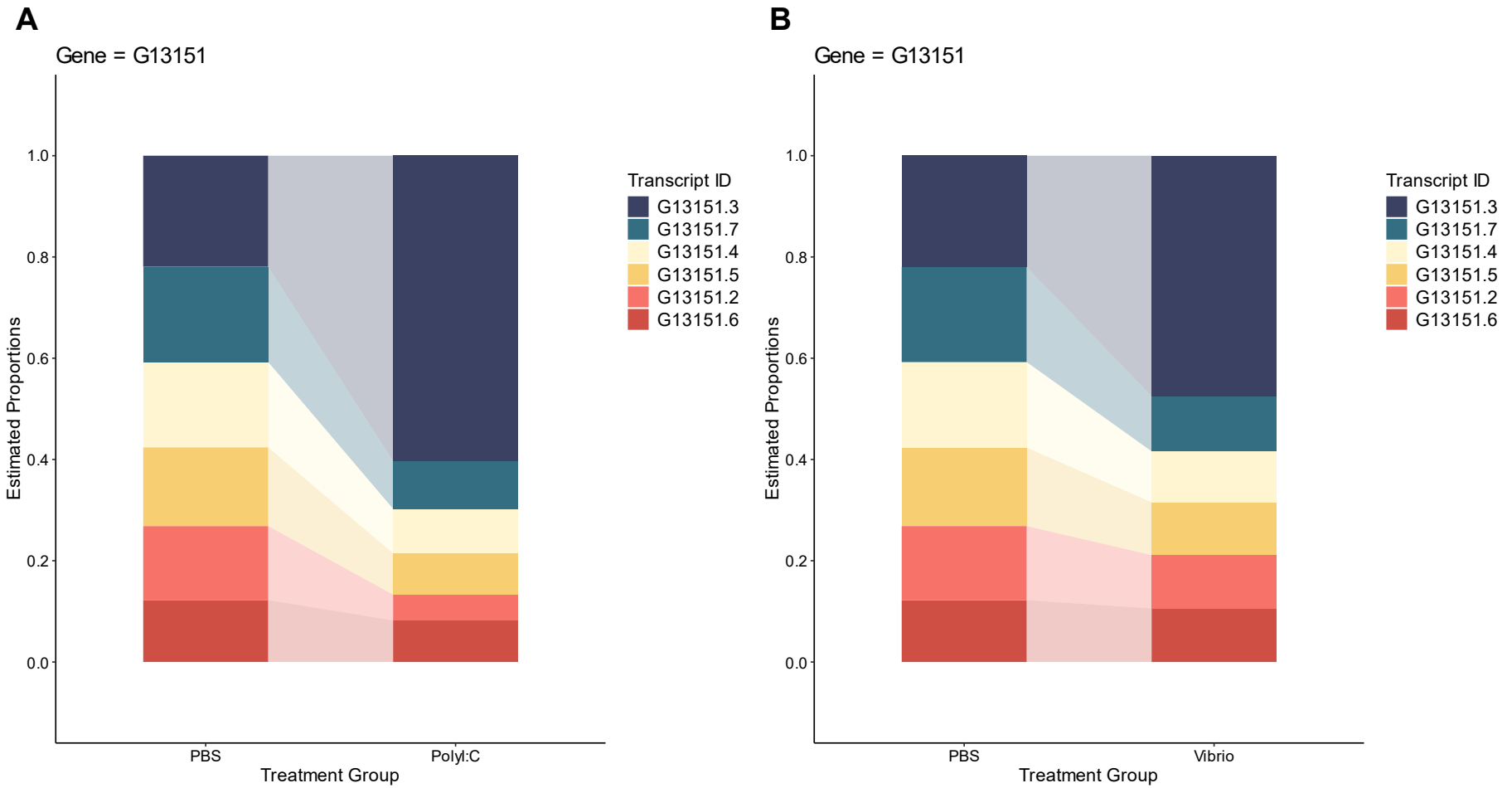


Figure 3.11: Ribbon plots showing changes in the proportion of transcripts expressed for gene G13151 between (A) control and poly I:C, and (B) control and Vibrio groups.

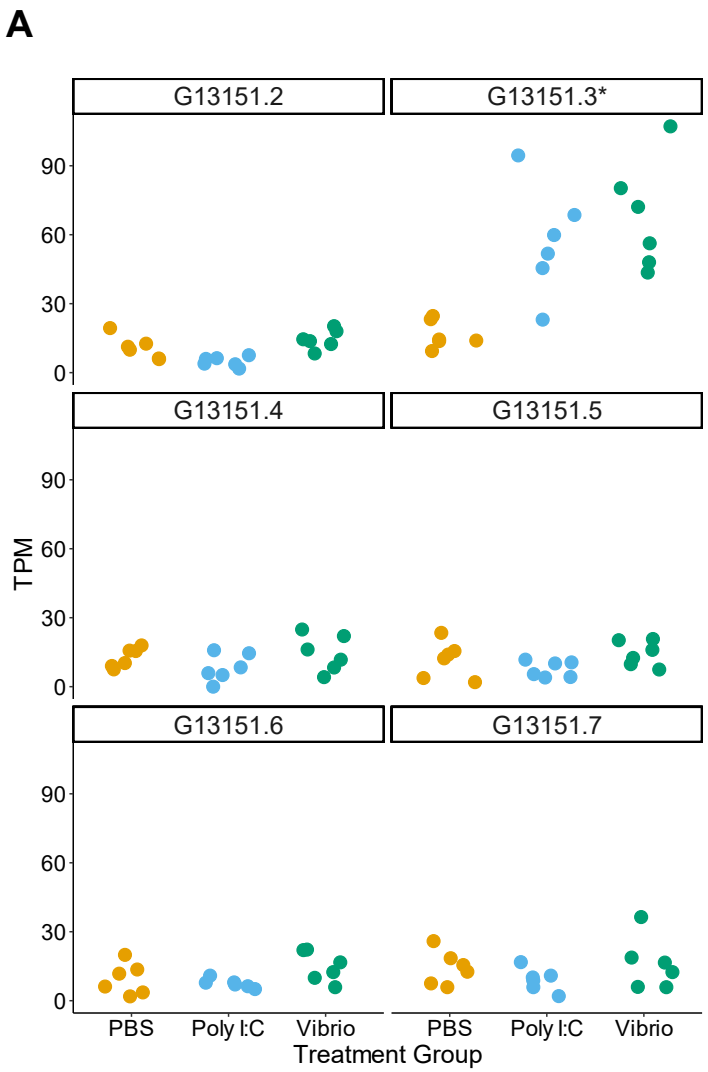


Figure 3.12: Visualisation of DETs (A) with matched transcript structures (B, see next page) for novel gene G13151.

*(A) Dotplots on left show TPMs (n=6) for all transcript models that passed expression filtering in the PBS, poly I:C and Vibrio groups. DETs are marked to highlight treatment groups; † = poly I:C only, ^ = Vibrio only, * = poly I:C and Vibrio.*

(B) Overleaf is a visualisation of transcript models for 1) Ensembl annotated transcripts up and downstream of G13151(dark red), 2) RefSeq annotation of transcripts in the same region, 3) DETs for G13151, coloured according to the treatment group for which they were differentially expressed, and 4) promoter regions from the Ensembl regulatory build (bright red). UTRs for Ensembl and RefSeq transcripts are displayed in white.

B

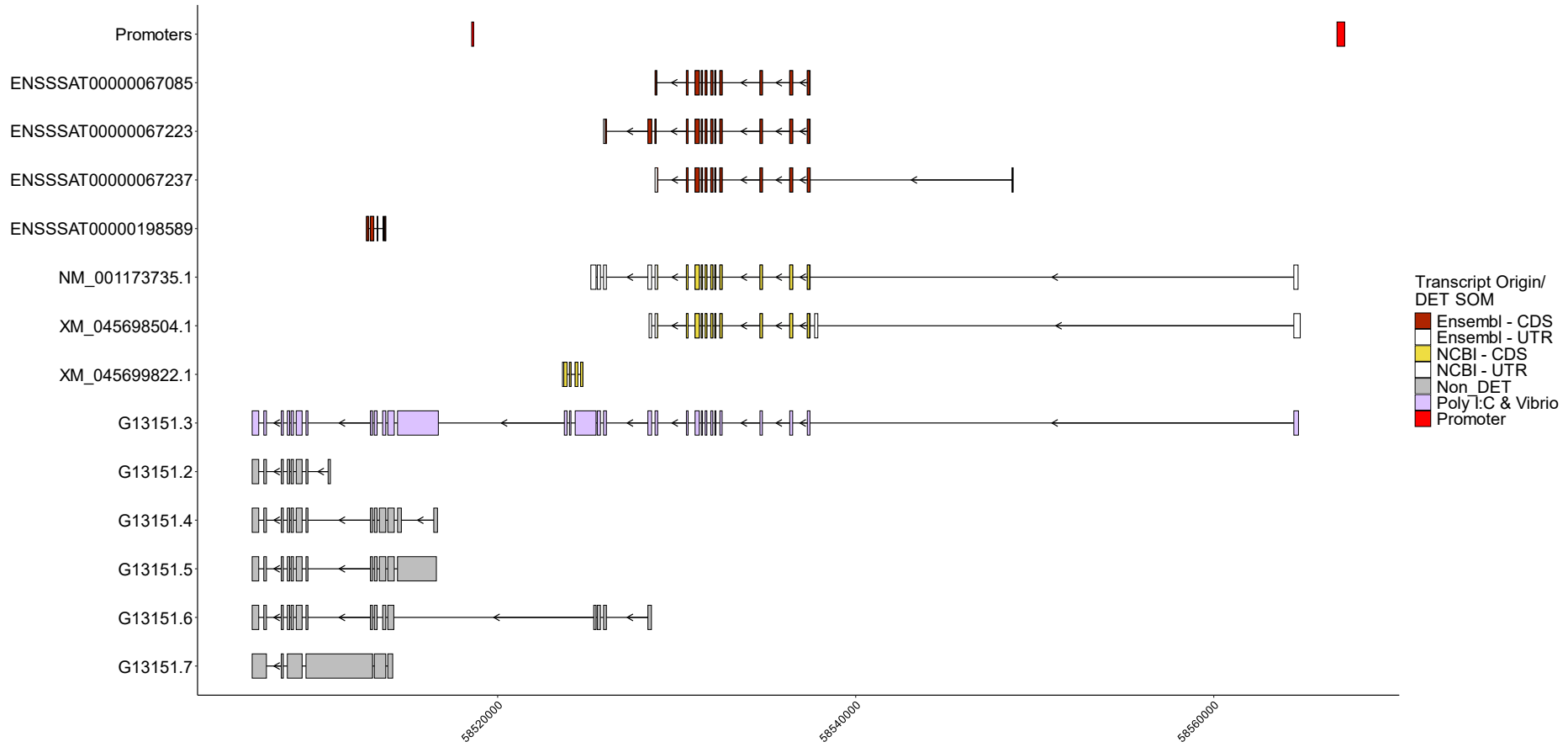


Figure 3.12 (continued): Visualisation of DETs (A) with matched transcript structures (B) for gene G13151.

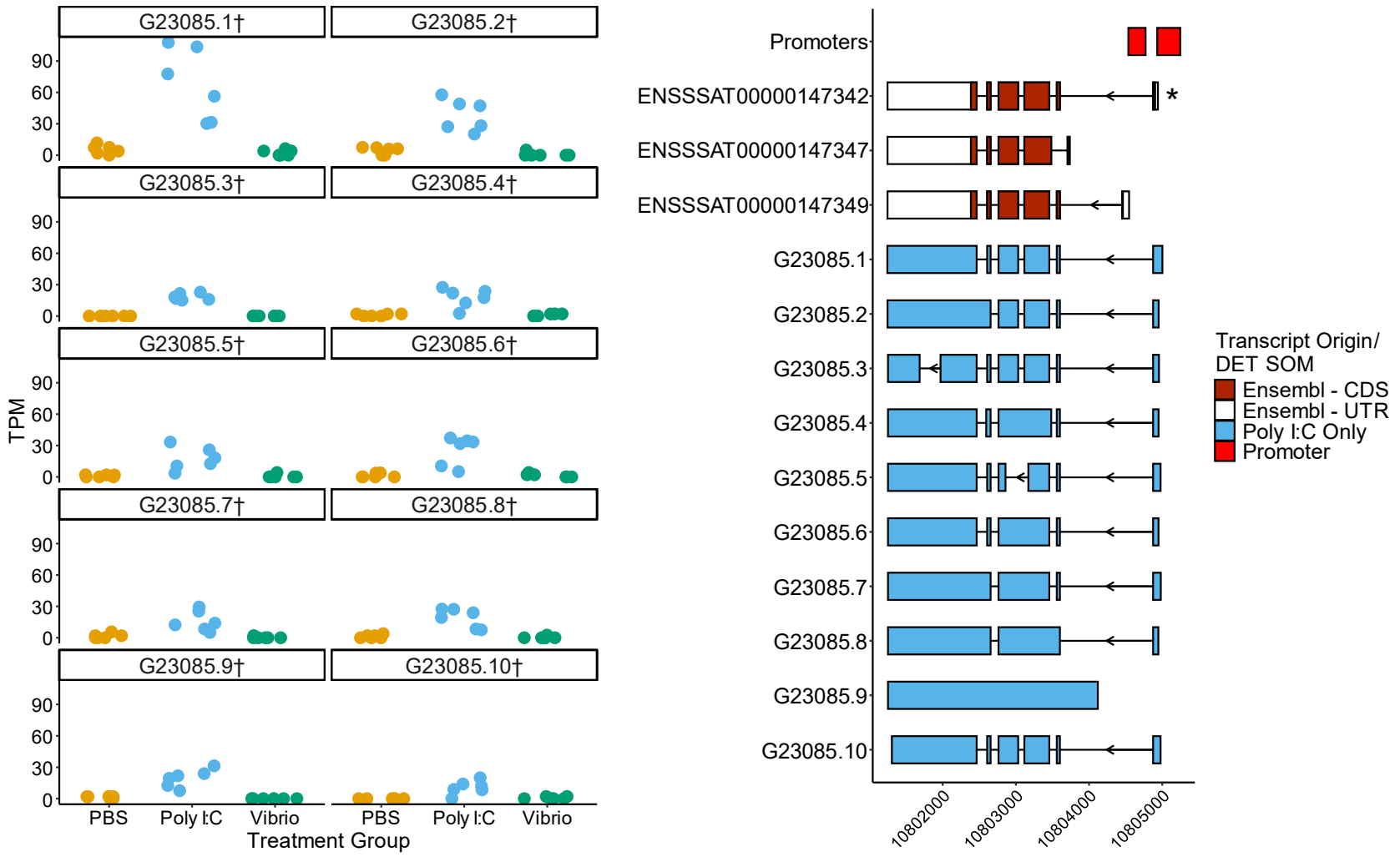


Figure 3.13: Visualisation of DETs (left) with matched transcript structures (right) for gene G23085 – lgals17…(Legend continued on next page)

(Legend continued) Visualisation of DETs (left) with matched transcript structures (right) for gene G23085 – lgals17.

*Dotplots on left show TPM values (n=6) for all transcript models that passed expression filtering in the three treatment groups; PBS, poly I:C and Vibrio. DETs are marked to highlight treatment groups where differentially expressed; † = poly I:C only, ^ = Vibrio only, * = poly I:C and Vibrio. On the right is a visualisation of transcript models for 1) the Ensembl annotation (dark red), 2) DETs for G23085, coloured according to treatment group, and 3) promoter regions from the Ensembl Ssal_v3.1 regulatory build (bright red). UTRs for Ensembl transcripts are displayed in white, while the canonical transcript is indicated by an asterisk.*

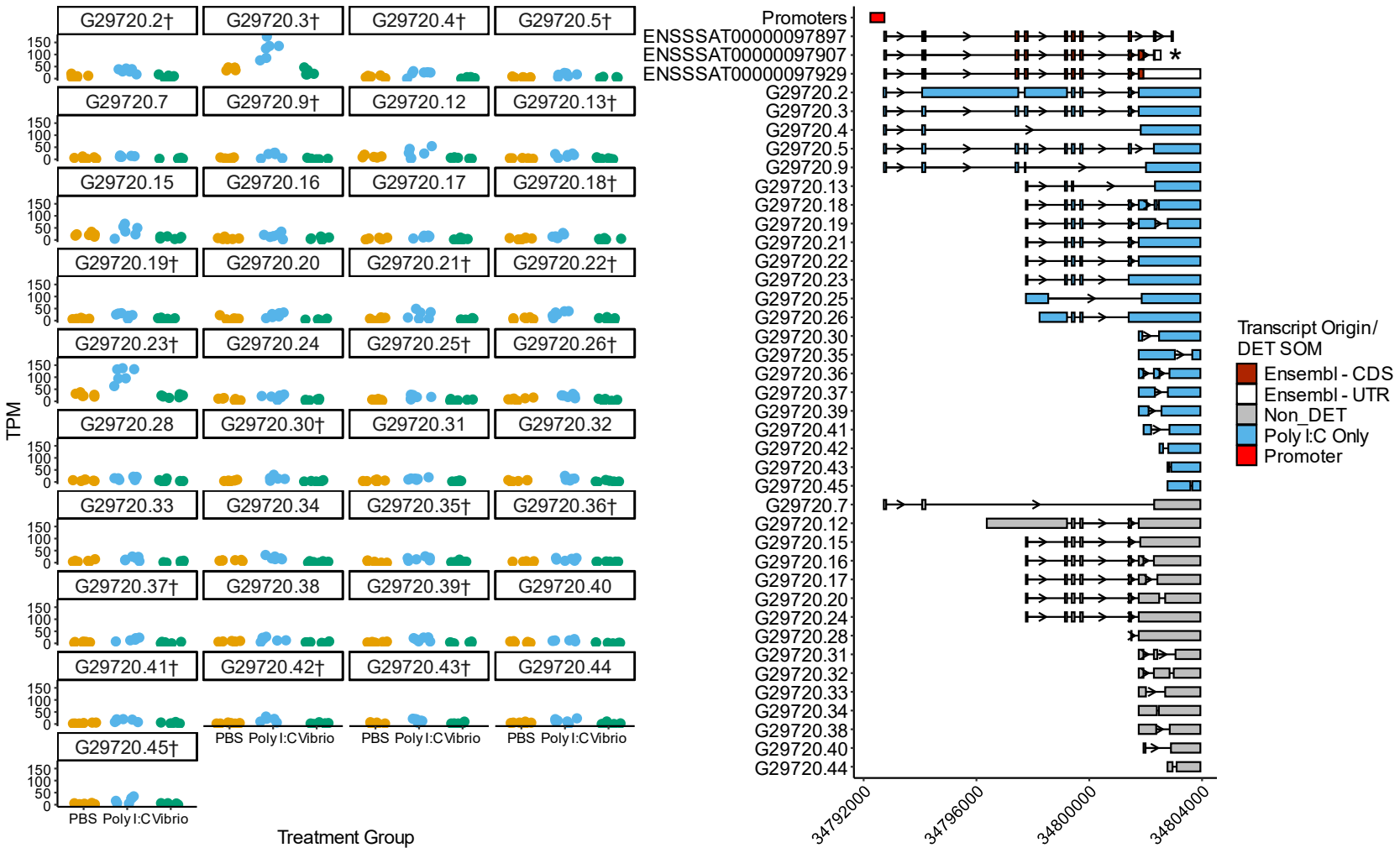


Figure 3.14: Visualisation of DETs (left) with matched transcript structures (right) for gene G29720 – cd9…(Legend continued on next page)

(Legend continued) Visualisation of DETs (left) with matched transcript structures (right) for gene G29720 – cd9.

*Dotplots on left show TPM values (n=6) for all transcript models that passed expression filtering in the three treatment groups; PBS, poly I:C and Vibrio. DETs are marked to highlight treatment groups where differentially expressed; † = poly I:C only, ^ = Vibrio only, * = poly I:C and Vibrio. On the right is a visualisation of transcript models for 1) the Ensembl annotation (dark red), 2) DETs for G29720, coloured according to treatment group, and 3) promoter regions from the Ssal_v3.1 regulatory build (bright red). UTRs for Ensembl transcripts are displayed in white, while the canonical transcript is indicated by an asterisk.*

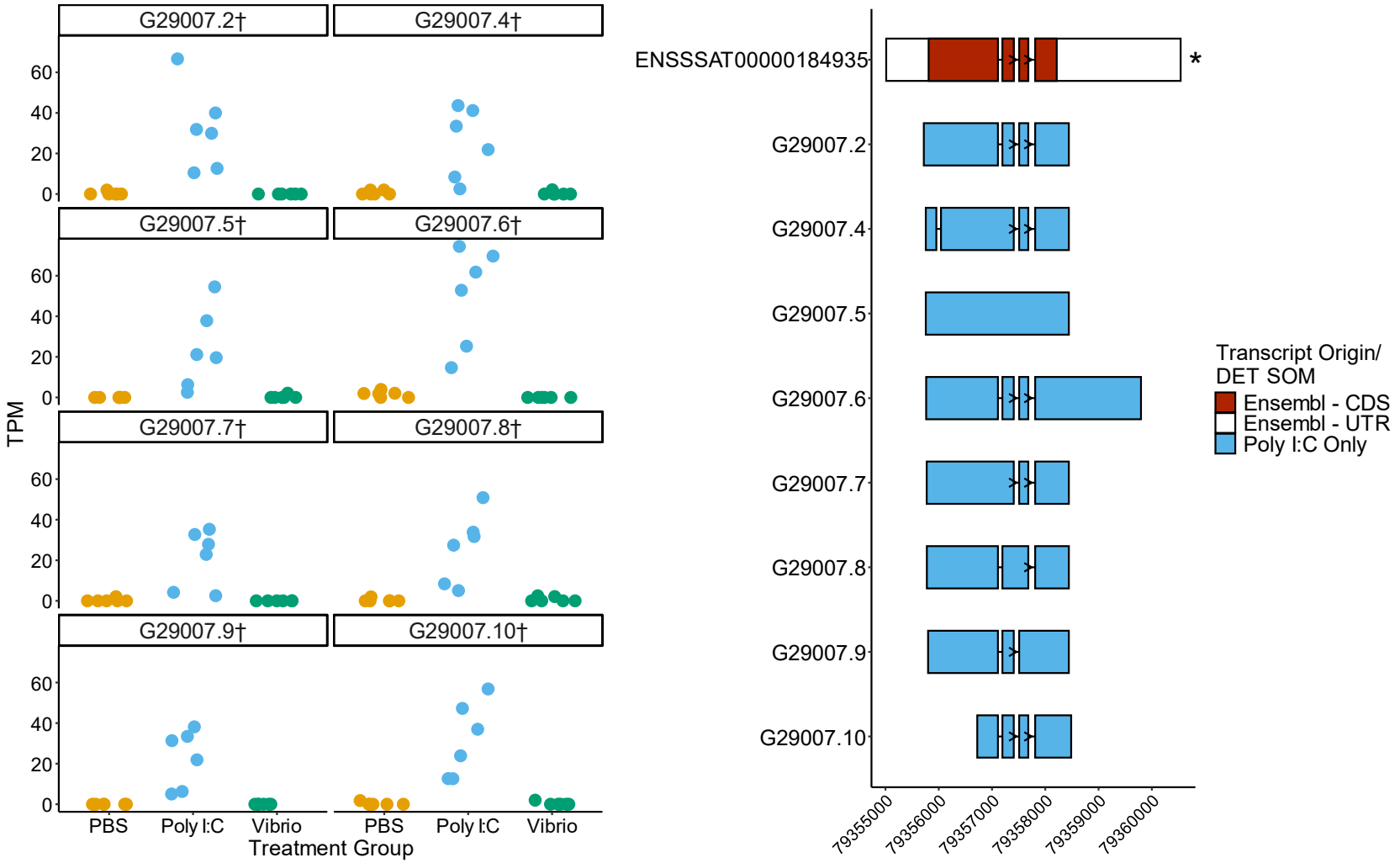


Figure 3.15: Visualisation of DETs (left) with matched transcript structures (right) for gene G29007 – aste1…(Legend continued on next page)

(Legend continued) Visualisation of DETs (left) with matched transcript structures (right) for gene G29007 – aste1.

*Dotplots on left show TPM values (n=6) for all transcript models that passed expression filtering in the three treatment groups; PBS, poly I:C and Vibrio. DETs are marked to highlight treatment groups where differentially expressed; † = poly I:C only, ^ = Vibrio only, * = poly I:C and Vibrio. On the right is a visualisation of transcript models for 1) the Ensembl annotation (dark red), and 2) DETs for G29007, coloured according to treatment group. UTRs for Ensembl transcripts are displayed in white, while the canonical transcript is indicated by an asterisk.*

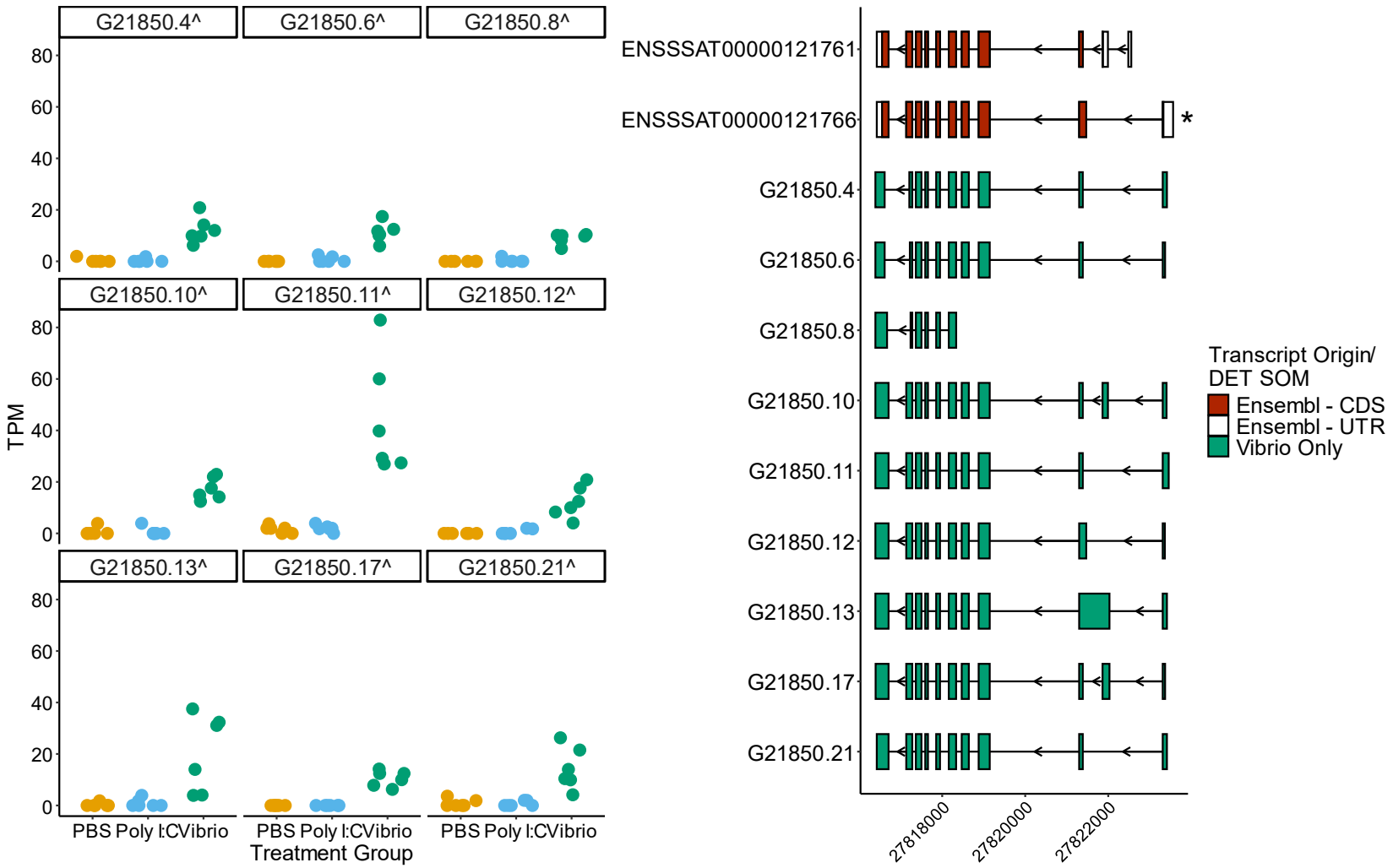


Figure 3.16: Visualisation of DETs (left) with matched transcript structures (right) for gene G21850 – il-rii …(Legend continued on next page)

(Legend continued) Visualisation of DETs (left) with matched transcript structures (right) for gene G21850 – il-rii.

*Dotplots on left show TPM values (n=6) for all transcript models that passed expression filtering in the three treatment groups; PBS, poly I:C and Vibrio. DETs are marked to highlight treatment groups where differentially expressed; † = poly I:C only, ^ = Vibrio only, * = poly I:C and Vibrio. On the right is a visualisation of transcript models for 1) the Ensembl annotation (dark red), and 2) DETs for G21850, coloured according to treatment group. UTRs for Ensembl transcripts are displayed in white, while the canonical transcript is indicated by an asterisk.*

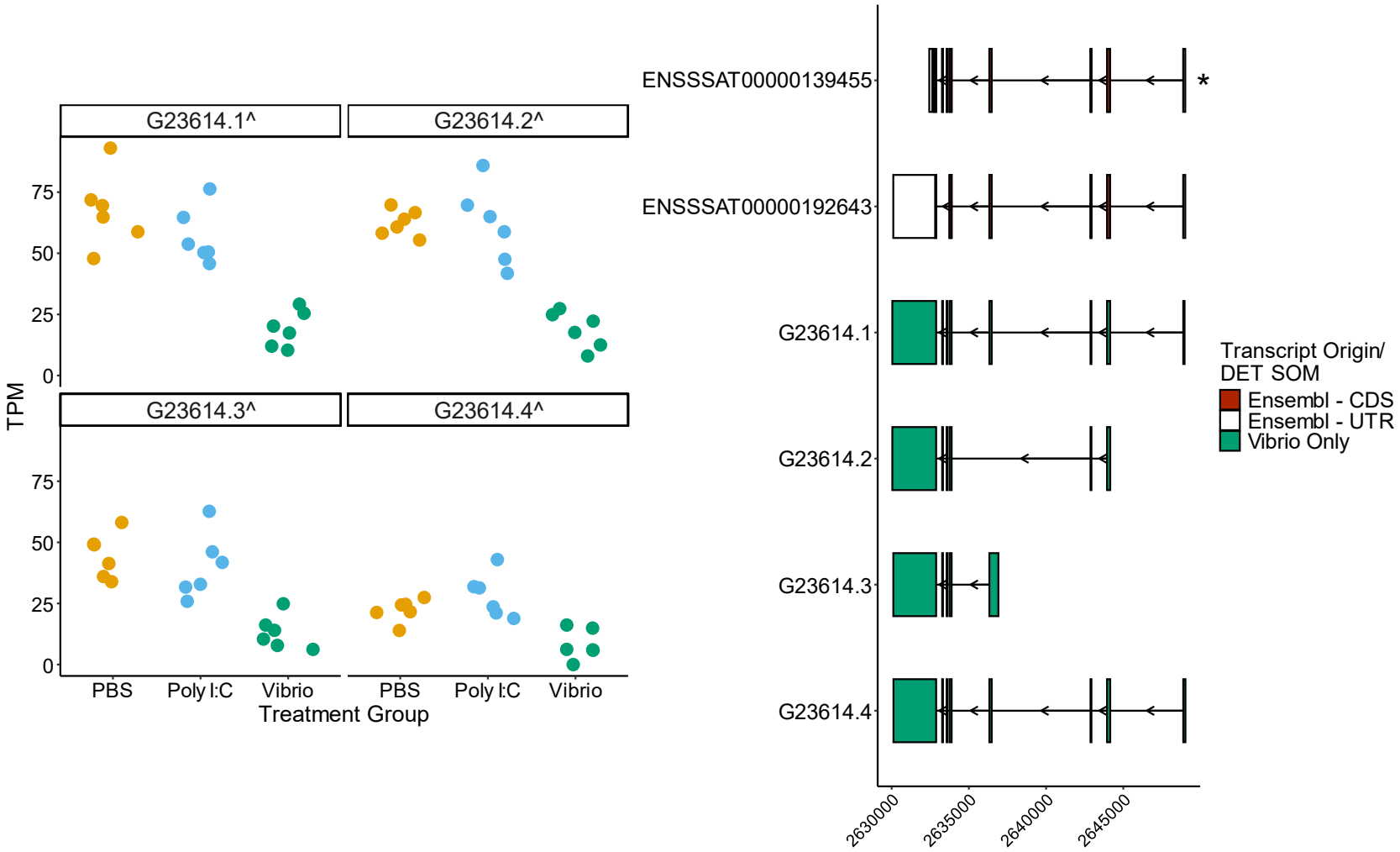


Figure 3.17: Visualisation of DETs (left) with matched transcript structures (right) for gene G23614 – tmem106a …(Legend continued overleaf)

(Legend continued) Visualisation of DETs (left) with matched transcript structures (right) for gene G23614 – tmem106a.

*Dotplots on left show TPM values (n=6) for all transcript models that passed expression filtering in the three treatment groups; PBS, poly I:C and Vibrio. DETs are marked to highlight treatment groups where differentially expressed; † = poly I:C only, ^ = Vibrio only, * = poly I:C and Vibrio. On the right is a visualisation of transcript models for 1) the Ensembl annotation (dark red), and 2) DETs for G23614, coloured according to treatment group. UTRs for Ensembl transcripts are displayed in white, while the canonical transcript is indicated by an asterisk.*

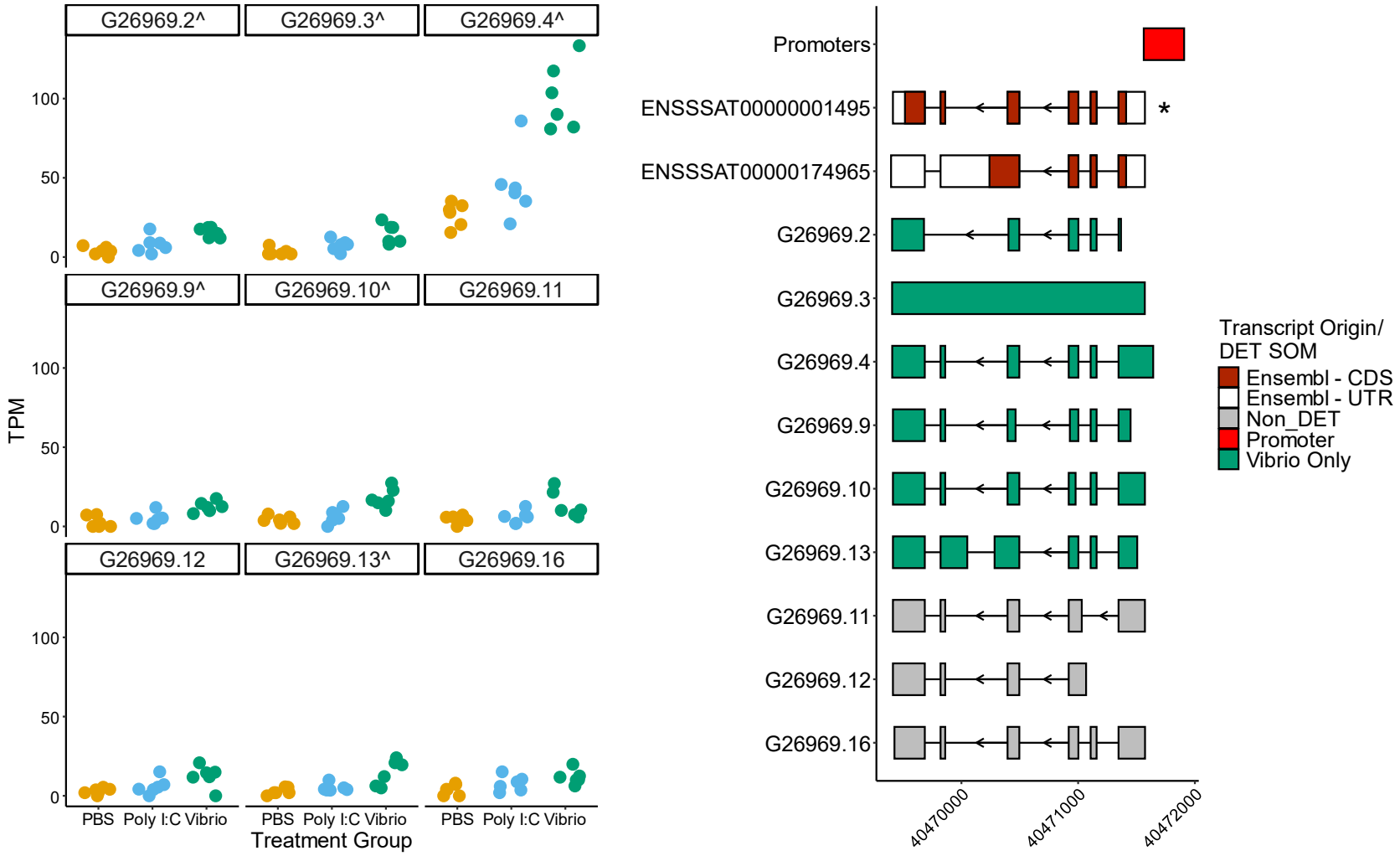


*Figure 3.18: Visualisation of DETs (left) with matched transcript structures (right) for gene G26969 – sat1…**(Legend continued on next page)*

(Legend continued) Visualisation of DETs (left) with matched transcript structures (right) for gene G26969 – sat1.

*Dotplots on left show TPM values (n=6) for all transcript models that passed expression filtering in the three treatment groups; PBS, poly I:C and Vibrio. DETs are marked to highlight treatment groups where differentially expressed; † = poly I:C only, ^ = Vibrio only, * = poly I:C and Vibrio. On the right is a visualisation of transcript models for 1) the Ensembl annotation (dark red), 2) DETs for G26969, coloured according to treatment group, and 3) promoter regions from the Ssal_v3.1 regulatory build (bright red). UTRs for Ensembl transcripts are displayed in white, while the canonical transcript is indicated by an asterisk.*

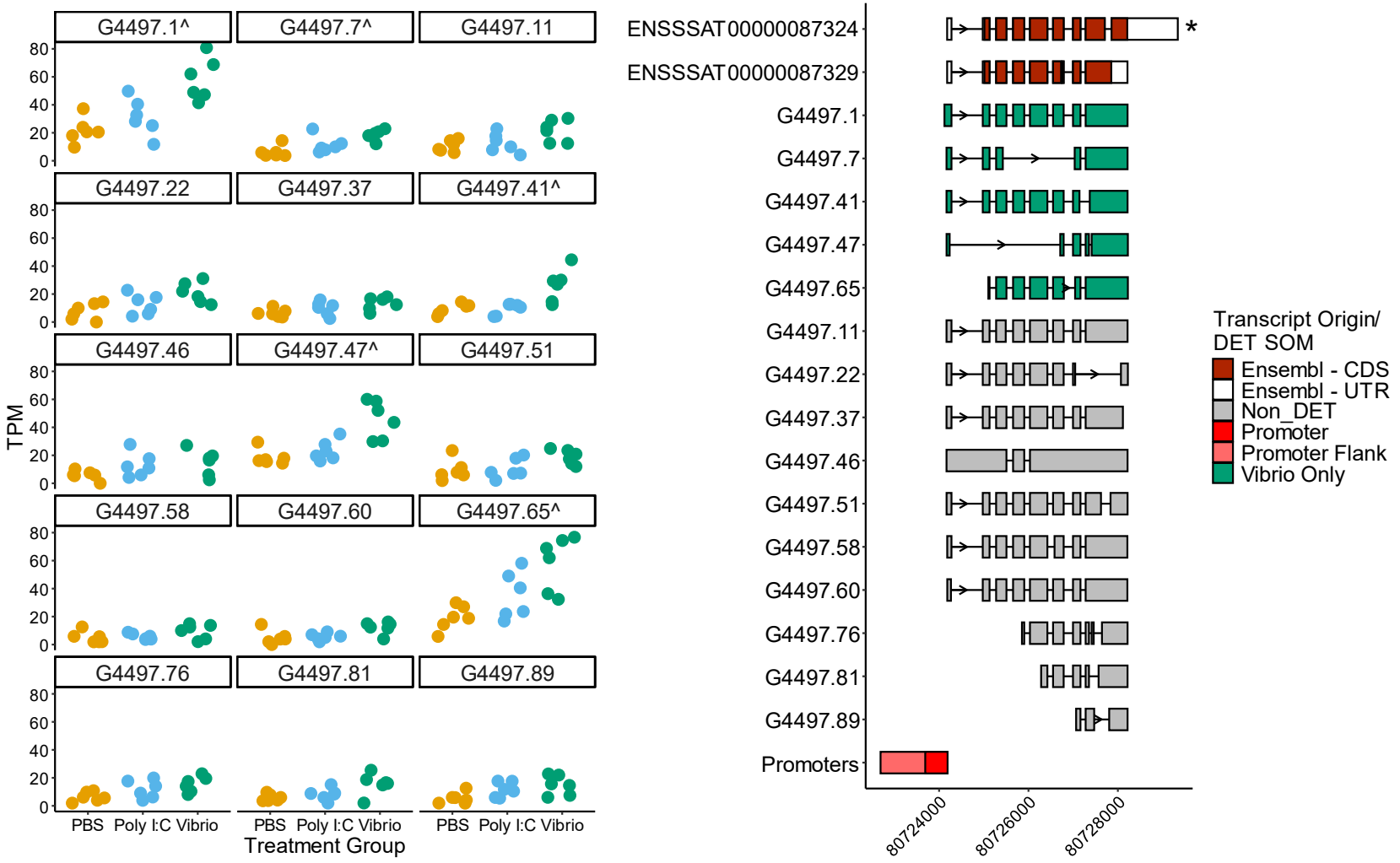


Figure 3.19: Visualisation of DETs (left) with matched transcript structures (right) for gene G4497 – hspa5…(Legend continued on next page)

(Legend continued) Visualisation of DETs (left) with matched transcript structures (right) for gene G4497 – hspa5.

*Dotplots on left show TPM values (n=6) for all transcript models that passed expression filtering in the three treatment groups; PBS, poly I:C and Vibrio. DETs are marked to highlight treatment groups where differentially expressed; † = poly I:C only, ^ = Vibrio only, * = poly I:C and Vibrio. On the right is a visualisation of transcript models for 1) the Ensembl annotation (dark red), 2) DETs for G4497, coloured according to treatment group, and 3) promoter regions from the Ssal_v3.1 regulatory build (bright red). UTRs for Ensembl transcripts are displayed in white, while the canonical transcript is indicated by an asterisk.*

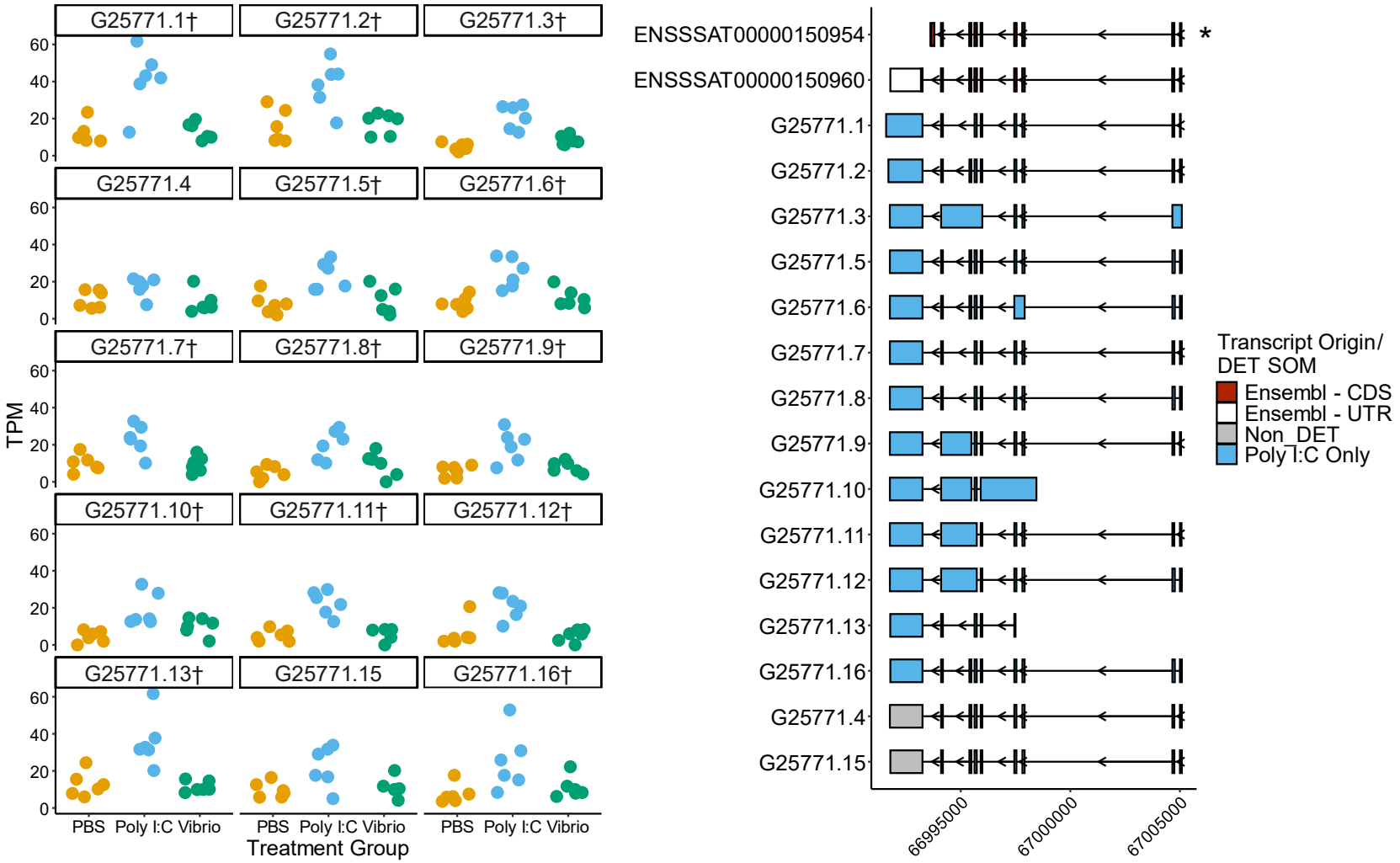


Figure 3.20: Visualisation of DETs (left) with matched transcript structures (right) for gene G25771 – cd9…(Legend continued on next page)

(Legend continued) Visualisation of DETs (left) with matched transcript structures (right) for gene G25771 – cd9.

*Dotplots on left show TPM values (n=6) for all transcript models that passed expression filtering in the three treatment groups; PBS, poly I:C and Vibrio. DETs are marked to highlight treatment groups where differentially expressed; † = poly I:C only, ^ = Vibrio only, * = poly I:C and Vibrio. On the right is a visualisation of transcript models for 1) the Ensembl annotation (dark red), and 2) DETs for G25771, coloured according to treatment group. UTRs for Ensembl transcripts are displayed in white, while the canonical transcript is indicated by an asterisk.*

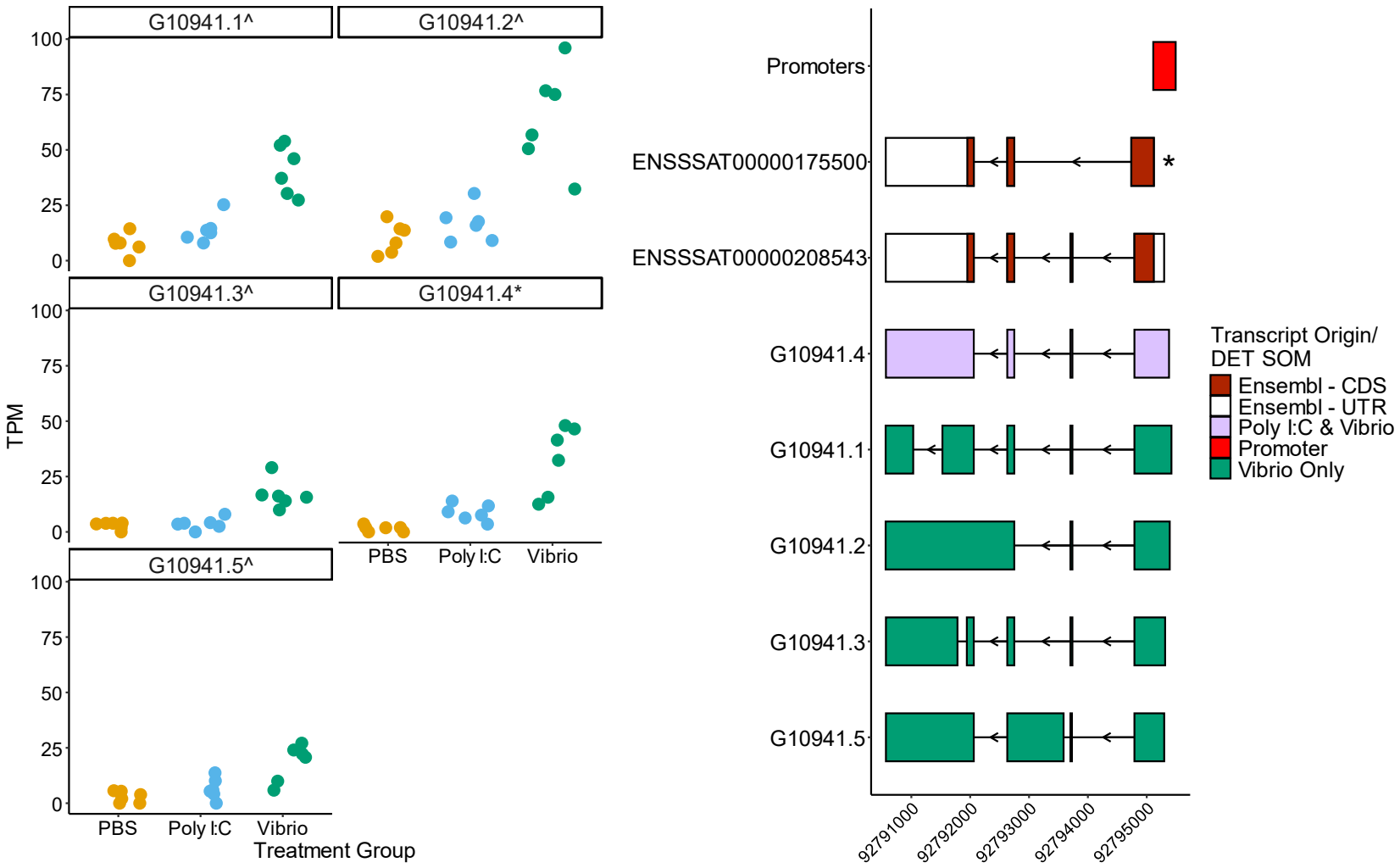


Figure 3.21: Visualisation of DETs (left) with matched transcript structures (right) for gene G10941 – igfbp6…(Legend continued on next page)

(Legend continued) Visualisation of DETs (left) with matched transcript structures (right) for gene G10941 – igfbp6.

*Dotplots on left show TPM values (n=6) for all transcript models that passed expression filtering in the three treatment groups; PBS, poly I:C and Vibrio. DETs are marked to highlight treatment groups where differentially expressed; † = poly I:C only, ^ = Vibrio only, * = poly I:C and Vibrio. On the right is a visualisation of transcript models for 1) the Ensembl annotation (dark red), 2) DETs for G10941, coloured according to treatment group, and 3) promoter regions from the Ssal_v3.1 regulatory build (bright red). UTRs for Ensembl transcripts are displayed in white, while the canonical transcript is indicated by an asterisk.*

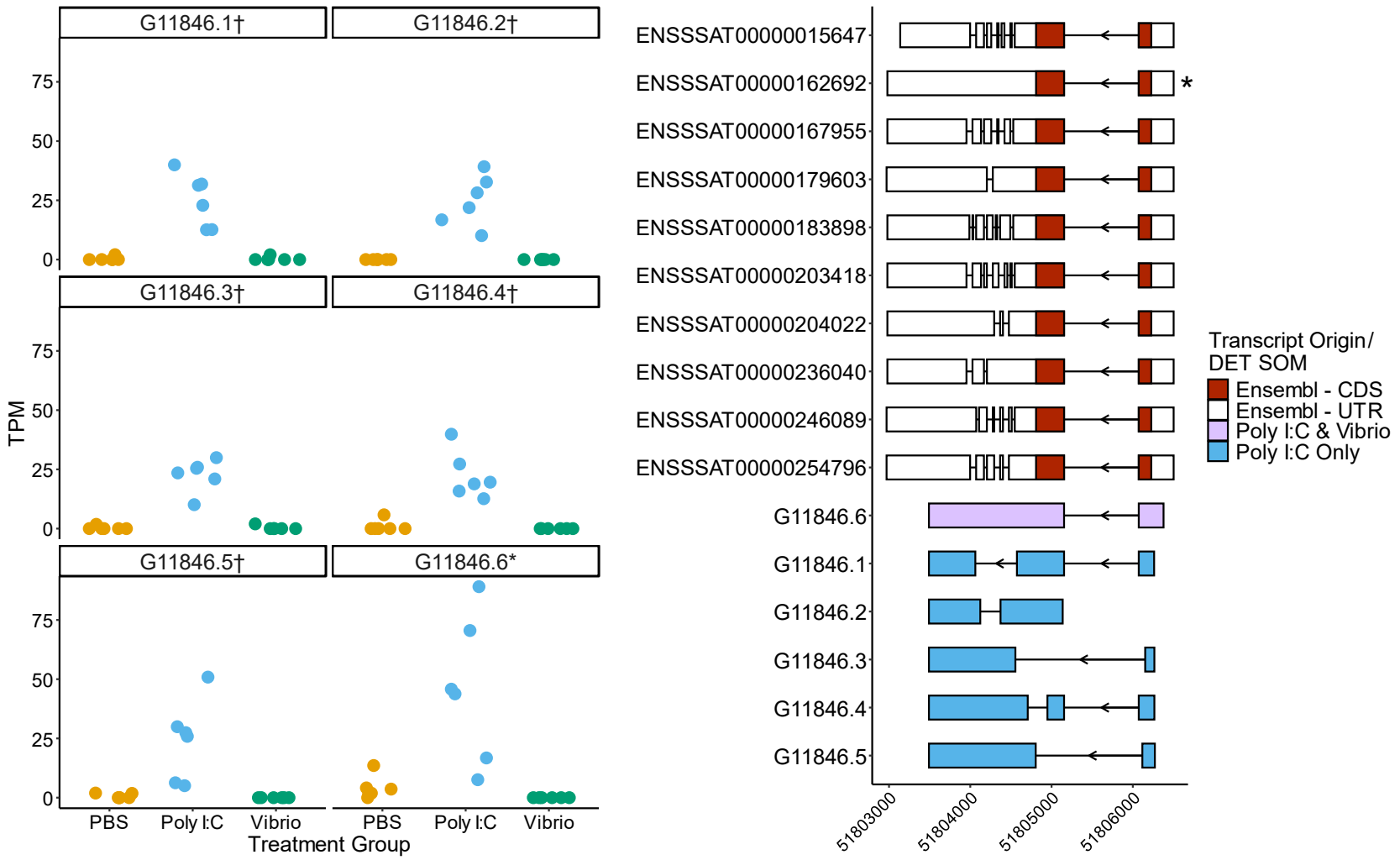


*Figure 3.22: Visualisation of DETs (left) with matched transcript structures (right) for gene G11846 – rtp2…**(Legend continued on next page)*

(Legend continued) Visualisation of DETs (left) with matched transcript structures (right) for gene G11846 – rtp2.

*Dotplots on left show TPM values (n=6) for all transcript models that passed expression filtering in the three treatment groups; PBS, poly I:C and Vibrio. DETs are marked to highlight treatment groups where differentially expressed; † = poly I:C only, ^ = Vibrio only, * = poly I:C and Vibrio. On the right is a visualisation of transcript models for 1) the Ensembl annotation (dark red), and 2) DETs for G11846, coloured according to treatment group. UTRs for Ensembl transcripts are displayed in white, while the canonical transcript is indicated by an asterisk.*

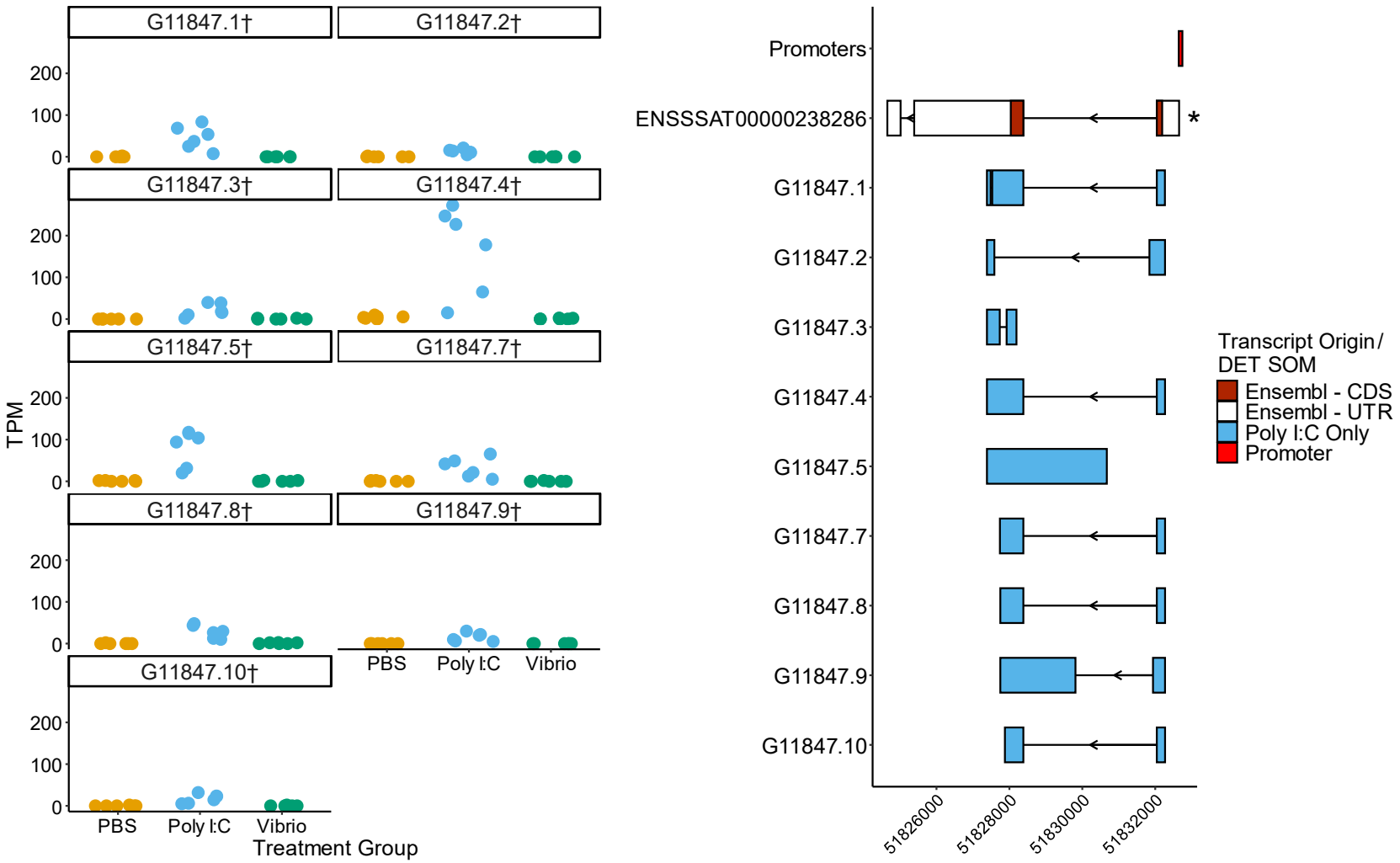


Figure 3.23: Visualisation of DETs (left) with matched transcript structures (right) for gene G11847 – rtp2…(Legend continued on next page)

(Legend continued) Visualisation of DETs (left) with matched transcript structures (right) for gene G11847 – rtp2.

*Dotplots on left show TPM values (n=6) for all transcript models that passed expression filtering in the three treatment groups; PBS, poly I:C and Vibrio. DETs are marked to highlight treatment groups where differentially expressed; † = poly I:C only, ^ = Vibrio only, * = poly I:C and Vibrio. On the right is a visualisation of transcript models for 1) the Ensembl annotation (dark red), 2) DETs for G11847, coloured according to treatment group, and 3) promoter regions from the Ssal_v3.1 regulatory build (bright red). UTRs for Ensembl transcripts are displayed in white, while the canonical transcript is indicated by an asterisk.*

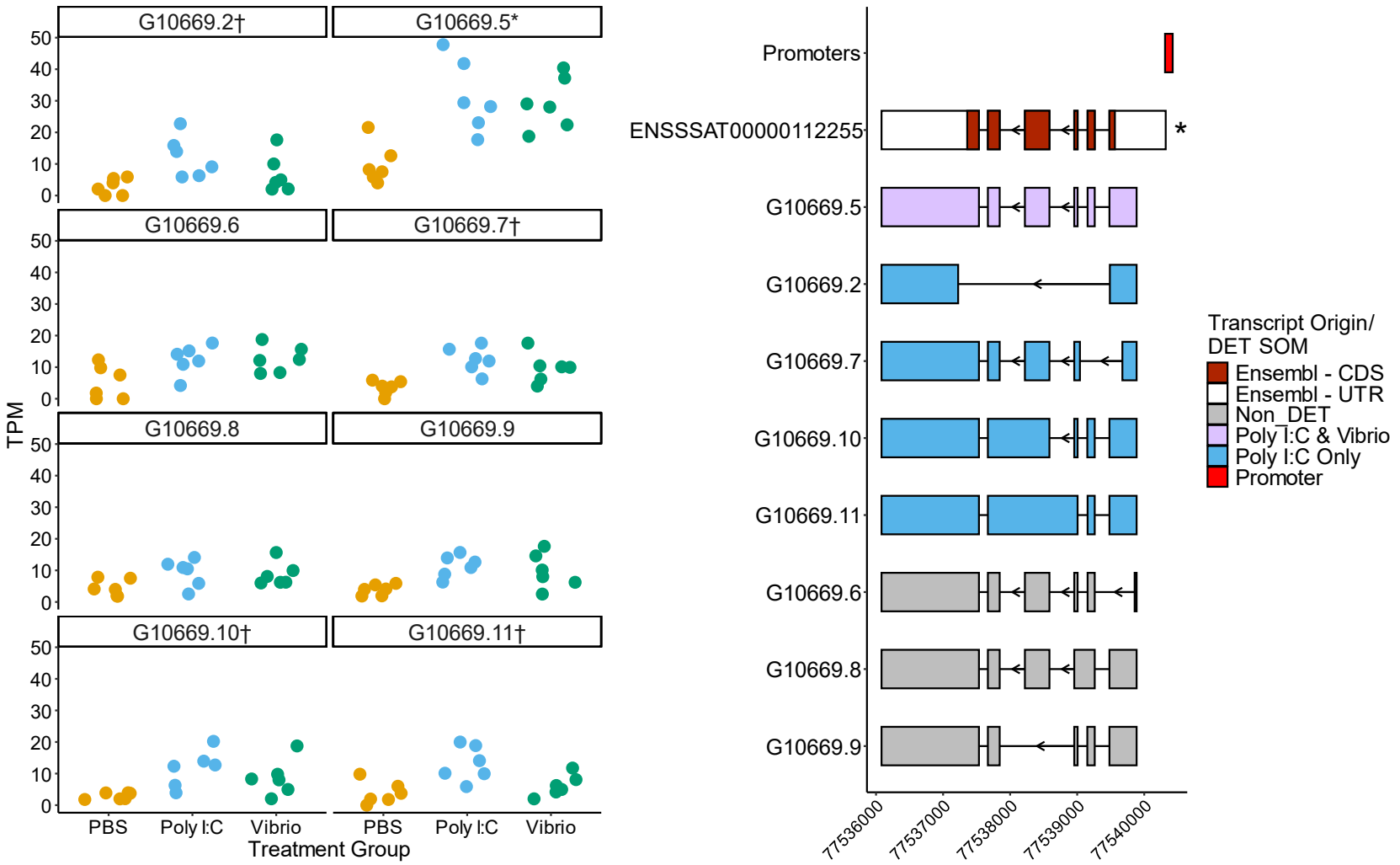


Figure 3.24: Visualisation of DETs (left) with matched transcript structures (right) for gene G10669 – pim1…(Legend continued on next page)

(Legend continued) Visualisation of DETs (left) with matched transcript structures (right) for gene G10669 – pim1.

*Dotplots on left show TPM values (n=6) for all transcript models that passed expression filtering in the three treatment groups; PBS, poly I:C and Vibrio. DETs are marked to highlight treatment groups where differentially expressed; † = poly I:C only, ^ = Vibrio only, * = poly I:C and Vibrio. On the right is a visualisation of transcript models for 1) the Ensembl annotation (dark red), 2) DETs for G10669, coloured according to treatment group, and 3) promoter regions from the Ssal_v3.1 regulatory build (bright red). UTRs for Ensembl transcripts are displayed in white, while the canonical transcript is indicated by an asterisk.*

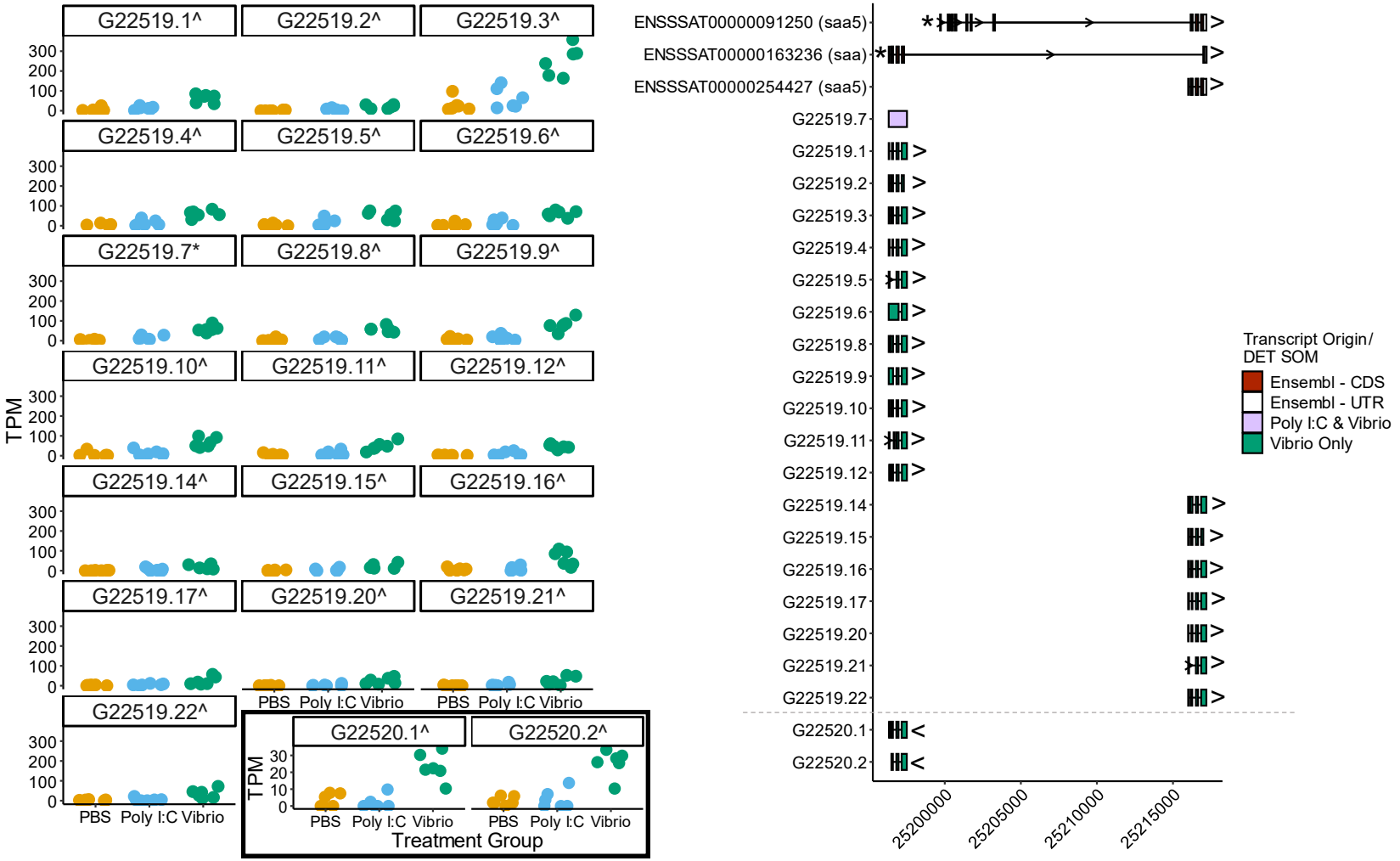


Figure 3.25: Visualisation of DETs (left) with matched transcript structures (right) for gene G25519 & G25520…(Legend continued overleaf)

(Legend continued) Visualisation of DETs (left) with matched transcript structures (right) for gene G25519 – saa & G25520 - novel gene antisense to saa/G25519.

*Dotplots on left show TPM values (n=6) for all transcript models that passed expression filtering in the three treatment groups; PBS, poly I:C and Vibrio. Dotplots inlayed box is for G25520. DETs are marked to highlight treatment groups where differentially expressed; † = poly I:C only, ^ = Vibrio only, * = poly I:C and Vibrio. On the right is a visualisation of transcript models for 1) the Ensembl annotation for saa and saa5 (dark red), and 2) DETs for G25519 and G25520, coloured according to treatment group. UTRs for Ensembl transcripts are displayed in white, while the canonical transcript for both Ensembl reference models is indicated by an asterisk. Dashed line indicates change between forward and reverse strand.*

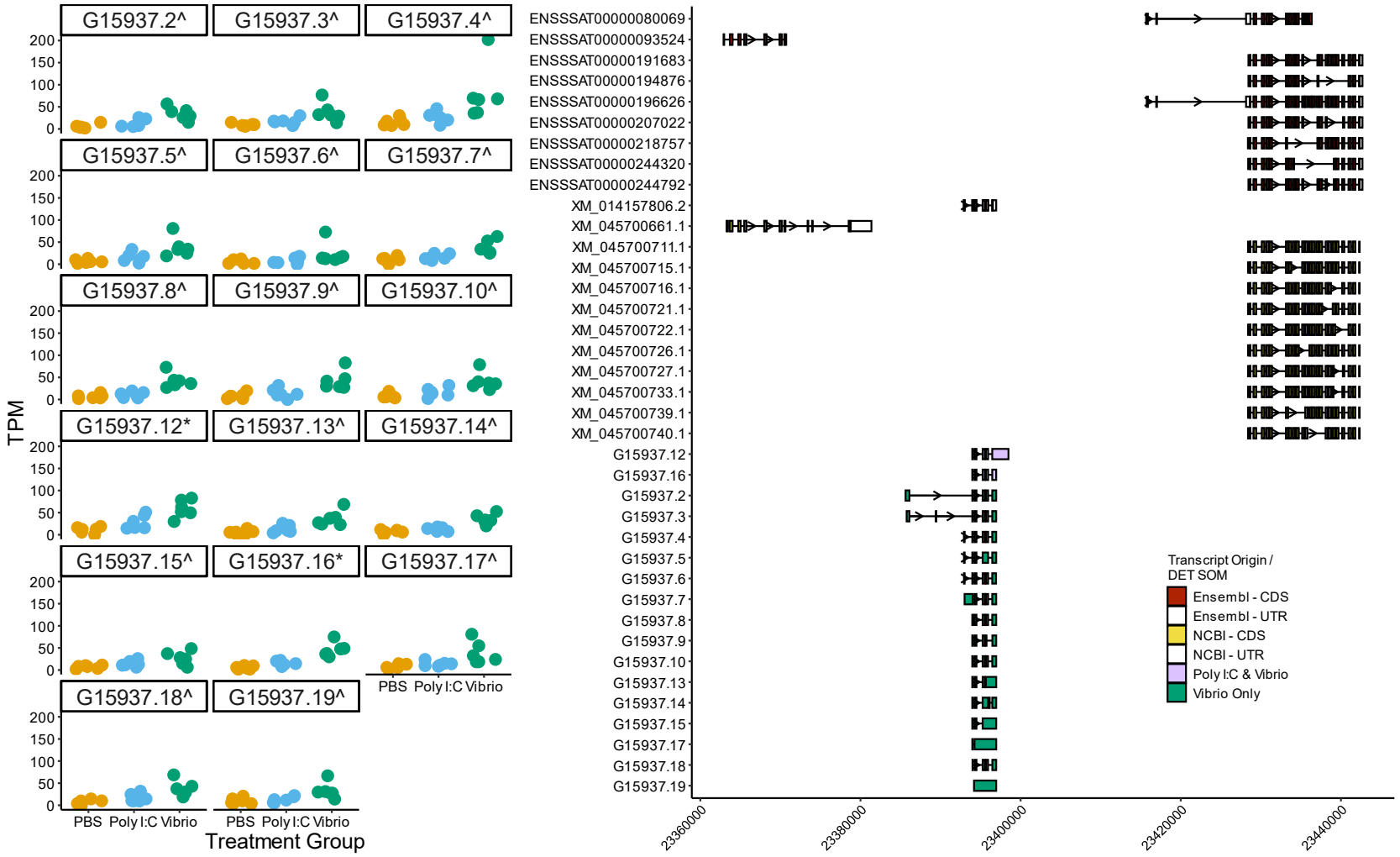


Figure 3.26: Visualisation of DETs (left) with matched transcript structures (right) for novel gene G15937…(Legend continued on next page)

(Legend continued) Visualisation of DETs (left) with matched transcript structures (right) for novel gene G15937.

*Dotplots on left show TPM values (n=6) for all transcript models that passed expression filtering in the three treatment groups; PBS, poly I:C and Vibrio. DETs are marked to highlight treatment groups where differentially expressed; † = poly I:C only, ^ = Vibrio only, * = poly I:C and Vibrio. On the right is a visualisation of transcript models for 1) the Ensembl transcripts located up and downstream of novel gene G15937 (dark red), 2) the NCBI RefSeq transcripts located in the same region, and 3) DETs for G15937, coloured according to treatment group. UTRs for Ensembl and RefSeq transcripts are displayed in white.*

Chapter 4: Transcript Resolved Expression and Alternative Usage During Atlantic Salmon Embryogenesis

Summary

This chapter develops a novel approach to profile transcript expression and alternative transcript usage across key stages of embryonic development in Atlantic salmon using ONT full-length RNA sequencing. Embryonic ONT reads were mapped to the transcriptome generated in Chapter 2, then quantified using a custom pipeline. Differential transcript expression was explored and a self-organizing map (SOM) approach adapted to cluster transcript co-expression through development. SOM clustering was adapted to identify genes showing alternative transcript usage at distinct developmental stages. This chapter advances our understanding of salmonid embryogenesis, elucidating the role of transcriptome diversity, while providing useful strategies to navigate DTE in complex timecourse datasets.

4.1 Introduction

Embryogenesis is an important stage of ontogeny where the cells and tissues defining the adult body plan are developed (Zhang et al., 2019). The regulation of gene expression is central to embryonic development in all species. However, much of our knowledge derives from studies of model species like mouse and zebrafish. Consequently, our understanding of gene expression during embryogenesis in non-model fish is limited. Short-read RNA-seq has been used extensively to study zebrafish embryogenesis, for example to profile transcriptome dynamics during the maternal-zygotic transition (Aanes et al., 2011) and describe lncRNA expression and impacts on development (Pauli et al., 2012).

Transcriptome diversity has a reported role in embryogenesis, including alternative splicing as mice and zebrafish embryos develop (Revil et al., 2010; Liu et al., 2022c). Long-read RNA-seq is effective for capturing transcriptome diversity (Kuo et al., 2020; Seki et al., 2021; Ramberg et al., 2021), however, few studies have applied this approach to study embryogenesis. A comprehensive long-read transcriptome of pre-

implantation mouse embryos was constructed with PacBio technology (Qiao et al., 2020), whilst ONT-based RNA-seq characterised 1,768 new genes and 79,810 transcripts expressed during embryogenesis of the fruit fly *Bactrocera oleae* (Bayega et al., 2021). These studies highlighted the power of long-read RNA-seq for studying embryonic transcriptomes. However, their focus was on transcriptome assembly, with limited quantification analysis. As such, no long-read RNA-seq studies have applied to studying transcript-level expression dynamics in embryogenesis of non-model teleosts.

This chapter builds on approaches detailed in Chapter 3 to examine transcript expression dynamics across a timecourse of Atlantic salmon development. My analysis covers six key stages of embryogenesis (late blastulation, mid gastrulation, early somitogenesis, mid somitogenesis, late somitogenesis and late-eyed stage), which span several key developmental transitions. As well as conducting DTE analysis across stages using a general linear model, I describe a novel method for exploring alternative transcript usage at distinct phases of embryogenesis.

4.2 Materials and Methods

4.2.1 Embryo Data Overview

173,423 transcripts constituting 26,539 genes in the long-read transcriptome (Chapter 2) were supported by at least 3 full-length reads in a minimum of 2 biological replicates per development stage. UpSet plots were generated with the aim to visualise which stages of development were contributing to each gene and transcript model. The UpSetR package (Conway et al., 2017) was used in RStudio (R4.3.3) to generate two UpSet plots showing the combination of developmental stages supporting each transcript and gene model in the final transcriptome.

4.2.2 Quantifying Transcript Expression

My approach to quantify transcript expression in the head kidney dataset (Chapter 3, section 3.2.2) was applied to the embryo dataset. In brief, full-length reads from all replicates for the six key stages of development were mapped against my long-read transcriptome using Minimap2 v2.24 (Li, 2021)

with options "`-ax map-ont N100 -t 8`". Non-primary alignments were removed from the resultant .bam alignment files with "`samtools view -b -F 2308`" before primary alignments were sorted using SAMtools (Danecek et al., 2021). Transcript expression was quantified per replicate with Salmon v1.8.0 (Patro et al., 2017) in 'alignment-based' mode with options "`-p 8, -l A, --ont`". The raw count data and TPMs for all biological replicates was imported into R for analyses described below.

4.2.3 Data Exploration and Quality Check

Counts from Salmon were imported into R and filtered using the "`filterByExpr()`" function included in the edgeR package v4.0.16 (Chen et al., 2024) with options: "`min.count = 5, min.total.count = 10, min.prop = 0.66`". This approach retained robustly evidenced transcripts which had five or more counts per million (CPM) in at least 2 out of 3 biological replicates for at least 1 out of 6 stages of development. The dataset was then transformed using the variance stabilising transformation (Anders & Huber, 2010) function, "`vst(, blind = FALSE`)" from the DESeq2 R package v1.42.1 (Love et al., 2014).

The "`plotPCA()`" function from DESeq2 was used to generate a PCA plot of the first two PCs using ggplot2 v3.5.0 (Wickham, 2016). DESeq2 was also used to calculate sample-to-sample Euclidean distances and perform hierarchical clustering. A sample-to-sample heatmap based on the clustering was plotted with the R package pheatmap v1.0.12 (Kolde, 2019).

4.2.4 DTE Analysis

For the embryonic timecourse data, DTE was carried out across all developmental stages using the generalised linear model and quasi-likelihood F-test function "`glmQLFTest()`" in the edgeR R package v4.0.16. This is a similar approach to that used by Harvey et al. (2024). Filtered transcript counts were normalised based on library sizes with "`normLibSizes()`" before dispersions were calculated with "`estimateDisp()`". FDR adjusted $p<0.05$ was the cut-off to define DETs, representing transcripts regulated during embryogenesis.

TPMs of DETs were scaled manually, before the pheatmap R package was used to generate a heatmap of transcript expression. Hierarchical clustering was used to split the DETs into 6 distinct clusters.

4.2.5 Clustering with Self-Organising Maps

SOM clustering is a dimensionality reduction method devised by Kohonen (1982), which retains topological relationships between clustered groupings of data. Thus, for a transcript expression dataset, transcripts showing similar expression patterns will be clustered together, with clusters showing similar expression existing close to one other in the topological space.

Before SOM clustering, TPM values for each transcript that passed the EdgeR filtering above were scaled using the "`scale()`" function provided by base R (R Core Team, 2024) with options "`scale = TRUE, center = FALSE`". This approach standardised the standard deviations of all transcripts to a value of 1, preserving relative changes in expression whilst removing absolute differences, hence limiting biases caused by highly expressed transcripts. SOM clustering was carried out on the scaled, filtered TPMs with the kohonen R package v3.0.12 (Wehrens & Kruisselbrink, 2018) using a 4x4 manifold with hexagonal topology and PCA initialisation.

SOM clusters were visualised as violin plots using a custom R script adapted from a PhD thesis (Perojil-Morata, 2024) to show average normalised expression profiles of each SOM cluster. Violin plots were displayed in order of development from early to late based on transcript expression within the clusters. Colours were generated for each SOM cluster in R using the inferno palette from the viridis package; constitutive clusters, that is clusters of transcripts displaying relatively similar expression across all stages of development were assigned a grey colour.

The complexHeatmap R package v2.18.0 (Gu, 2022) was used to generate heatmaps to visualise variation in transcript expression across and within SOM clusters through development. The heatmap was coloured using a colourblind-friendly palette (cividis) from the R package viridis v0.6.5 (Garnier et al., 2024). Each SOM cluster in the heatmap was manually annotated with

its respective SOM cluster colour used in the violin plots. Additional colours were added to the top of the heatmap to annotate each of the 6 stages of development.

Transcript expression dynamics across developmental stages were visualised using the Uniform Manifold Approximation and Projection (UMAP) algorithm (McInnes et al., 2020). UMAPs is a dimensionality reduction approach that provides two-dimensional representations of data with complex underlying structures and patterns. The uwot R package v0.2.2 (Melville, 2024) was used for UMAP creation. Each UMAP data point represents a scaled, per-stage mean of TPMs across all replicates for a single transcript; points were coloured using the same colour assigned to the SOM cluster in the violin plots and heatmap above.

The transcript-to-gene ratio for each SOM cluster was calculated by counting the total number of transcripts in each cluster and dividing by the number of genes in the same cluster, and plotted using ggplot2.

4.2.6 Gene Ontology Analysis

Gene ontology (GO) enrichment analysis was performed on each SOM cluster to determine biological processes overrepresented among transcripts according to their expression during embryogenesis. First, Ensembl transcript IDs overlapping my transcript models (identified by SQANTI3, Chapter 2) were imported into R and then assigned to their overlapping long-read transcript in the edgeR-filtered subset with a custom R script. Next, the R package AnnotationForge v1.44.0 (Carlson & Pagès, 2024) was used to create a custom R .db object from the Ensembl annotation extracted using biomaRt v2.58.2 (Durink et al, 2005; Durink et al, 2009). GO enrichment was carried out on the unique set of filtered transcripts for each SOM cluster using the "`compareCluster()`" function in the clusterProfiler R package v4.10.1 (Yu et al., 2012; Wu et al., 2021). The full set of filtered transcripts possessing associated Ensembl transcript IDs was set as the background and the analysis restricted to 'Biological Processes' ontologies.

The "`dotplot()`" function, in conjunction with the R package ggplot2, was used to plot the top 5 enriched GO terms represented in each SOM cluster.

4.2.7 Differential Transcript Usage

Whilst tools for identifying DTU, e.g. IsoformSwitchAnalyzeR (Vitting-Seerup & Sandelin, 2019) are emerging, they are designed for pairwise comparisons, hence unsuitable for identifying DTU in complex timecourse datasets.

As an alternative approach to prioritise DTUs in this chapter, I focussed on identifying transcripts derived from the same genes that showed: 1) significant differential expression across embryogenesis (i.e. they were DETs in the above edgeR GLM analysis) and 2) showed distinct expression patterns during embryogenesis according to SOM clustering (i.e. transcripts belonging to distinct clusters). Thus, an additional round of scaling and SOM clustering was conducted on the TPMs for the DETs identified by edgeR. A 2x3 manifold was used to produce 6 master clusters each representing a major stage of embryonic development in the timecourse. A master list of DETs derived from the same genes and located in different SOM clusters was generated using a custom R script. Genes with only 1 DET or with DETs placed in a single SOM cluster were removed. The UpSetR package was used to visualise genes with DETs located across distinct combination of SOM clusters.

Ensembl gene IDs, transcript IDs, and gene names were extracted from the Ssal_v3.1 annotation using biomaRt. For genes lacking a gene name, a search for orthologous gene names was conducted in rainbow trout, zebrafish, mouse or human in that order of priority (as done in Chapter 3, section 3.2.4).

For candidate genes displaying evidence of DTU across multiple SOM clusters, the expression levels of DETs, as well as a mean expression level for all non-DETs was plotted using ggplot2. For each DET, TPM values were averaged across the 3 replicates at each stage of development, whilst for the non-DETs the TPMs for all replicates and transcripts were averaged. Mean TPM was plotted as a line plot with ggplot2, using a custom colour palette; for the DETs, SD error bars were added, whilst a ribbon showing SD was used for the non-DET mean expression to aid visualisation. The R package

ggtranscript v0.99.9 (Gustavsson et al., 2022) was used to visualise transcript structures for each gene of interest, its associated gene in the Ensembl Ssal_v3.1 annotation, and to display predicted promoters from the Ssal_v3.1 regulatory build. Coordinates of each feature were extracted from gtf and gff3 files with a custom "`awk`" script, before being imported into R with the rtracklayer R package v1.62.0 (Lawrence et al., 2009) for visualisation. UTRs were partitioned in the Ensembl annotations to show coding-sequences in the reference models, and the Ensembl canonical transcript marked with an asterisk. The same custom colour palette in the DET SOM clustering was used to annotate the DETs based on SOM cluster membership, whilst the non-DETs were coloured in grey. Both the line plots and transcript model plots were exported as svg files and imported into InkScape to combine into a single figure.

A second round of GO enrichment analysis was carried out on all genes showing DTU across multiple SOM clusters using clusterProfiler v4.10.1. In Chapter 2, if an overlap was found between a long-read transcript and an Ensembl reference gene, SQANTI3 detailed the Ensembl gene ID in its classification output file. For this second round of GO enrichment analysis, all associated Ensembl gene IDs for the transcripts in the DTU subset were extracted from the SQANTI3 classification file. The same custom .db object produced in section 3.2.5 was used for this round of GO analysis. The "`enrichGO()`" function was used for GO analysis with the background consisting of all genes with an associated Ensembl ID (as determined by SQANTI3) in the filtered dataset, and GO terms restricted to "Biological Processes". GO results were summarised in tabular format with the flextable R package v0.9.6 (Gohel & Skintzos, 2024).

4.3 Results

4.3.1 Data Overview and Quality Assessment

Of the 266,222 transcripts and 35,480 gene models described in the long-read transcriptome (see Chapter 2), 173,423 and 26,539 were supported by the embryonic dataset, respectively. Most of these gene models (53.8%) were supported by reads from all stages of development (Figure 4.1A), with

4.4% supported by a single stage. However, approximately half of all transcripts supported by embryonic reads were stage-specific (Figure 4.1B). Transcripts supported by embryonic reads common to all development stages constituted only 10.5% of all transcripts supported by embryonic reads. Just under half the samples in the embryo dataset (8/18) possessed greater than 600,000 full-length reads (Table 4.1), which is considered an indicator of “good” data quality by the ENCODE long-read RNA-seq data standards (ENCODE Project Consortium, 2025). Another 8 of the samples had 400,000-600,000 full-length reads and are deemed “acceptable” for read depth by the same standards. 2 samples had significantly fewer reads than the other samples pushing them into the “poor” category (ENCODE Project Consortium, 2025).

33,604 transcripts produced by 8,524 unique genes passed filtering with edgeR. A PCA using this filtered dataset revealed tight grouping of biological replicates for each stage of development, with sequential stages of embryogenesis clustering proximally (Figure 4.2). Most variation was explained by PC1 and PC2, 89.0% and 6.7%, respectively. PC1 mainly separates late somitogenesis and late-eyed stages, whilst PC2 separates the earlier stages of development. Hierarchical clustering similarly showed that biological replicates clustered together within their respective stage, and that consecutive stages clustered together (Figure 4.3).

4.3.2 Transcript Expression Patterns Resolved with SOM Clustering

SOM clustering of transcripts that passed edgeR filtering resulted in 16 distinct clusters (Figure 4.4). A heatmap visualisation revealed cluster-specific patterns of expression (Figure 4.5); 13 SOM clusters showed stage-specific expression patterns, whilst 3 showed more constitutive expression across stages. SOM1 and SOM2 captured transcripts with highest expression in the late blastula stage, with a subsequent decrease in expression as embryogenesis progresses. SOM4 and SOM5 captured transcripts showing high expression in both late blastulation and mid gastrulation. Transcripts upregulated specifically during gastrulation were represented by SOM13, which also shows a lower expression level in early

somitogenesis. SOM9 appears to capture high expression across all three of the early development stages, with a slight reduction at mid-gastrulation. SOM14, 15, 16 and 12 captured highest expression during somitogenesis, whilst SOM4 captured low expression outside of the late-eyed stage. Developmental expression dynamics were further demonstrated by UMAP visualisation, which captured the trajectory of expression as the embryo transitions from a blastula to a juvenile fish possessing differentiated tissues and delineated organs (Figure 4.6).

A variable transcript-to-gene ratio was observed in each SOM cluster (Figure 4.7). Whilst SOM4, possessing transcripts expressed the latest in the development series, had the greatest transcript-to-gene ratio, the second highest ratio was seen in SOM11, a constitutive cluster. In general, low transcript-to-gene gene ratios were observed in early-stage SOM clusters and in constitutive cluster SOM7, with the lowest transcript-to-gene ratio observed in SOM13, primarily associated with mid-gastrulation.

Transcript-level GO analysis revealed biological processes overrepresented in each SOM cluster (Figure 4.8). Early development clusters SOM1 and SOM2 were enriched for GO terms “GO:0007049 - cell cycle” and “GO:0044085 - cellular component biogenesis”, consistent with processes involved in cell cleavages. Terms associated with the commencement of metabolic processes are enriched in SOM clusters 15, 12, and 3, comprising transcripts highly expressed during somitogenesis, where primordial organs are forming (Gorodilov, 1996). GO terms associated with transport and localisation such as “GO:0015669 gas transport” and “GO:0051179 localization” are enriched in SOM4, a late-eyed-specific SOM cluster. Constitutive clusters SOM11 and SOM10 showed enrichment of terms associated with amide and peptide processes, as did SOM15 which is comprised of transcripts with high expression across the three somitogenesis stages. There were no enriched GO terms in SOM14, a cluster showing high transcript expression at early somitogenesis, and the constitutive cluster SOM7.

4.3.3 Characterisation of DET Expression Profiles with SOM Clustering

DTE analysis with edgeR identified 13,471 DETs constituting 4,211 genes across the embryonic timecourse. DTE was used to identify DETs which were then taken forward into a second round of SOM clustering for examining within-gene differential transcript usage. SOM clustering of DETs identified 6 clusters and as expected, no constitutive clusters (Figure 4.9). Two highly stage-specific stages were represented by SOM2 and SOM6 containing DETs with high expression almost exclusively in late blastulation and late-eyed stages respectively. SOM clusters separating SOM2 and SOM6 captured transcripts showing variable expression over multiple development stages. For example, SOM4 clustered DETs with high expression across all 3 somitogenesis stages, whilst SOM5 clusters DETs upregulated most strongly in late somitogenesis, but showing high expression in mid somitogenesis and late-eyed stages. DETs with high expression in both late blastulation and mid gastrulation were clustered in SOM1. SOM3 contains DETs expressed from mid-gastrulation to mid-somitogenesis with greatest expression displayed in early somitogenesis. The 6 SOM clusters of DETs were used as the basis for DTU analysis.

4.3.4 Identification of Genes Showing DTU During Salmon Embryogenesis

641 genes comprising 6,047 DETs showed membership of alternative transcripts to at least 2 SOM clusters. 571 (89%) had transcripts expressed in similar neighbouring SOM clusters only (e.g. SOM2 and SOM1; Figure 4.10). 70 genes had transcripts that clustered in either non-neighbouring SOM clusters (e.g. 4 genes had DETs in both SOM2 and SOM4) or multiple neighbouring clusters (e.g. 3 genes had DETs in SOM2, SOM1 and SOM3). The majority of the 6 DET SOM clusters represent transcripts with expression in multiple stages of development and thus do not represent hard barriers between consecutive development stages. As such, I decided to focus on examples of DTU from genes possessing DETs in non-neighbouring SOM clusters.

Out of the 641 genes identified as having signatures of DTU, 605 had associated Ensembl gene IDs. GO enrichment analysis revealed 6 biological

processes over-represented among these genes (Table 4.2). Lipid transport (GO:0006869) and lipoprotein metabolic processes (GO:0015671) were explained by genes including *apoeb* and *apob*, which code for lipid carrying apolipoproteins with roles in transporting yolk and dietary liquids around the developing fish embryo (Otis et al., 2015). Amide metabolic processes (GO:0043603) were explained by genes coding for ribosomal proteins including *rpl3l, rps19* and *rl29.* Translation (GO:0006412) was also enriched, sharing all 42 of its supporting genes with amide metabolic processes. Ribosome protein heterogeneity has been linked to the regulation of translation required for the proper progression of embryogenesis (Norris et al., 2021). Finally, oxygen transport (GO:0015671) was enriched in the DTU gene set, predictably explained by genes encoding haemoglobin proteins such as *hbb* and *hbb-bh1.*

4.3.5 Genes Showing DTU During Atlantic Salmon Embryogenesis

4.3.5.1 DTU in Transgelin Gene

The gene model G7805 was found to overlap the Ensembl gene *tagl* (ENSSSAG00000074856), which codes for transgelin, an actin-binding protein family member, essential for cellular cytokinesis and intracellular transport (Pollard, 2016). Transgelin was shown to be involved with smooth muscle formation during embryogenesis (Santoro et al., 2009; Hsieh & Jin, 2023) and cytoskeleton formation in humans (Elsefadi et al., 2016).

G7805 possesses 2 DETs, G7805.3 and G7805.6, which were expressed in SOM2 (early development) and SOM4 (mid-late development) respectively (Figure 4.9). The mean per-stage TPM expression supports the SOM clustering classification, with G7805.3 displaying highest expression in late blastulation, followed by decreasing expression towards the eyed stage (Figure 4.11). In contrast, G7805.6 shows low expression in early development, with expression peaking during early somitogenesis. Expression remains high throughout somitogenesis before decreasing towards the late-eyed stage (Figure 4.11). In total, G7805 has 9 transcripts in the long-read annotation, 7 of which were not identified as significantly expressed. Plotting all (Figure 4.11) for G7805 and comparing them with the

three Ensembl transcripts reveals that the two isoforms contributing to DTU differ in their TSS and exon structure. The promoter region identified in the Ensembl regulatory build supports 3 out of 9 long-read transcripts, including G7805.6, which is most expressed in somitogenesis. G7805.3 is not supported by the promoter annotation, however, several models in the long-read annotation possess TSS approximately 3kb downstream of the promotor region. The TSS of the canonical Ensembl transcript is not supported by either promoter or long-read annotations.

4.3.5.2 DTU in Ribosomal Protein L3

Ribosomal protein genes constitute a significant proportion of the genes in the DTU subset (section 4.3.4). G26082 overlaps *rpl3l* (ENSSSAG00000001684), which codes for an L3-like ribosomal protein. This model showed three DETs clustered into two distinct SOM clusters. G26082.3 was captured by SOM1, whilst G26082.7 and G26082.10 were captured by SOM6. Plotting transcript expression supports these classifications with a sharp increase in expression of G26082.7 and G26082.10 at the eyed stage, whilst G26082.3 shows highest expression in late blastulation, with a subsequent reduction during early-mid somitogenesis and further decrease towards late somitogenesis (Figure 4.12). The three DET models differ in their TSS, with G26082.10 possessing one fewer exon (Figure 4.12). TSSs for all three models fall within Ensembl annotated UTRs. There are two promoters associated with this gene, and both support the long-read transcript models.

G26082 has an ohnologue pair (G29398) retained from Ss4R (Chapter 1, section 1.4.2) in collinear blocks 3q-6p (Lien et al., 2016) that also displayed evidence of DTU. G29398 has 6 transcripts, all significantly upregulated during development. G29398.3, G29398.18, G29398.22 and G29398.23 were classified into SOM6, whilst G29398.24 and G29398.25 were classified into SOM5 and SOM1 respectively. The expression patterns of these transcripts support their SOM classifications (Figure 4.13) and indeed match those of G26082 (inlayed Figure 4.13). However, the 6 DETs for G29398 display striking differences in exon structure. G29398.3 shows a large intron

retention event encompassing exon 1, or exons 1 and 2, of the other 4 late development isoforms (.18, .22, .23, .24). G29398.25, expressed in SOM1, has a different exon chaining pattern, with no exonic sequence in the intron-retention region encompassed by G29398.3 (Figure 4.13). These 5' end differences again overlap the UTR regions predicted by Ensembl.

4.3.5.3 Exon Skipping in a Phosphate Carrier Protein

G31138 has 3 isoforms contributing to its DTU signature and associated with Ensembl gene *slc25a3b* (ENSSSAG00055931), which encodes a phosphate carrier protein located in the inner mitochondrial membrane involved in oxidative phosphorylation (Boulet et al., 2018; Peoples et al., 2021). G31138.6 and G31138.8 are in SOM2 and SOM1, respectively, which show highest expression in early development, followed by a gradual decline in expression as development progresses (Figure 4.14). G31138.10 is a member of SOM5, characterised by low expression during blastulation and an increase in expression towards late somitogenesis and the eyed stage. All three of these transcripts comprise 8 exons, with differences evident in exon 3, which is coding sequence according to Ensembl (Figure 4.14). The canonical Ensembl transcript possesses a TSS not shared by any of the long-read models, nor by the Ensembl promoter. All but 2 of the 21 G31138 transcripts have a TSS within the bounds of the Ensembl promoter.

4.4 Discussion

This chapter reports a new approach to apply long-read transcriptomics to explore the dynamics of transcript expression and usage across embryonic development in a non-model fish. The methods reported, namely DTE analysis followed by SOM clustering of DETs, have proved powerful for prioritising genes displaying DTU in a timecourse dataset, successfully differentiating between the expression of structurally distinct transcript isoforms. My approach uses publicly-available tools and is transferable to many experimental designs involving multiple comparisons where clustering is necessary to summarise complex expression dynamics.

4.4.1 Potential Capture of Zygotic Genome Activation

Zygotic genome activation signals the cessation of maternal transcriptional control during embryogenesis (Jukam et al., 2017). The timecourse experiment in this chapter was designed with the intention of being able to capture this change in transcription in the late blastulation. SOM2 of the DET SOM clustering (Figure 4.9) captured transcripts showing highest expression in late blastulation samples, with a clear reduction in expression in mid gastrulation and minimal expression in subsequent development stages. These patterns are consistent with previous studies which report rapid degradation of maternal RNA post-blastulation (Aanes et al., 2011; Walser & Lipshitz, 2011; Vastenhouw et al., 2019; Solnica-Krezel, 2020; Perojil-Morata 2024).

Early embryonic development relies on translation of maternal mRNA by maternally-deposited ribosomal proteins (Cenik et al., 2019; Breznak et al., 2023). Genes encoding ribosome component proteins were highly common among those showing DTU. Indeed, *rpl3l* was shown to express distinct transcripts in SOM2 and SOM6, the earliest and latest development clusters. In both ohnologues of *rpl3l,* only a single isoform was expressed in early development SOM clusters, whereas there was an increase in transcript diversity expressed in later development from the same gene. This result may have captured the switch between maternal ribosomal machinery, to that of the zygote, which needs to express diverse ribosomal proteins as the embryo undergoes differentiation and organogenesis (Pollard, 2016).

The maternal RNA repertoire deposited into the oocyte has been linked with successful development of embryos, and proposed as a marker of egg quality in farmed fish (Sullivan et al., 2015; Reading et al., 2018; Weber et al., 2021). Increasing our knowledge of the diversity of transcripts involved in maternal transcriptional control during early embryonic development may allow better understanding of the impact of maternal RNA on the processes underlying successful embryonic progression in aquaculture species, while also offering novel markers of egg quality in the future.

4.4.2 Characterisation of Structural Changes in Alternative Transcripts

DETs used in distinct development stages were differentiated by structural variations including exon skipping and alternative TSSs. In the cases of G7805 (Figure 4.11), G26082 (Figure 4.12), and G29398 (Figure 4.13), alternative TSSs were employed by transcripts expressed in later development. This is consistent with Nepal et al. (2013), who revealed that after ZGA, new transcript variants with alternative TSSs can be expressed from the same loci as maternally-derived transcripts.

The identification of a plethora of novel transcript structures in this chapter, including changes to predicted coding regions such as that captured in G31138, is important for understanding the transcriptional mechanisms governing embryonic development. Structural changes in protein-coding regions of transcripts can potentially lead to the production of proteins with distinct biological functions (Wright et al., 2022). However, the role of transcript diversity is poorly understood in salmonids. The data in this chapter represent a push towards characterising and annotating transcript diversity in early development of salmonids.

4.4.3 Flexibility of SOM Clustering Approach

Perojil-Morata (2024) demonstrated the ability of SOM clustering to profile gene expression during Atlantic salmon embryogenesis using short-read RNA-seq. This chapter builds on those foundations and successfully applies the method to long-read RNA-seq data to examine transcript-level expression dynamics. In comparison with traditional pairwise methods of differential expression analysis, SOM clustering can be scaled to accommodate complex experimental designs such as the timecourse detailed here, making it suitable for a wide variety of study designs and applicable to both model and non-model species. This chapter in particular demonstrates that long-read RNA-seq and subsequent SOM clustering analysis is a powerful tool for alternative transcript discovery and characterisation of transcript-level expression patterns.

4.4.4 Reflections on Normalisation Approach

This chapter employed TPM normalisation followed by z-score scaling to produce normalised transcript expression values between samples. Whilst an established approach for both DTE and SOM clustering analyses (Perojil-Morata, 2024), potential improvements in false-discovery rates could be yielded by using an alternative normalisation approach. For example, using a spike-in control RNA sample to perform conventional normalisation techniques on, then subsequently applying a spike-in-based normalisation factor to the experimental data (Evans et al., 2018) may be more appropriate for DTE where the focus is on transcript expression rather than the cumulative expression forming total gene expression (e.g., Byrne et al., 2017).

In the DTE analysis, lowly expressed transcripts were removed if they did not possess at least 5 counts per million in at least 2/3 samples within a single developmental stage. Whilst this threshold was set to be commensurate with previous analyses and short-read methods, this meant that some transcripts were retained if they were expressed in only 2/18 total samples. Increasing this threshold to retaining transcripts expressed in at least 3/3 samples in a single treatment group, or introducing a stricter across all sample read count filter (e.g 20 reads or more across all samples), may improve the robustness of the DTE and SOM clustering.

4.4.5 Concluding Words

The use of long-read RNA-seq for characterising transcript diversity and expression is limited for studying embryogenesis. In this chapter, I successfully identify significant differential transcript expression across a timecourse of 6 embryonic stages in Atlantic salmon. This data enhances our understanding of transcriptional regulation during Atlantic salmon and serves as a valuable resource for further research into embryonic transcriptome expression. In addition, I present a versatile methodology that may be transferred to other species and experimental designs.

In relation to this thesis, this chapter builds on quantitative approaches for transcript-level expression analysis introduced in Chapter 3 and the data

herein forms a foundation to better understand transcriptional regulation displayed in salmonid embryonic development.

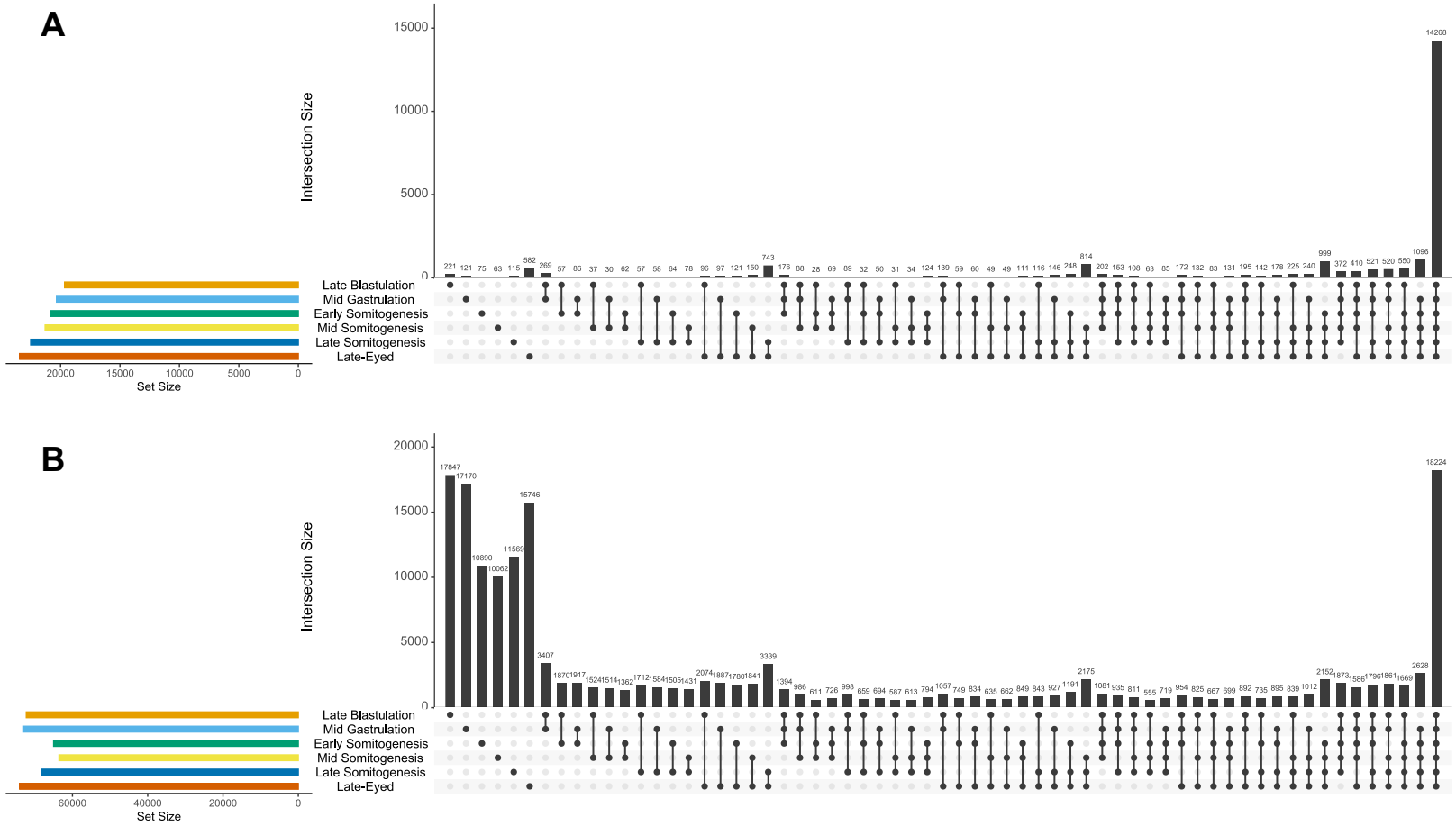


Figure 4.1: UpSet plots showing the number of genes (A) and transcripts (B) supported by full-length embryonic reads. The sets on the left represent stages of development, dots at the bottom combinations of stages, whilst bars show the number of reads in each combination of stages.

Table 4.1: Number of reads for each embryo sample during transcriptome assembly pipeline

Sample	Development Stage	Number of Filtered Reads (q>7)	Number of Full-Length Reads	Number of Primary Alignments
LB_R1	Late Blastulation	2,001,768	542,482	497,792
LB_R2		2,235,605	675,788	596,872
LB_R3		929,058	271,961	229,113
MG_R1	Mid Gastrulation	2,129,501	668,162	626,798
MG_R2		1,032,196	328,056	281,486
MG_R3		2,337,840	695,255	633,523
ES_R1	Early Somitogenesis	1,805,464	538,582	469,632
ES_R2		2,018,411	577,174	530,395
ES_R3		1,633,852	493,427	395,569
MS_R1	Mid Somitogenesis	1,598,574	459,779	413,289
MS_R2		1,867,737	561,247	530,224
MS_R3		1,849,907	531,086	488,157
LS_R1	Late Somitogenesis	1,912,119	537,673	504,174
LS_R2		2,181,588	703,152	608,789
LS_R3		1,957,175	624,001	560,389
LE_R1	Late-Eyed	2,783,932	896,237	782,815
LE_R2		2,421,376	773,813	685,295
LE_R3		2,699,711	831,780	780,549

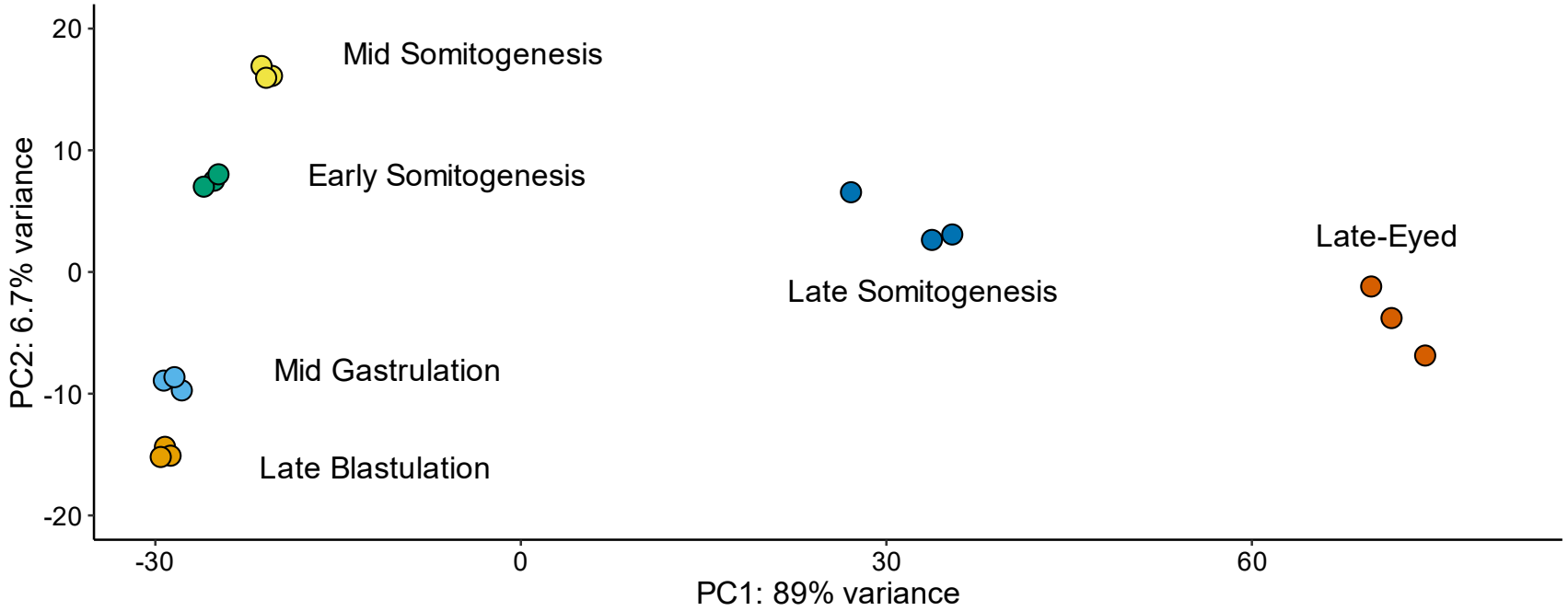


Figure 4.2: PCA plots for Atlantic salmon embryonic development timecourse. 6 stages of development (n=3) were sequenced with long-read Nanopore RNA-seq. PCs generated from vst-transformed counts for long-read transcript models possessing a minimum of 5 reads in at least 2/3 replicates in at least one stage of development. Count filtering conducted with edgeR. Dots represent individual samples, colours represent stages of development.

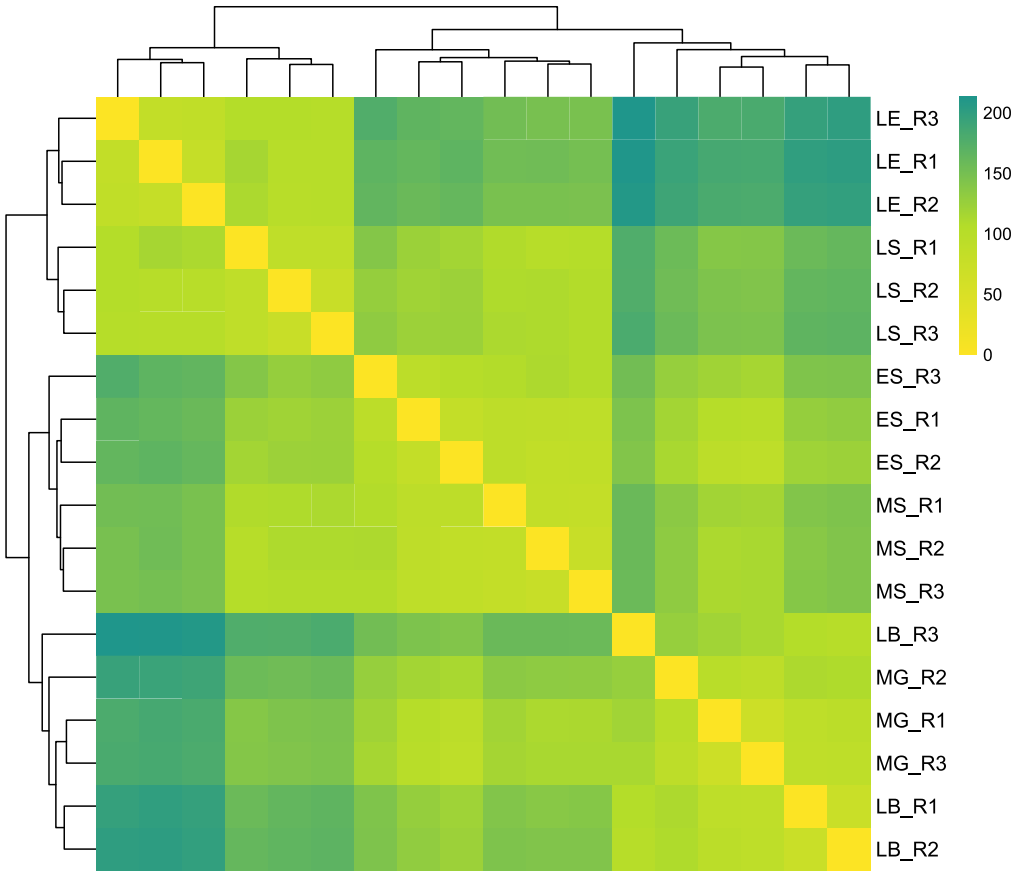


Figure 4.3: Sample similarity matrix plot of vst-transformed counts for Atlantic salmon embryo transcripts sampled at 6 stages of embryogenesis. Transcripts were filtered with edgeR based on read counts and retained if they possessed a minimum 5 reads in at least 2/3 replicates in at least one stage of development. Data displayed is Euclidean distance and dendrogram based on hierarchical clustering. Sample names shown on right-hand side indicate development stage: LB = Late Blastulation, MG = Mid Gastrulation, ES = Early Somitogenesis, MS = Mid Somitogenesis, LS = Late Somitogenesis, LE = Late-Eyed. Heatmap shows tight within-stage clustering of replicates except LB_R3 which clusters with the mid gastrulation group. Additionally, stages cluster sequentially from early to late (bottom to top).

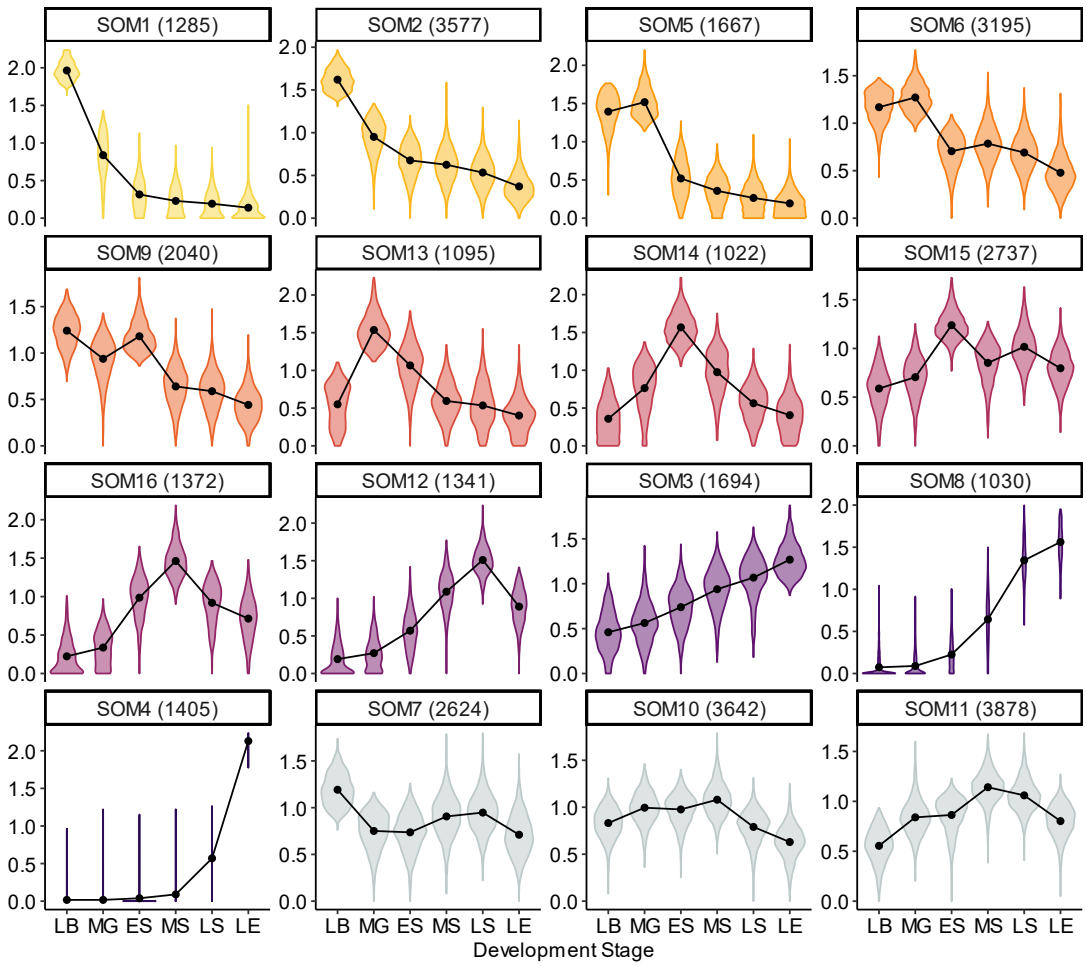


Figure 4.4: Violin plots of transcript expression following standard deviation standardisation for each SOM cluster. The lines in each plot denote expression trends, whilst the violins show the expression distribution for each stage of development. The numbers aside each "SOM" heading show the number of transcripts within that SOM cluster. SOM clusters are ordered from early development cluster (SOM1) to late development cluster (SOM4) from left to right reading down the rows. SOM clusters 7, 10 and 11 represent constitutive expression, i.e. clusters of transcripts showing relatively similar expression levels across all stages of development. SOM clusters are coloured according to the progression of development with clusters showing early-stage expression coloured in yellower hues, middle stage clusters coloured in orange hues and late-stage clusters coloured in purple hues. Constitutive clusters are coloured in grey.

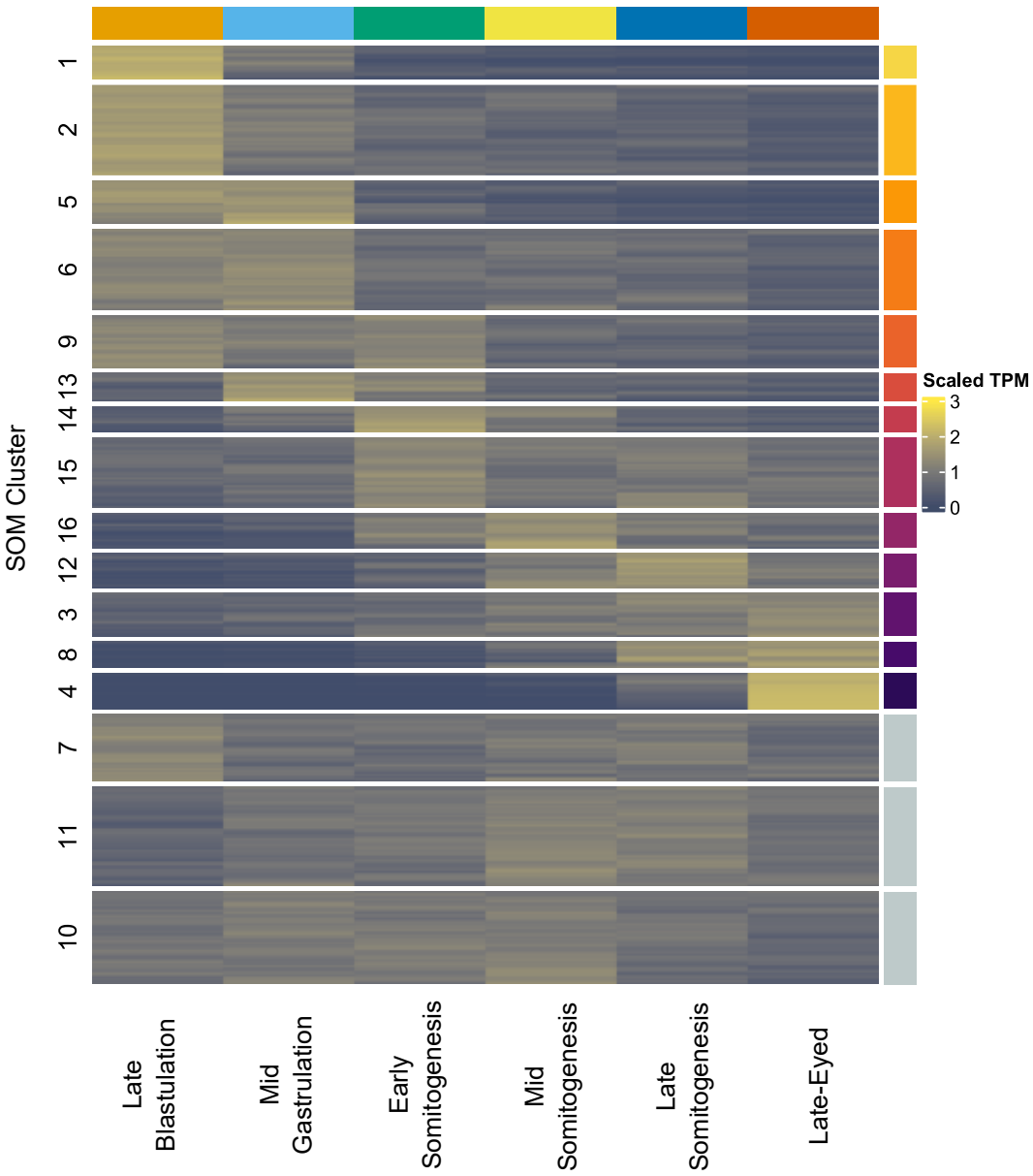


Figure 4.5: Heatmap of transcript expression showing normalised TPM values. Cluster numbers generated through SOM are displayed on the left. The colours on the right denote the type of cluster, with stage-specific clusters in vibrant colours matching Figure 4.4, whilst constitutive clusters are de-saturated. The column colours denote the development stage and match Figures 4.1 and 4.2. SOM clusters are coloured according to the progression of development with clusters showing early-stage expression coloured in yellower hues, middle stage clusters coloured in orange hues and late-stage clusters coloured in purple hues. Constitutive clusters are coloured in grey.

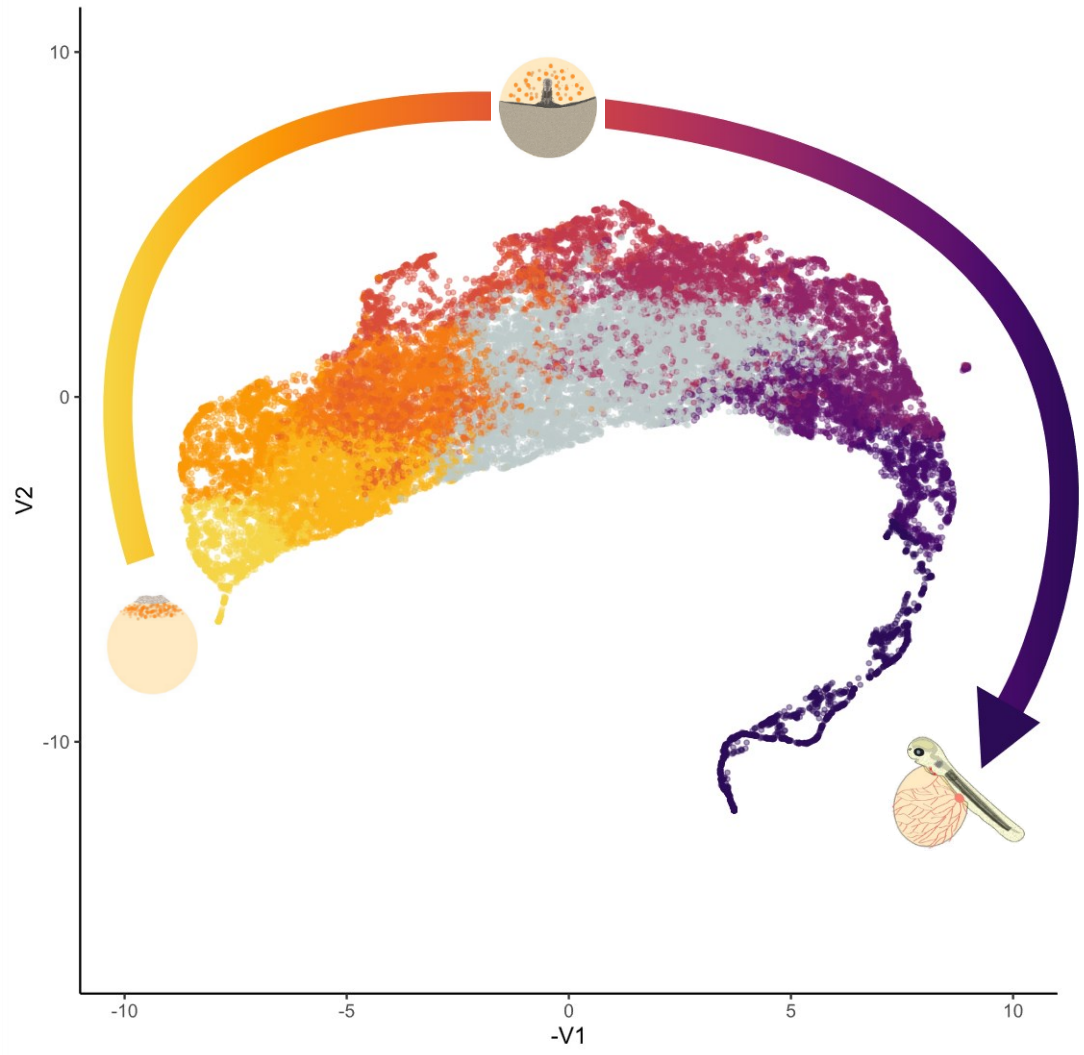


Figure 4.6: UMAP of transcript expression of the edgeR filtered long-read dataset across development. Each point in the UMAP is an individual transcript coloured according to which SOM cluster it belongs to. UMAP is a dimensionality reduction method and as such the x-axis and y-axis are labelled V1 and V2 which denotes the first two components of reduced dimensional space. SOM clusters are coloured according to the progression of development with clusters showing early-stage expression coloured in yellower hues, middle stage clusters coloured in orange hues and late-stage clusters coloured in purple hues. Constitutive clusters are coloured in grey. The arrow shows direction of development from blastulation to late-eyed stage. Diagrams are a visualisation representing the developmental progression path in the UMAP from blastulation, through somitogenesis and ending with the late-eyed stage.

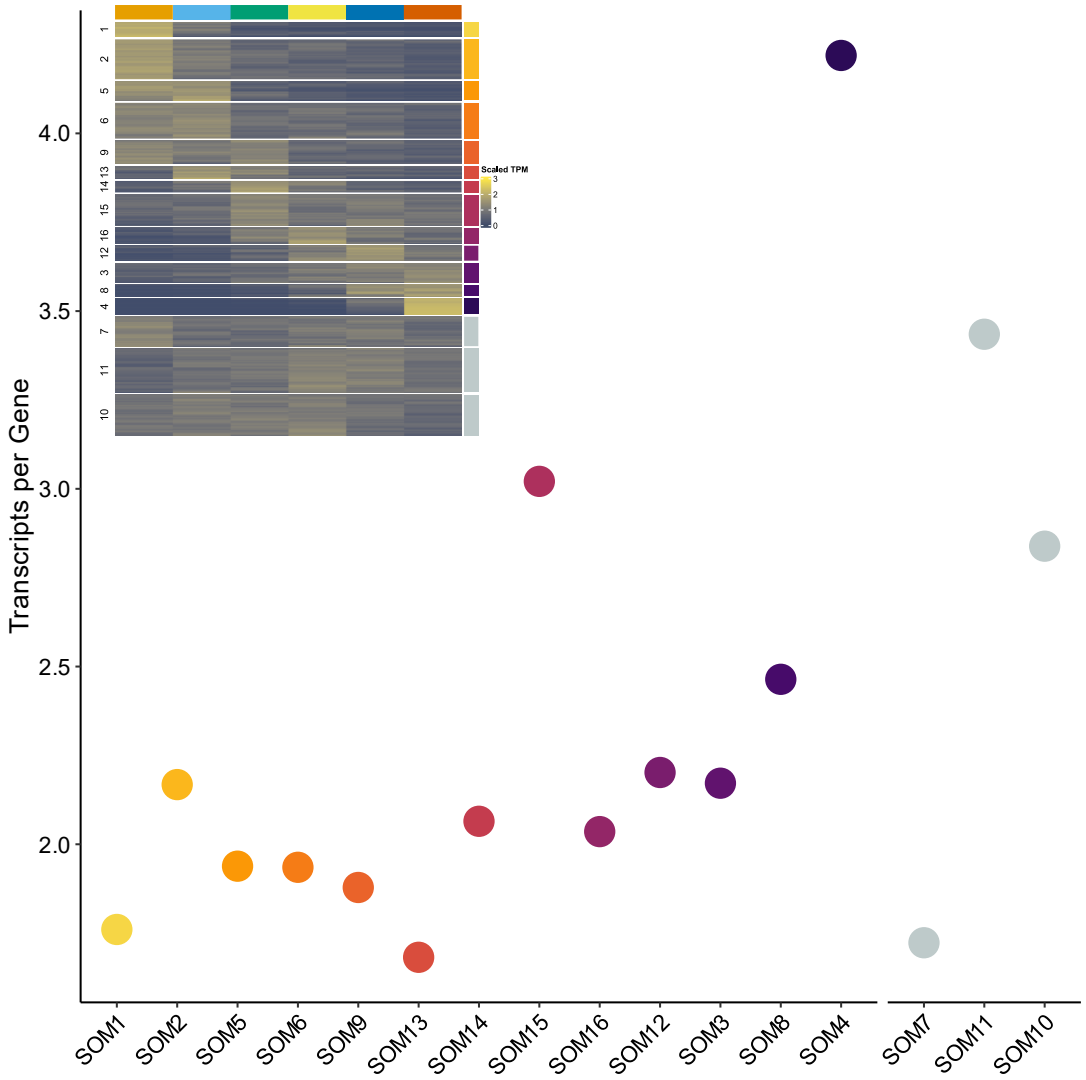


Figure 4.7: Transcript-to-gene ratio within each SOM cluster. Transcript-to-gene ratio was calculated by dividing the total number of transcripts by the total number of genes in each SOM cluster. The x-axis shows the SOM clusters ordered by development, early-to-late with a split indicating constitutive clusters. SOM clusters are coloured as per Figure 4.5 showing development from early (yellow) to late (purple) stages. Constitutive clusters are coloured grey as per Figure 4.5.

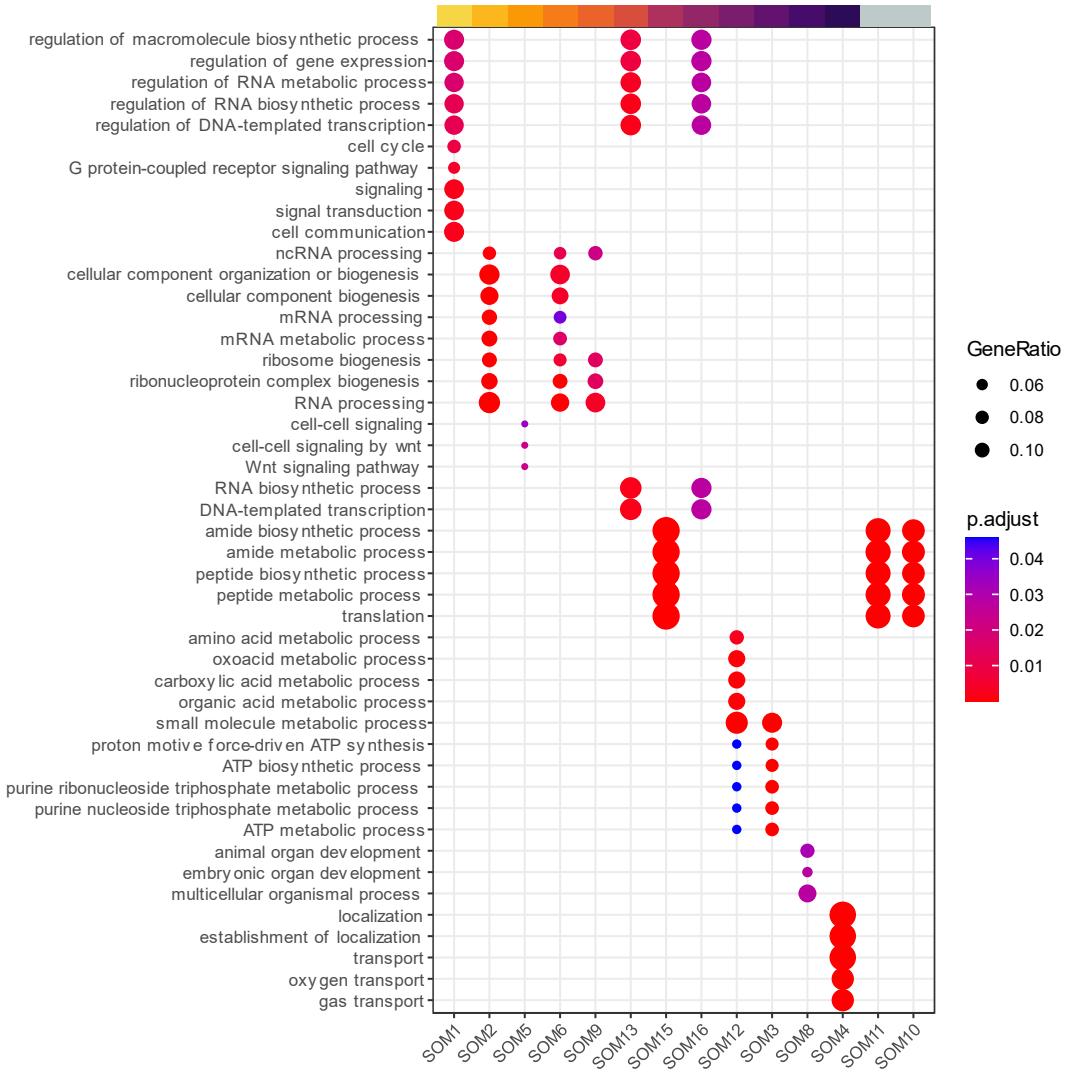


Figure 4.8: Dotplot of enriched GO terms ("Biological Processes") for each SOM cluster. A maximum of 5 distinct GO terms were plotted for each SOM cluster based on FDR adjusted p-value. The clusters on the x-axis are arranged in order of development, from early to late stages. SOM11 and SOM10 are constitutive clusters. The y-axis denotes the GO term with the size of each dot indicating gene ratio and colour indicating the adjusted p-value. The coloured bar at the top indicates SOM cluster membership, and the gradient from yellow to purple is as used in Figures 4.5, 4.6 and 4.7.

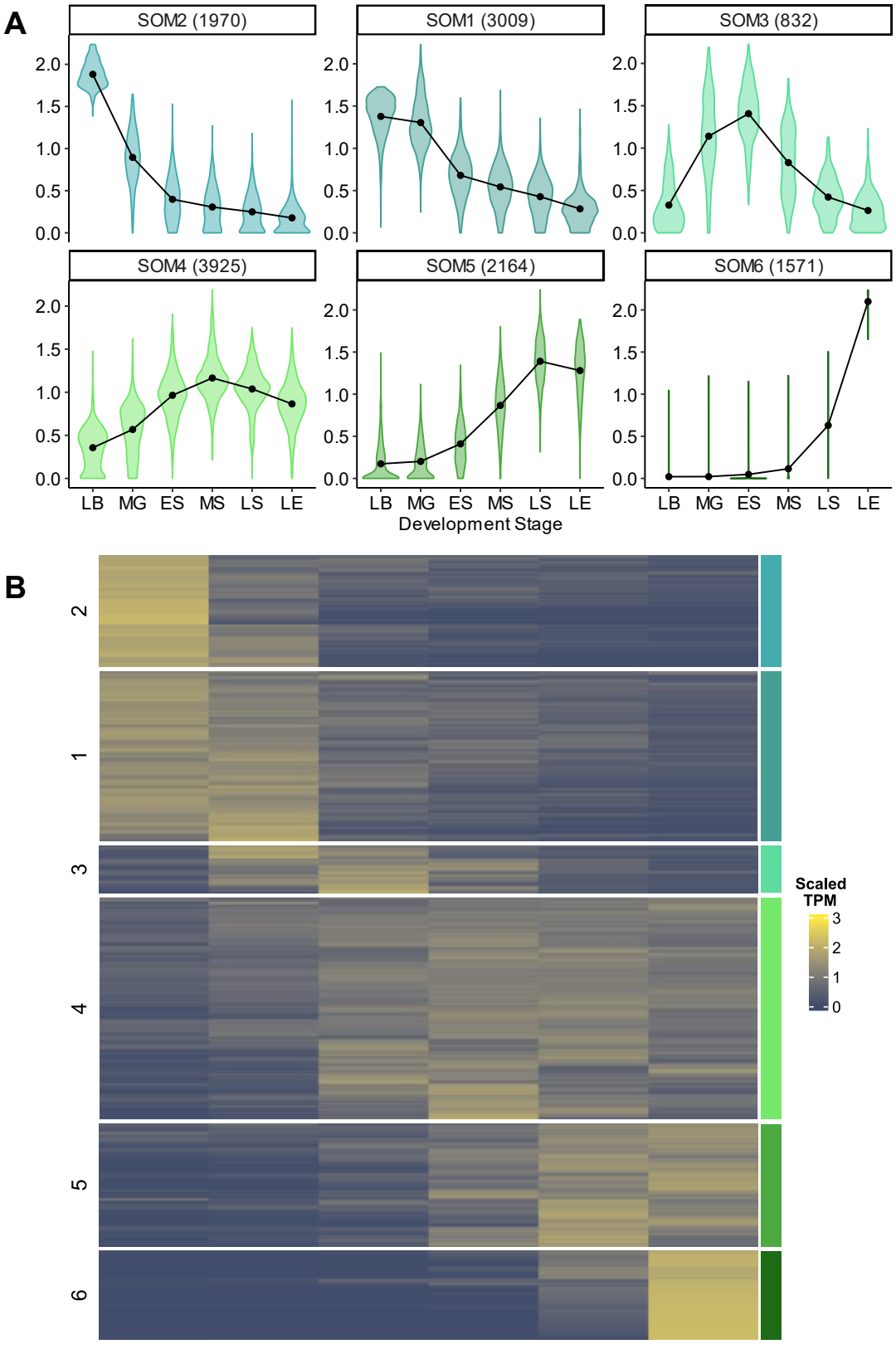


Figure 4.9: SOM clustering of DETs identified by edgeR. (A) violin plots of the 6 master clusters,(B) heatmap of the scaled transcript expression of the DETs. Colours indicate SOM Cluster from early (light blue), to late development (dark green).

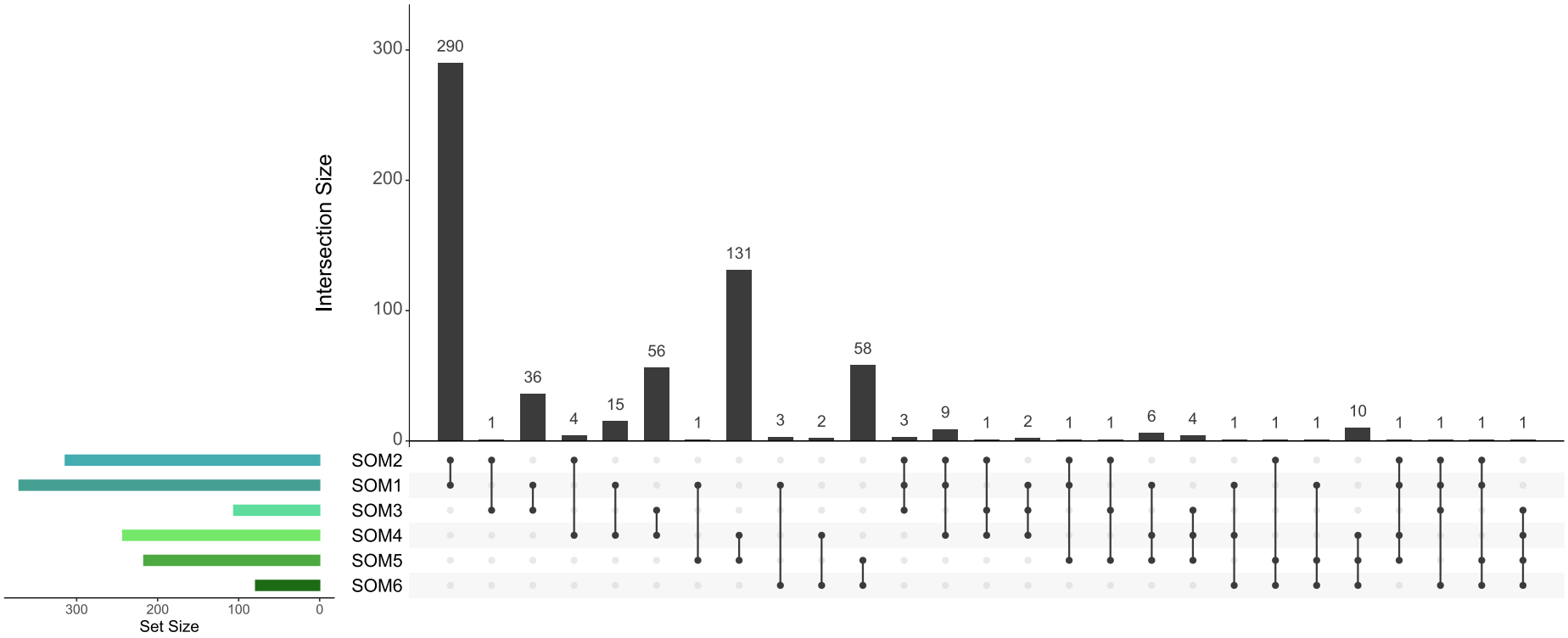


Figure 4.10: UpSet plot showing the number of genes with DETs in multiple SOM clusters. SOM clusters are displayed on the left in order of early to late development along the y-axis (top to bottom, light blue = early development, dark green = late development). Dots indicate DTEs shared among SOM clusters and bars denote the number of DTEs shared by combinations of SOM clusters displayed by the dots.

Table 4.2: Gene Ontology (GO) enrichment results for genes displaying evidence of differential transcript usage.

Term	Gene Count	Adjusted P-Value	Gene Ratio	Background Ratio	Homologous Gene Names (if present)	Contributing Genes (Ensembl ID)
GO:0006869 - lipid transport	12	0.000093	12/278	28/3522	*apoeb* *pg12b* *apoa-i-1* *afp4* *apoeb* *apoa1* *npc2.1* *apob*	ENSSSAG00000086296 ENSSSAG00000085443 ENSSSAG00000010289 ENSSSAG00000047184 ENSSSAG00000115227 ENSSSAG00000103268 ENSSSAG00000007048 ENSSSAG00000120281 ENSSSAG00000097945 ENSSSAG00000003002 ENSSSAG00000065331 ENSSSAG00000102777
GO:0015671 - oxygen transport	10	0.00013	10/278	21/3522	*hbb* *hbba2* *hbb-bh1* *hbae3* *hbbe1*	ENSSSAG00000065233 ENSSSAG00000103747 ENSSSAG00000093862 ENSSSAG00000065398 ENSSSAG00000086616 ENSSSAG00000087487 ENSSSAG00000119552 ENSSSAG00000045065 ENSSSAG00000065229 ENSSSAG00000087439
GO:0042157 - lipoprotein metabolic process	8	0.014	8/278	22/3522	*apoeb* *apoa1* *apoa-i-1* *afp4* *apoc-lb*	ENSSSAG00000003126 ENSSSAG00000086296 ENSSSAG00000085443 ENSSSAG00000115227 ENSSSAG00000007048 ENSSSAG00000010260 ENSSSAG00000120281 ENSSSAG00000097945
GO:0043603 - amide metabolic process	45	0.039	45/278	335/3522	*rpl3l* *rl18* *rpl39* *rps23* *rpl4* *rpl27a* *rps19* *rpl10a* *rpl9* *eef1b2* *rpl26* *rl17* *rpl19* *rl29*	ENSSSAG00000051618 ENSSSAG00000072884 ENSSSAG00000041178 ENSSSAG00000046726 ENSSSAG00000010392 ENSSSAG00000067773 ENSSSAG00000075399 ENSSSAG00000071589 ENSSSAG00000066273 ENSSSAG00000007090 ENSSSAG00000047422 ENSSSAG00000008596 ENSSSAG00000098459 ENSSSAG00000076201 ENSSSAG00000078842 ENSSSAG00000089657

Term	Gene Count	Adjusted P-Value	Gene Ratio	Background Ratio	Homologous Gene Names (if present)	Contributing Genes (Ensembl ID)
					rps13	ENSSSAG00000039434
					rpl10	ENSSSAG00000006614
					rps24	ENSSSAG00000077735
					eef1g	ENSSSAG00000024307
					ef1a	ENSSSAG00000001684
					rpl11	ENSSSAG00000046051
					rpsa	ENSSSAG00000002154
					rpl35a	ENSSSAG00000072444
					zgc: 103559	ENSSSAG00000002140
					rpl35	ENSSSAG00000008448
					rpl13	ENSSSAG00000008668
					uba52	ENSSSAG00000046508
					ef1a	ENSSSAG00000109892
					acer1	ENSSSAG00000045463
					rpl31	ENSSSAG00000058507
					rpl37a	ENSSSAG00000032246
					rps8a	ENSSSAG00000051583
					rps11	ENSSSAG00000071292
					rpl36	ENSSSAG00000018086
					rps16	ENSSSAG00000066907
					rpl32	ENSSSAG00000054819
					cers5	ENSSSAG00000056402
					eef1da	ENSSSAG00000068533
						ENSSSAG00000006077
						ENSSSAG00000005944
						ENSSSAG00000053970
						ENSSSAG00000062937
						ENSSSAG00000063001
						ENSSSAG00000077892
GO:0006412 - translation	42	0.040	42/278	328/3522	*rpl3l*	ENSSSAG00000051618
					rl18	ENSSSAG00000072884
					rpl39	ENSSSAG00000041178
					rps23	ENSSSAG00000046726
					rpl4	ENSSSAG00000010392
					rpl27a	ENSSSAG00000067773
					rps19	ENSSSAG00000075399
					rpl10a	ENSSSAG00000071589
					rpl9	ENSSSAG00000066273
					eef1b2	ENSSSAG00000007090
					rpl26	ENSSSAG00000047422
					rl17	ENSSSAG00000008596
					rpl19	ENSSSAG00000098459
					rl29	ENSSSAG00000076201
					rps13	ENSSSAG00000078842
					rpl10	ENSSSAG00000039434
					rps24	ENSSSAG00000006614
						ENSSSAG00000077735
						ENSSSAG00000024307
						ENSSSAG00000001684
						ENSSSAG00000046051

Term	Gene Count	Adjusted P-Value	Gene Ratio	Background Ratio	Homologous Gene Names (if present)	Contributing Genes (Ensembl ID)
					eef1g	ENSSSAG00000002154
					ef1a	ENSSSAG00000002140
					rpl11	ENSSSAG00000008448
					rpsa	ENSSSAG00000008668
					rpl35a	ENSSSAG00000046508
					zgc: 103559	ENSSSAG00000109892
					rpl35	ENSSSAG00000045463
					rpl13	ENSSSAG00000058507
					uba52	ENSSSAG00000032246
					ef1a	ENSSSAG00000051583
					acer1	ENSSSAG00000071292
					rpl31	ENSSSAG00000018086
					rpl37a	ENSSSAG00000066907
					rps8a	ENSSSAG00000054819
					rps11	ENSSSAG00000056402
					rpl36	ENSSSAG00000006077
					rps16	ENSSSAG00000005944
					rpl32	ENSSSAG00000053970
					cers5	ENSSSAG00000062937
					eef1da	ENSSSAG00000063001
						ENSSSAG00000077892

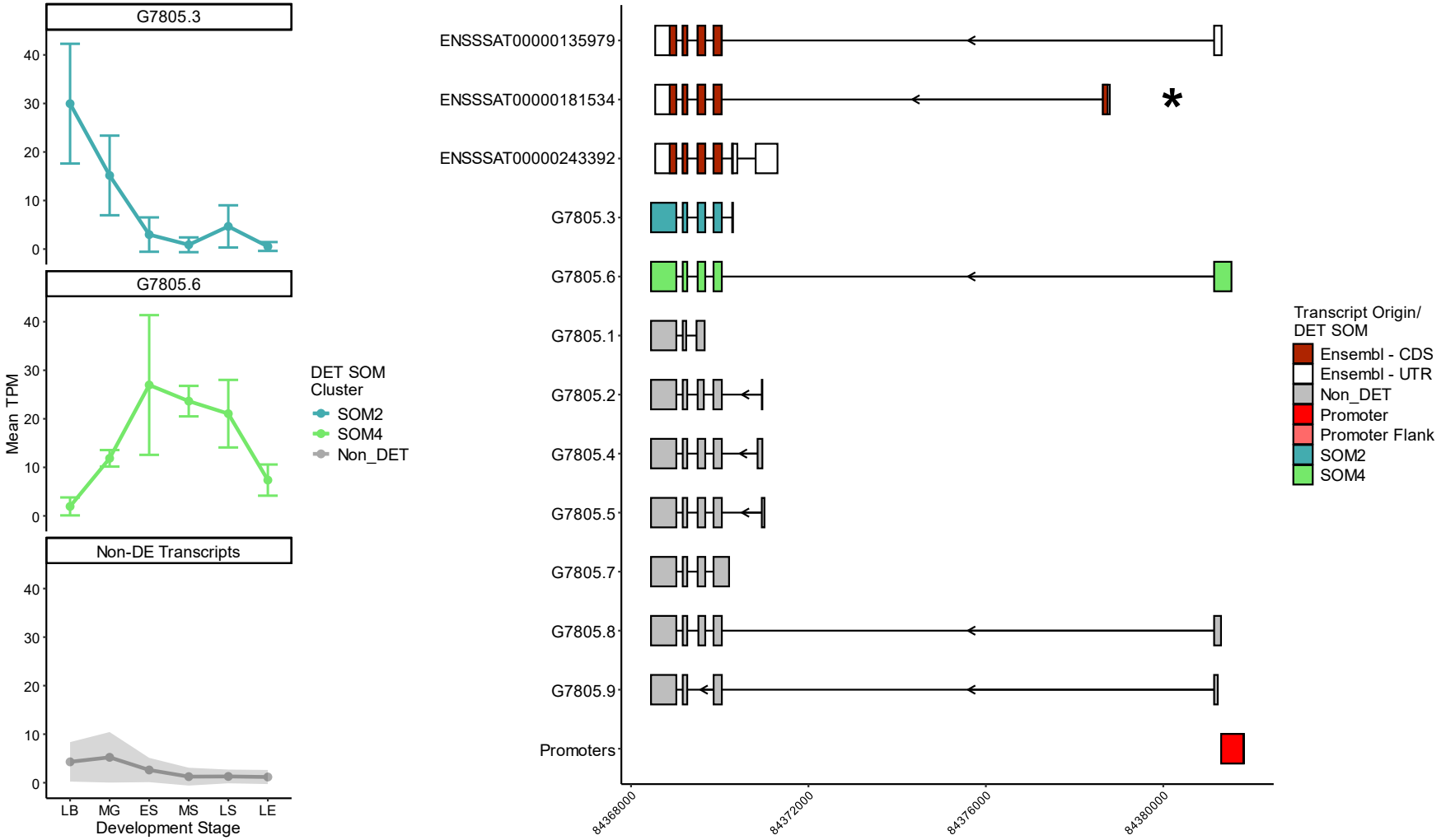


Figure 4.11: Visualisation of DETs (left) with matched transcript structures (right) for gene G7805 – tagl…(Legend continued on next page)

(Legend continued) Visualisation of DETs (left) with matched transcript structures (right) for gene G7805 – tagl.

Lineplots on left show TPMs (n=3) for all DETs over the 6 stages of development; late blastulation (LB), mid gastrulation (MG), early somitogenesis (ES), mid somitogenesis (MS), late somitogenesis (LS) and late-eyed (LE). Error bars are standard deviation and DETs are coloured to indicate DTU SOM cluster membership as per Figure 4.9. Average TPM for all replicates for all non-DETs is plotted in grey; ribbon shows standard deviation. On the right is a visualisation of transcript models for 1) the Ensembl reference annotation transcripts (dark red) for ENSSSAG00000074856, 2) DETs for G7805, coloured according to SOM cluster membership, 3) the non-DETs for gene G7805, and 4) promoter regions from the Ssal_v3.1 regulatory build (bright red). UTRs for Ensembl transcripts are displayed in white, while the canonical transcript is indicated by an asterisk.

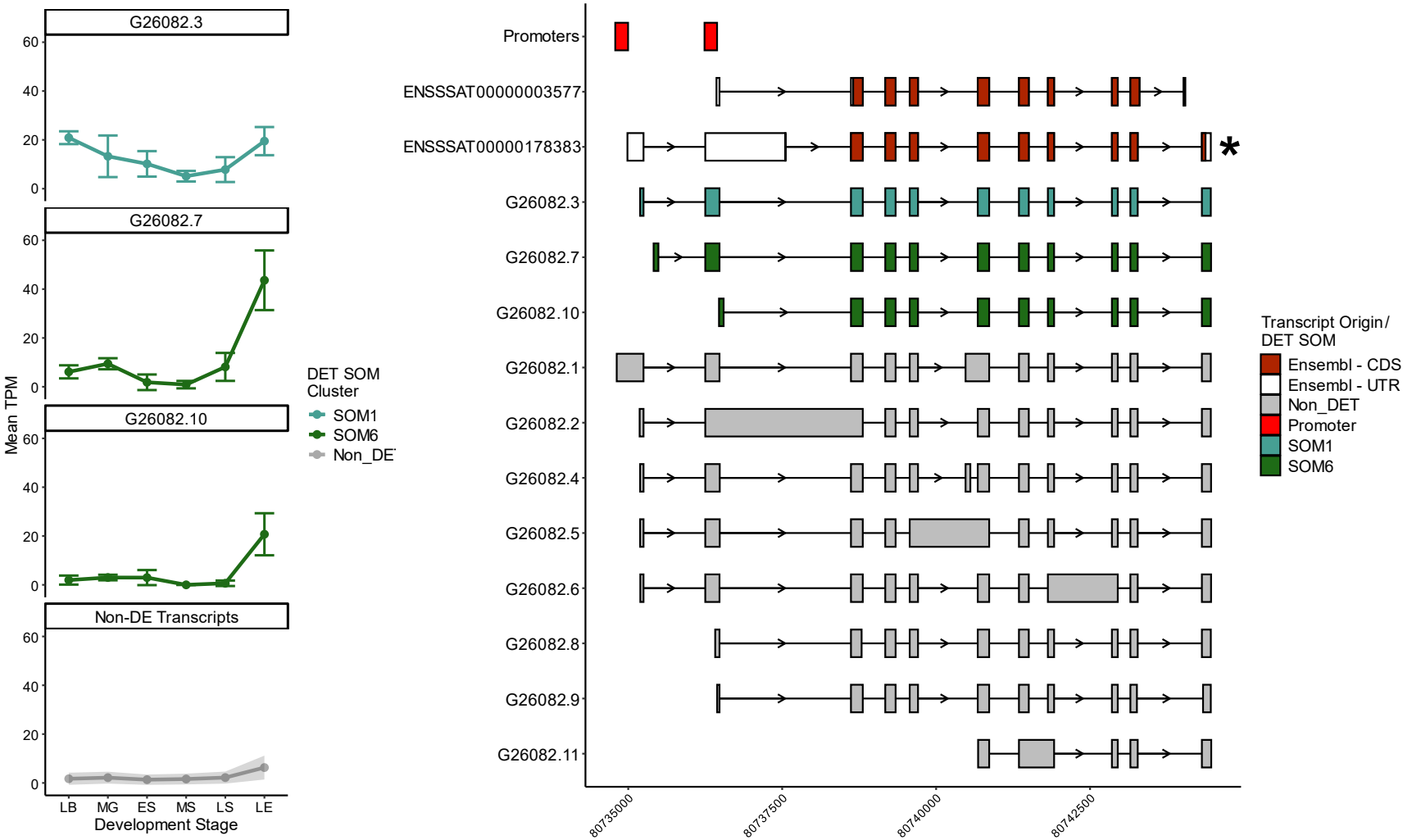


Figure 4.12: Visualisation of DETs (left) with matched transcript structures (right) for gene G26082 – rpl3l…(Legend continued on next page)

(Legend continued) Visualisation of DETs (left) with matched transcript structures (right) for gene G26082 – rpl3l.

Lineplots on left show TPMs (n=3) for all DETs over the 6 stages of development; late blastulation (LB), mid gastrulation (MG), early somitogenesis (ES), mid somitogenesis (MS), late somitogenesis (LS) and late-eyed (LE). Error bars are standard deviation and DETs are coloured to indicate DTU SOM cluster membership as per Figure 4.9. Average TPM for all replicates for all non-DETs is plotted in grey; ribbon shows standard deviation. On the right is a visualisation of transcript models for 1) the Ensembl reference annotation transcripts (dark red) for ENSSSAG00000001684, 2) DETs for G26082, coloured according to SOM cluster membership, 3) the non-DETs for gene G26082, and 4) promoter regions from the Ssal_v3.1 regulatory build (bright red). UTRs for Ensembl transcripts are displayed in white, while the canonical transcript is indicated by an asterisk.

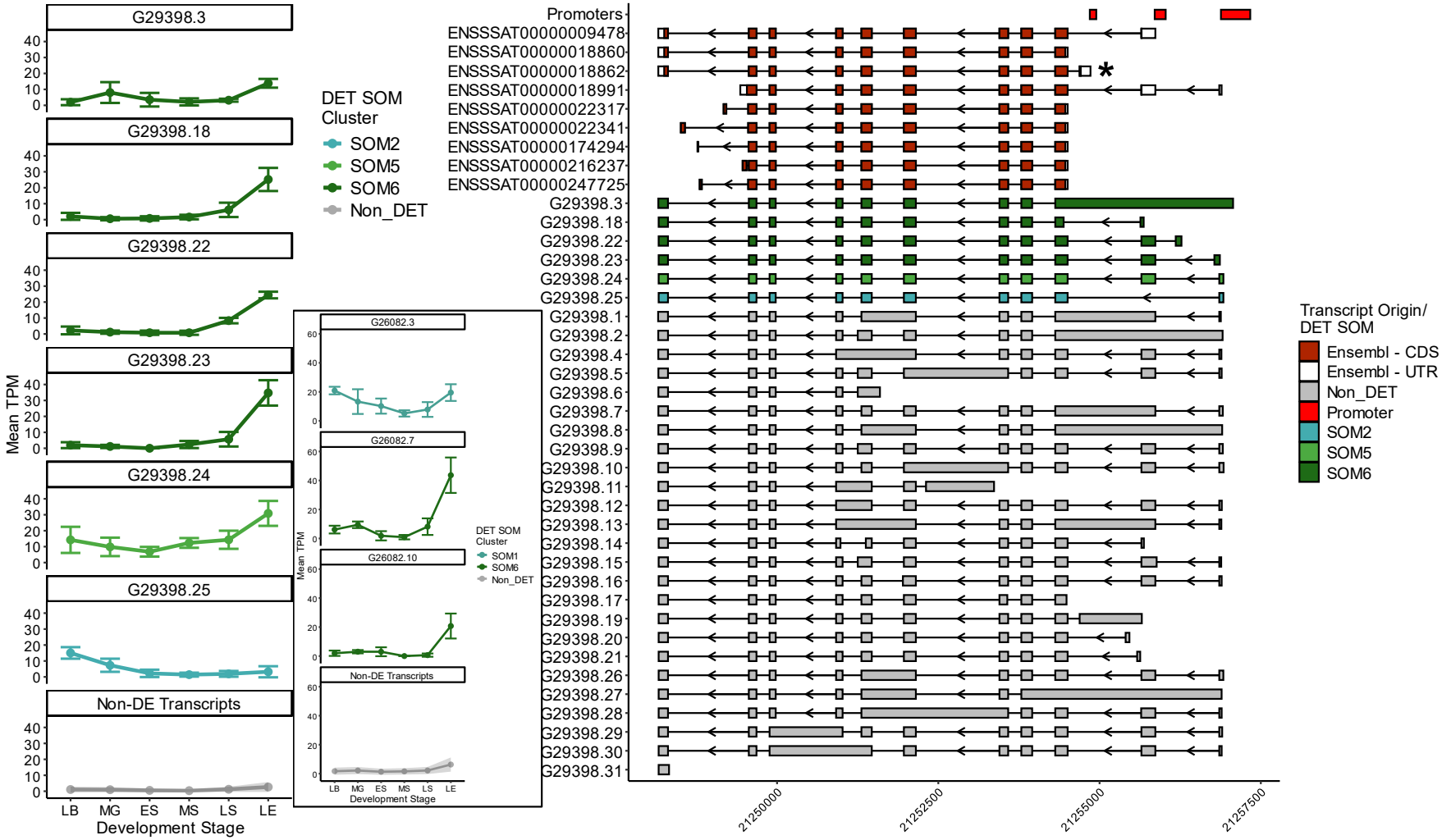


Figure 4.13: Visualisation of DETs (left) with matched transcript structures (right) for gene G29398 – rpl3l…(Legend continued on next page)

(Legend continued) Visualisation of DETs (left) with matched transcript structures (right) for gene G29398 – rpl3l. G29398 is the ohnologue pair of G26082 displayed in Figure 4.12 (collinear blocks 3q-6p; Lien et al., 2016)

Lineplots on left show TPMs (n=3) for all DETs over the 6 stages of development; late blastulation (LB), mid gastrulation (MG), early somitogenesis (ES), mid somitogenesis (MS), late somitogenesis (LS) and late-eyed (LE). Error bars are standard deviation and DETs are coloured to indicate DTU SOM cluster membership as per Figure 4.9. Average TPM for all replicates for all non-DETs is plotted in grey; ribbon shows standard deviation. Inlayed in the box is the TPM plot for G26082, the ohnologue of G29398, extracted from Figure 4.12. On the right is a visualisation of transcript models for 1) the Ensembl reference annotation transcripts (dark red) for ENSSSAG00000008448, 2) DETs for G29398, coloured according to SOM cluster membership, 3) the non-DETs for gene G29398, and 4) promoter regions from the Ssal_v3.1 regulatory build (bright red). UTRs for Ensembl transcripts are displayed in white, while the canonical transcript is indicated by an asterisk.

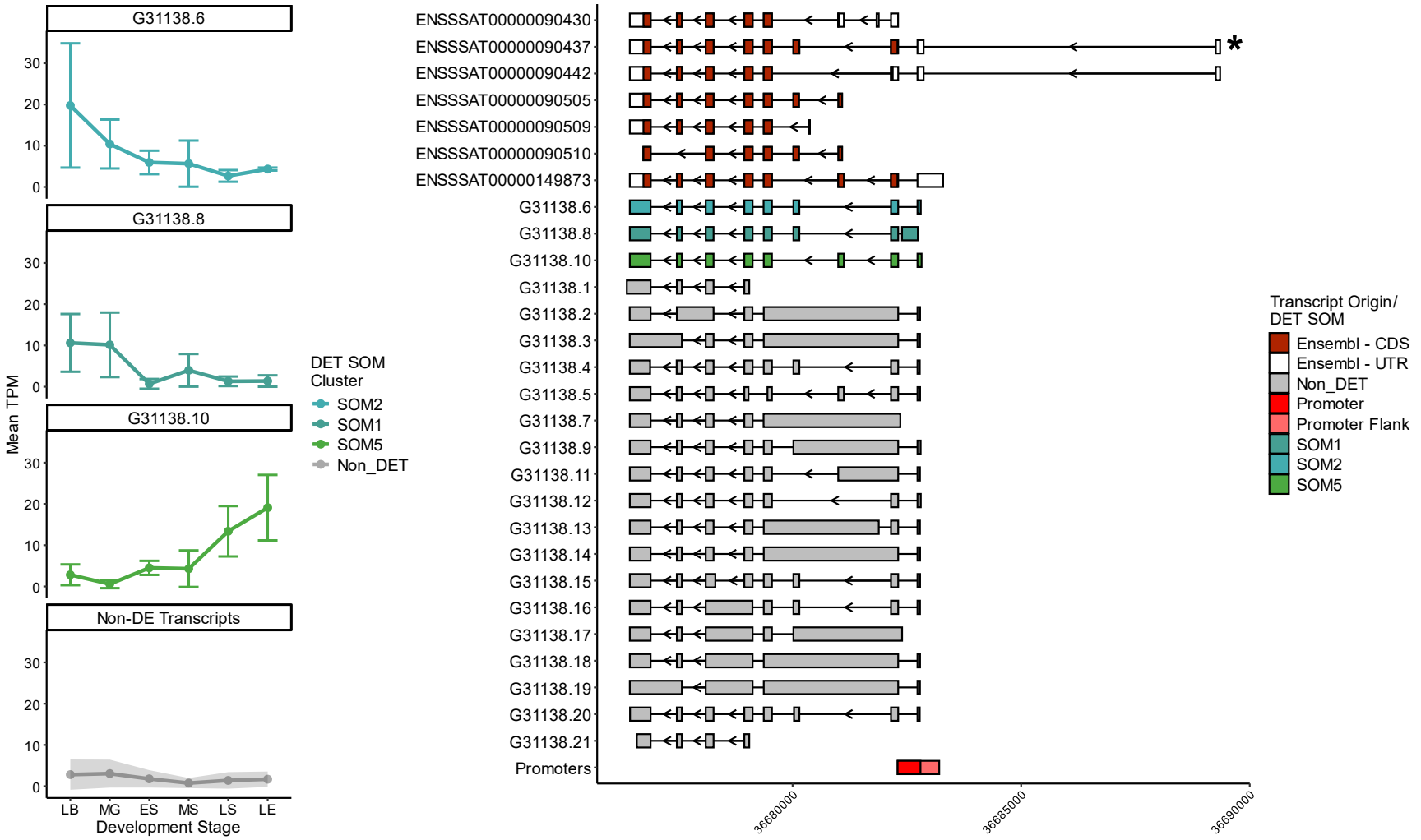


Figure 4.14: Visualisation of DETs (left) with matched transcript structures (right) for gene G31138 – slc25a3b…(Legend continued overleaf)

(Legend continued) Visualisation of DETs (left) with matched transcript structures (right) for gene G31138 – slc25a3b.

Lineplots on left show TPMs (n=3) for all DETs over the 6 stages of development; late blastulation (LB), mid gastrulation (MG), early somitogenesis (ES), mid somitogenesis (MS), late somitogenesis (LS) and late-eyed (LE). Error bars are standard deviation and DETs are coloured to indicate DTU SOM cluster membership as per Figure 4.9. Average TPM for all replicates for all non-DETs is plotted in grey; ribbon shows standard deviation. On the right is a visualisation of transcript models for 1) the Ensembl reference annotation transcripts (dark red) for ENSSSAG00000055931, 2) DETs for G31138, coloured according to SOM cluster membership, 3) the non-DETs for gene G31138, and 4) promoter regions from the Ssal_v3.1 regulatory build (bright red). UTRs for Ensembl transcripts are displayed in white, while the canonical transcript is indicated by an asterisk.

Chapter 5: Characterisation of Mono-Exonic Transcript Models in the Atlantic Salmon Genome

Summary

A high prevalence of novel mono-exonic transcript models was observed in the Atlantic salmon long-read transcriptome generated in Chapter 2. Often considered transcriptional noise or sequencing artifacts, mono-exonic transcripts are commonly excluded from transcriptomic analysis. However, recent long-read RNA-seq studies have captured mono-exonic transcripts with greater fidelity and highlighted their potential biological relevance in eukaryotes. In this chapter, I characterise 15,072 mono-exonic transcripts described in Chapter 2, most of which were not annotated by Ensembl. Approximately a third have putative protein-coding potential and some may represent enhancer RNAs. Additionally, a novel family of highly repetitive, expressed retrogene elements was discovered.

5.1 Introduction

Historically, most transcriptome research has focussed on protein-coding mRNAs produced by alternative splicing of multi-exonic RNA. However, many classes of RNA lack coding potential and instead have potential regulatory functions, including long non-coding RNA (lncRNA), microRNA (miRNA), and circular RNAs (Wang et al., 2011; Lu & Rothenberg, 2018; Nemeth et al., 2024), among diverse others. Such discoveries raise many questions on the function of non-coding regions, which make up the majority of eukaryotic genomes (Mattick, 2001; Alexander et al., 2010). Mono-exonic transcripts, unlike typical mRNAs, lack introns and are often produced by genes solely producing mono-exonic transcripts (Aviña-Padilla et al., 2021). They are captured in small numbers by short-read RNA-seq and often considered transcriptional noise or artifacts; thus, they are often filtered from short-read datasets (e.g. Pertea et al., 2018; Torre et al., 2023).

Recent studies employing long-read RNA-seq have revealed extensive unannotated mono-exonic transcript diversity with potential impacts on gene expression and regulation (Kuo et al., 2020). Using PacBio RNA-seq, Kuo et

al. (2020) identified over 2,000 novel mono-exonic lncRNAs in humans, whilst approximately 9,000 non-coding mono-exonic genes were identified in *Arabidopsis thaliana* (Zhang et al., 2022). Nanopore RNA-seq identified 22,934 mono-exonic transcripts in Hereford cattle, *Bos taurus*, predominantly expressed in brain, with just 5% having putative coding potential (Halstead et al., 2021). PacBio sequencing in chicken captured 14,831 non-coding and 5,533 coding mono-exonic transcripts (Kuo et al., 2017). These studies suggest that mono-exonic transcripts are widespread, suggesting their biological relevance, and are readily captured by long-read RNA-seq. Despite this, many long-read RNA-seq studies retain the practice of filtering them as transcriptional noise (Su et al., 2024).

Similar to other long-read RNA-seq studies, I identified 15,072 mono-exonic transcripts in my Atlantic salmon transcriptome (Chapter 2). In this chapter, after comparing them with mono-exonic transcripts annotated by Ensembl, I investigated their genomic and functional properties and established their association with repetitive regions and active regulatory elements, to better understand the potential biological importance and evolutionary history of these enigmatic transcripts.

5.2 Materials and Methods

5.2.1 Extraction of Mono-Exonic Transcripts

Part of the classification output file generated by SQANTI3 (reported in Chapter 2) (Tardaguila et al, 2018; Pardo-Palacios et al., 2024) assigned a tag to each transcript model indicating whether it was multi-exonic or mono-exonic. A custom "`awk`" script was used to subset the mono-exonic transcript models from the long-read transcriptome assembly .bed file based on this tag. Then, sequences of all mono-exonic transcripts were extracted by first converting the subsetted .bed file to .gtf format with the TAMA GO format converter (Kuo et al., 2020) and then extracting the sequences from the Atlantic salmon Ensembl annotation (Ssal_v3.1) based on genomic coordinates with the "`getfasta`" function from the BEDtools suite v2.30.0 (Quinlan & Hall, 2010).

In Chapter 2, the TAMA suite was used to create a file detailing the number of reads supporting each transcript model in the long-read transcriptome. This was subsetted based on the transcript model ID to form a read-support file for the mono-exonic subset. Finally, transcript model lengths were imported into RStudio (R4.3.3) by extracting column 11 of the mono-exonic .bed file. A histogram of mono-exonic transcript lengths was generated with ggplot2 v3.5.0 (Wickham, 2016).

5.2.2 Clustering Mono-Exonic Subset

All mono-exonic transcript sequences were processed using CD-HIT v4.8.1 (Li & Godzik, 2006; Fu et al., 2012) with options "`-c 0.95 -n 10 -d 0 -M 16000 -T 2`" to cluster them into groups based on similarity. A sequence identity threshold of 95% (`-c 0.95`) was used. Sequence identity is calculated by totalling the number of identical nucleotides and dividing them by the length of the shortest sequence in the comparison. For CD-HIT clusters containing >10 members, I extracted chromosome coordinates from the transcriptome annotation files using a custom "`awk`" script.

5.2.3 Overlap with Ensembl Reference Mono-Exonic Models

To compare mono-exonic transcript models from the long-read transcriptome with those annotated by Ensembl, I first extracted the reference models using a custom bash script and then converted the reference subset to .bed format with the "`gtf2bed`" function in the BEDOPS suite v2.4.41 (Neph et al., 2012). BEDtools "`intersect`" was used to run a reciprocal overlap between the long-read and Ensembl reference mono-exonic models with these options: "`-wo -f 0.7 -r -a`". The "`-f`" option sets how much of the query model needs to overlap the reference model to be considered a true overlap (as a decimal percentage), whilst the "`-r`" option denotes this overlap needs to be reciprocal (two way), i.e. in this case, 70% of the length of the long-read transcript model needs to overlap the reference model AND 70% of the reference model needs to overlap the long-read query model. Reciprocal overlaps of 50%, 20%, 10%, and any overlap (1bp) were also tested for comparison.

5.2.4 Predicting Protein-Coding Function

TransDecoder v5.7.1 (Haas, 2023) was used to check the mono-exonic models for protein-coding potential. First, open-reading frames (ORFs) were predicted from the mono-exonic sequences using "`TransDecoder.LongOrfs`" with a minimum peptide length cut-off of 80 amino acids (aa). TransDecoder categorises ORFs into 4 subcategories; "complete" if ORFs contains both a start and stop codon, "5' partial" if no start codon is detected but ORFs contains a stop codon, "3' partial" if a start codon is present but no stop codon is identified, and "incomplete" if the predicted ORFs lacks both start and stop codons. The number of ORFs belonging to each subcategory was imported into RStudio and plotted as a barchart with ggplot2.

Secondly, all predicted ORFs were tested for homology to known proteins in the combined database of proteins extracted from the Atlantic salmon, rainbow trout, zebrafish, mouse and human UniProt annotations generated in Chapter 3, section 3.2.4 using DIAMOND v2.1.9 (Buchfink et al., 2021). The options "`--evalue 1e-5 --max-target-seqs 1`" were used to retain only the top hit for each mono-exonic model. Finally, "`TransDecoder.Predict`" with options "`--single_best_only`" was used to filter hits down to the single best match per mono-exonic transcript model.

I finally extracted and counted the length of aa sequences from both the Ensembl reference and my long-read transcriptome mono-exonic models. This data was imported into RStudio and a histogram plotted with ggplot2 for comparison.

5.2.5 Overlapping Repeat Genomic Regions

Mono-exonic transcript models were overlapped with repeat regions masked by Ensembl in the Ssal_v3.1 reference genome to determine if any were repetitive elements. I generated a .bed file including coordinates of the masked regions using a custom "`awk`" script. A one-way BEDtools intersect was used to overlap mono-exonic transcripts and masked regions with

option: "`-wo`", outputting the length of each mono-exonic model overlapping a repeat region. The output intersect .bed file was imported to RStudio.

There were multiple instances of masked repeats existing in close proximity to other repeat regions with few bases separating them. In some cases, mono-exonic transcript models spanned these gaps, creating intersections spanning consecutive repeat regions. In these instances, I summed the lengths of all individual intersections for each transcript ID in R and then calculated the percentage overlap of the total transcript model length. The percentage overlap was then plotted in a barplot with ggplot2.

5.2.6 Overlapping with Known Enhancer Regions

The AQUA-FAANG project produced a comprehensive list of putative enhancer regions in the Atlantic salmon genome, made publicly available through the Ensembl Ssal_v3.1 regulatory build (Johnston et al. 2024). Enhancer RNAs (eRNA) are non-coding RNAs transcribed from enhancer regions, with hotly debated functional relevance (Natoli & Andrau, 2012). To identify candidate eRNAs among my dataset, I overlapped the mono-exonic transcripts with the Ensembl predicted enhancers. First, I downloaded the regulatory build from Ensembl and converted it to .bed format with the "`gff2bed`" function in the BEDOPS suite v2.4.41 (Neph et al., 2012). I then specifically extracted enhancer coordinates with a custom "`awk`" script and used BEDtools intersect, employing a non-reciprocal overlap of 50% to identify mono-exonic model showing ≥50% overlaps with these regions. Finally, bed files of mono-exonic models overlapping Ensembl enhancers were imported into RStudio, alongside a gff3 file containing the Ensembl regulatory build, and structures plotted using ggtranscript v0.99.9 (Gustavsson et al., 2022).

5.2.7 Identification of Retrogenes

Retrogenes originate from the re-insertion of retrotransposed mature mRNAs into the genome (Kaessmann et al., 2009). The result is a mono-exonic copy of an existing gene at a different locus (Figure 5.1). Retrogenes initially possess a genomic polyA signature immediately following their transcription

termination site (TTS; Casola & Betrán, 2017). To search for expressed retrogenes in the mono-exonic transcripts, these sequences were aligned to the Ensembl transcriptome annotation using the "`blastn`" function within the BLAST+ suite v2.15.0 (Camacho et al., 2009), applying a percentage identity filter ≥95% ("`-perc_identity 95`") to retain only transcripts with highly similar sequences. BLAST results were further filtered to only retain mono-exonic transcript queries and Ensembl transcripts showing reciprocal coverage ≥95% of their respective lengths. The rationale is that recently inserted retrogenes should initially comprise the full length of their expressed parent transcript. A further step was conducted to remove hits associated with a known gene in the Ensembl transcriptome, which presumably represent novel transcripts from existing genes, rather than retrogenes, which should have distinct coordinates in the genome.

One of the outputs of the SQANTI3 classification file generated in Chapter 2 is a 20bp sequence following the TTS. To determine how many of the mono-exonic transcripts showed evidence of a (genomic) polyA tail, I used a custom bash script to extract the first 12bp downstream of the TTS from the SQANTI3 classification. Mono-exonic models where ≥10bp of their post-TTS genomic sequence represented a string of either A or T bases were considered to show evidence of a genomic polyA tail.

Following filtering, the remaining mono-exonic models were sorted into two retrogene families, each containing all the mono-exonic transcripts associated with a set of paralogous parent genes. Then, for each retrogene family, Ensembl IDs of putative parent transcripts and their genes of origin for the filtered mono-exonic models containing a genomic polyA tail (i.e. candidate Atlantic salmon retrogenes) were extracted from the BLAST results. BioMart (Smedley et al., 2009) was then used (Ensembl release 112) to extract cDNA sequences for all Atlantic salmon transcripts derived from candidate parent genes and their paralogues predicted by Compara (Herrero et al. 2016). BioMart was further used to extract cDNA sequences for all brown trout (*Salmo trutta*) Compara-predicted orthologues for the candidate parent genes and their paralogues. Sequences for the Atlantic salmon mono-exonic models, their parent genes and predicted paralogues, and their

predicted orthologues from brown trout, were aligned using MAFFT version 7 (Katoh et al., 2019) with command "`mafft --thread 8 --threadtb 5 --threadit 0 --reorder --adjustdirection --auto input > output`". The resultant alignment was trimmed to retain only blocks of conserved alignment between the mono-exonic transcript sequences and the other aligned sequences. The alignment was submitted to IQTREE (Nguyen et al. 2015; Trifinopoulos et al., 2016) to generate a maximum-likelihood phylogenetic tree using the best-fitting substitution model estimated by ModelFinder (Kalyaanamoorthy et al. 2017). Branch support was determined using the ultrafast bootstrap algorithm (Minh et al., 2013). The phylogenetic tree in Newick format was imported into FigTree v1.4.3 (Rambaut, 2016: http://tree.bio.ed.ac.uk/software/figtree/) for viewing and annotation.

5.3 Results

The 15,092 mono-exonic transcript models derived from 11,601 unique genes and had a mean length of 1,350bp and N50 of 2,319bp (Figure 5.2). Many derived from genes that produced multi-exonic transcripts (see Figure 5.3). Specifically, 5,828 genes produced only mono-exonic transcripts and of these, 5,151 produced a single transcript for a total of 6,722 transcripts. Thus, just over half of the mono-exonic transcript models derived from genes that also produced multi-exonic transcripts (55%; 8,350/15,072).

11,586 mono-exonic models (derived from 11,127 unique genes) are currently annotated by Ensembl. 10,589 (95%) of these genes only produce mono-exonic transcripts. Comparing my long-read mono-exonic transcripts with the Ensembl set revealed that 505 overlapped with ≥50% reciprocal overlap, whilst 1,681 showed any overlap (min. 1bp). Therefore, much of the mono-exonic transcripts captured by Nanopore long-read RNA-seq is currently unannotated.

5.3.1 Coding Potential of Mono-Exonic Transcripts

Approximately 35% of the mono-exonic transcript models (5,349/15,072) contained an ORF sequence ≥80aa, indicative of protein-coding function (Figure 5.4). Of these, 68.5% were deemed complete ORFs containing both a start and stop codon. The next most common category was 5' partial.

Around 16% (1,845/11,586) of the Ensembl mono-exonic transcripts are classified as coding. However, only 286 of the 5,349 protein-coding mono-exonic models in my transcriptome overlapped the Ensembl set.

Predicted amino acid sequences from the long-read mono-exonic models displayed a similar length distribution to the Ensembl set (Figure 5.5). Of note, 78 of the Ensembl predicted proteins were smaller than 80aa in length, the cut-off I deployed for the long-read mono-exonic ORFs.

5.3.2 Clustering of Mono-Exonic Models

CD-HIT clustered the long-read mono-exonic transcript models into 11,298 unique clusters with members sharing >95% similarity. Only 36 clusters had ≥10 members, representing 986 mono-exonic transcript models, indicating that much of the captured mono-exonic transcript repertoire represents unique sequence (Figure 5.6).

Determining the genomic coordinates for transcripts in the 36 clusters revealed some interesting patterns. In several cases, e.g. CD-HIT cluster 29, all members were mono-exonic transcripts originating from the same gene. At the other end of the spectrum, CD-HIT cluster 157 contained 60 transcript models derived from 50 loci spanning 31 chromosomes and unplaced scaffolds (Table 5.1).

5.3.3 Mono-Exonic Transcripts Overlap Repetitive Genomic Regions

3,814 of the 15,092 total mono-exonic transcripts had at least 50% of their length overlapping a repeat region in the Atlantic salmon genome (Figure 5.7). 814 had their entire length contained within a repeat region. 264 transcripts showing 90% or greater overlap with repeat regions were found within the 26 CD-HIT clusters containing 10 or more members, consistent with them representing distinct groups of expressed repeats sharing high similarity.

5.3.4 Identification of Candidate eRNAs

164,105 putative enhancer regions are annotated in the Ensembl regulatory build. 696 mono-exonic transcripts overlapped at least 50% of the length of

an enhancer region. 128 models showed reciprocal overlap ≥50% with an enhancer and are strong candidates to represent eRNAs.

For example, gene G971 produces 4 mono-exonic transcripts of this type (Figure 5.8), all showing support from reads deriving solely from embryo samples. G971.1 is highly supported by 54 reads whilst G971.18, G971.21 and G971.55 are supported by 4 to 6 reads each. Another highly supported mono-exonic eRNA is transcript G3837.1, with 66 reads supporting it from both embryo and head kidney datasets (Figure 5.9). Most of the 128 candidate eRNA mono-exonic transcripts were well-supported, with 109 (85%) and 63 (49%) having read support of ≥5 and 10, respectively. High read support was also observed for the 696 non-reciprocally overlapped mono-exonic transcripts with 553 (79%) and 332 (48%) having read support of ≥5 and 10, respectively.

5.3.5 Salmonid-Specific Retrogene Family Identified in Mono-Exonic Transcripts

37 mono-exonic transcript models showed BLASTn hits with ≥95% identity and ≥95% coverage against multi-exonic Ensembl reference transcripts and were found in different genomic locations. 28 out of 37 possessed a putative polyA tail (e.g. Figure 5.10) indicative of a retrogene, all derived from CD-HIT cluster 157 (Table 5.1). These 28 transcripts had BLAST hits against 4 unique Ensembl transcripts, each derived from a unique gene (ENSSSAT00000105940 from ENSSSAG00000104719 (16 hits); ENSSSAT00000022212 from ENSSSAG00000010098 (1 hit); ENSSSAT00000098881 from ENSSSAG00000086375 (1 hit); ENSSSAT00000174305 from ENSSSAG00000089398 (11 hits).

These candidate retrogene parent genes are closely related paralogues according to Ensembl Compara, both with each other, in addition to other genes, comprising a family of 27 unique Ensembl genes. Furthermore, these genes were predicted by Ensembl to be orthologous to a gene family with 16 members in the brown trout genome, each with a single mono-exonic transcript. A single orthologous Ensembl gene exists in the rainbow trout

genome, with no other predicted paralogues, consistent with a retrogene family expansion specific to *Salmo*.

A maximum likelihood phylogenetic analysis was used to investigate the evolution of the 28 mono-exonic transcripts representing candidate retrogenes. This was done in a framework containing the homologous gene families annotated in the Ensembl Atlantic salmon and brown trout genomes. Most of the 28 sequences branched closely to each other, in a polytomy excluding most other Atlantic salmon and all brown trout sequences, implying they arose very recently (Figure 5.11). In addition, most of the brown trout orthologues form a monophyletic group, which supports that the expansion of this gene family was largely species-specific. However, two brown trout orthologues branch with Atlantic salmon sequences, suggesting distinct retrogene expansions have occurred from genes present in the common ancestor of *Salmo*.

Finally, these 28 retrogenes were expressed from 21 unique genomic locations. 27 completely overlap masked regions, whilst G33119.1 overlaps a repeat region with 98% of its length. Overall, the data indicates long-read-RNA-seq has captured a highly expressed repetitive element widely distributed throughout the Atlantic salmon genome.

5.4 Discussion

Mono-exonic transcripts are an underexplored feature of transcriptomes gaining recognition for their potential roles in gene expression and regulation. Captured with greater frequency by long-read RNA-seq, this chapter supports and expands on work conducted in other species (Kuo et al., 2017; Kuo et al, 2020; Halstead et al., 2021) utilising a variety of approaches to elucidate mono-exonic transcript characteristics.

5.4.1 Identification of Retrogene Families

Retrogenes were historically thought to be non-functional copies of other similar genes. However, an increasing body of research has recognised the potential roles of retrocopies for regulating gene expression (Kubiak & Makałowska, 2017; Cheetham et al., 2020; Ciomborowska-Basheer et al.,

2021; Elbarbary et al., 2016). Retrogenes are often pseudogenised (Salmena, 2021), but in some instances they remain active and elicit novel protein functions, called neofunctionalisation (Kubiak & Makałowska, 2017). In some cases where retrogenes retain activity, protein-coding function is lost and instead these genes may assume regulatory roles; for example, lncRNA deriving from retrogenes can bind to mRNA of the parent gene to inhibit translation (Bryzghalov et al., 2016) whilst sense retrogenes can form truncated coding sequences that are translated into proteins with biological activity (Dennis et al., 2012).

Retrotransposition was first observed in salmonids in the 1980s (Matsumoto et al., 1986) and a diverse repertoire of retrotransposons have been characterised since (Matveev & Okada, 2009). However, little research has been conducted on salmonid retrogenes in the post-genomic era. A significant discovery in this chapter was the identification of mono-exonic retrogene families including a common repeat element expressed in 21 different loci throughout the genome. This demonstrates the ability of my long-read RNA-seq approach for identifying retrogene diversity in complex genomic regions. Improved annotation of retrogene elements in Atlantic salmon is a first step towards understanding their potential functional roles in salmonids.

5.4.2 Mono-Exonic Transcripts Derived from Multi-Exonic Genes

5.7% of the long-read transcriptome consists of mono-exonic transcripts compared with 6.2% in the Ensembl reference. Despite this high-level similarity, ~55% of the long-read mono-exonic transcripts originated from genes that also produce multi-exonic transcripts, which was ten-fold less common in the Ensembl annotation. It seems that these mono-exonic transcripts originating from multi-exonic genes also explained the two-fold higher proportion of predicted coding mono-exonic transcripts in the long-read annotation. A recent short-read RNA-seq study found limited evidence for protein-coding mono-exonic transcripts derived from multi-exonic genes (Aviña-Padilla et al., 2021). This can be attributed to challenges resolving exon-chaining and intron-retention events using short-read data (Dong et al.,

2021). Recent long-read RNA-seq studies also support my observations, though the majority of mono-exonic transcripts have previously been annotated as non-coding (Kuo et al., 2020; Zhang et al., 2022). In fact, it is difficult to find information about mono-exonic transcripts deriving from multi-exonic genes in the literature. Protein-coding mono-exonic transcripts were identified in *Bos taurus* with high prevalence in brain tissue (Halstead et al., 2021), whereas maternal genes captured in advance of zygotic genome activation in *Drosophila* embryos either contained intronic regions or were purely mono-exonic (Riemondy et al., 2023). This chapter highlights that many mono-exonic transcripts transcribed from multi-exonic genes retain protein-coding potential, with more work needed to identify and elucidate their functions.

5.4.3 Long-Read RNA-Seq Resolves Expression of Repetitive Loci

Long-read RNA-seq is effective at resolving stretches of repetitive genomic DNA during genome assembly (Kono & Arakawa, 2019) and, at the transcriptional level, can distinguish between highly-similar repetitive RNA molecules (Reggiardo et al., 2023). In this chapter, I successfully identified many mono-exonic transcripts overlapping repeat regions. It is often postulated that expressed repetitive elements constitute transcriptional noise or so-called ‘junk’ RNA (discussed in Lee et al., 2019). Indeed, it is possible that many of the long-read mono-exonic transcripts lack function and fit such definitions. Currently, more investigation is required to fully understand the repetitive transcripts in all species, not just salmonids. Given that the Atlantic salmon genome has one of the highest levels of repeat content in eukaryotes (Lien et al., 2016), the ability of long-read RNA-seq to better resolve RNA expression from repetitive stretches of the genome will allow improved characterisation of expressed repetitive RNAs in salmonids.

5.4.4. Enhancer RNA Expression

They study of eRNAs is an emerging field within transcriptomics, broadly implicated in human disease (Ren et al., 2017). Expression of eRNAs is correlated with the induction of mRNA transcription of neighbouring genes (Arner et al., 2015; Andersson et al., 2014) and overexpression of eRNA

leads to an increase in expression of its target gene (Jiao et al., 2018). In addition, eRNAs can influence enhancer function, which in turn affects the expression of enhancer target genes (Ntini & Marsico, 2019). However, it is yet unknown if these regulatory effects are common to all eRNAs and consequently there is a focus on identifying which eRNAs have impact, and to what extent, on mRNA expression (Arnold et al., 2020). My long-read method was able to identify candidate eRNAs in the Atlantic salmon genome. Twinned with other functional genomic tools aimed at describing regulatory elements including ChIP-seq, this method, with sufficient sequencing depth and sample representation, has potential to improve understanding of full-length eRNA structures and potentially lead to better understanding of the mechanisms by which cis-regulatory elements drive gene expression (Panigrahi & O'Malley, 2021).

5.4.5 Potential for DNA Contamination

Mono-exonic transcript models are often considered transcriptional noise or artifacts and subsequently filtered from RNA-seq datasets (Torre et al., 2023). In this chapter, I have described a potential myriad of characteristics and potential functions of mono-exonic transcripts including enhancer RNA and retrogenes. However, some additional factors should be taken into account when interpreting these results. Firstly, there is a possibility that the mono-exonic transcript set may include signatures derived from genomic DNA contamination as no DNase treatment was used during RNA extraction from the embryo samples. Whilst polyA selection was used to capture the mRNA and would in theory reduce genomic DNA prevalence, it is possible that poly(dT) priming to genomic runs of polyA could still occur (Li et al., 2015) thereby producing apparently mono-exonic transcript models. It is possible that these artifacts could be filtered out of the dataset in future analyses by examining the genome sequence downstream of the TTS, however, applying a strict filter on this feature this may lead to the reduction in capture of elements of interest like retrogenes. Taking this into account, we might expect genomic contamination to be of greater length than true mono-exonic transcripts. However, comparing the mean lengths of the 15,072 mono-exonic transcripts with the multi-exonic transcript models

reveals that the mono-exonic transcript are substantially shorter than the multi-exonic models (1,349.76bp vs 2,298.95bp), providing minimal evidence for genomic DNA contamination. To mediate for genomic DNA contamination, DNase treatment could have been applied to the embryo samples during RNA extraction to eliminate possible DNA contamination

Finally, due to the possibility of genomic DNA contamination, it is possible that a portion of the identified retrogenes may not be truly transcribed and are instead artifacts captured through intra-polyA priming. Retrogene insertion occurs randomly (Kaessmann et al., 2009) and thus the retrogene may not be inserted into a region downstream of a promoter. Transcription of genes without associated promoters is rare, which makes an expressed retrogene an even rarer occurrence.

5.4.6. Concluding Thoughts

In this chapter, I successfully investigated the origin and characteristics of mono-exonic transcripts, an often-overlooked component of the transcriptome that remains poorly understood in salmonids. The description of 15,072 mono-exonic transcripts, most unannotated by Ensembl, offers a significant advance to functional annotation of the Atlantic salmon genome. Additionally, the methods developed in this study provide useful resources for further research into mono-exonic transcripts.

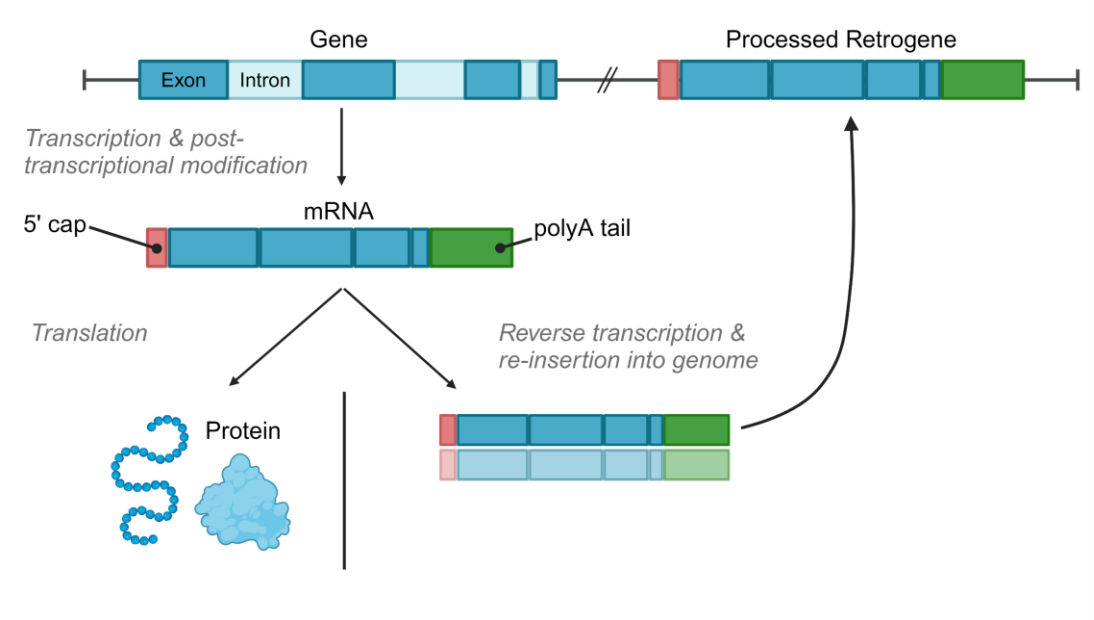


Figure 5.1: Retrogene formation process. Genes are transcribed into mRNAs which are usually translated into proteins. A retrogene is formed when a mature mRNA is reverse-transcribed and re-inserted into the genome in a new locus forming a mono-exonic copy of the original multi-exonic gene with a stretch of genomic polyA at the 3' end. Created with BioRender.com, adapted from Oshima (2013).

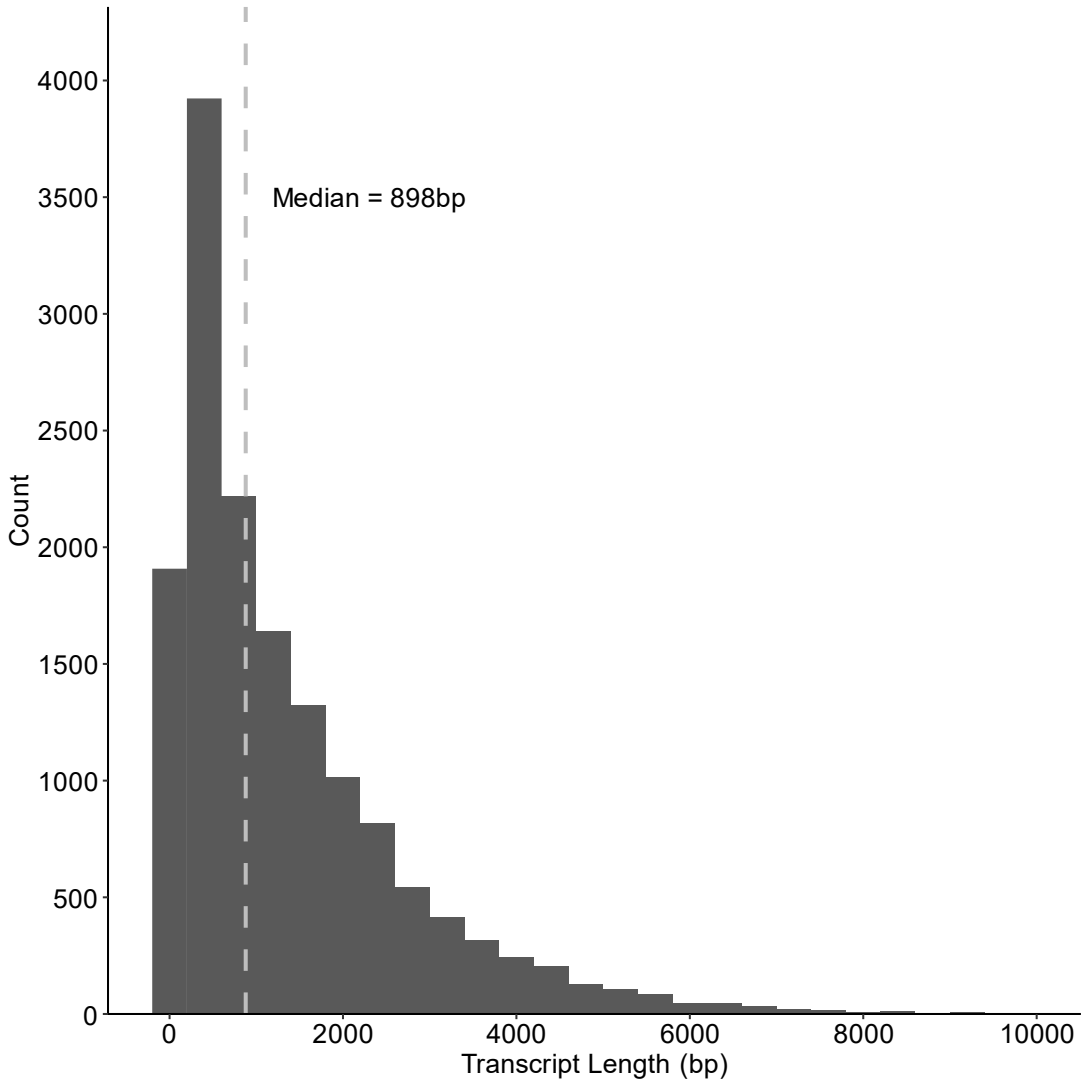


Figure 5.2: Histogram showing length distribution of the 15,072 mono-exonic transcript models defined in my long-read transcriptome.

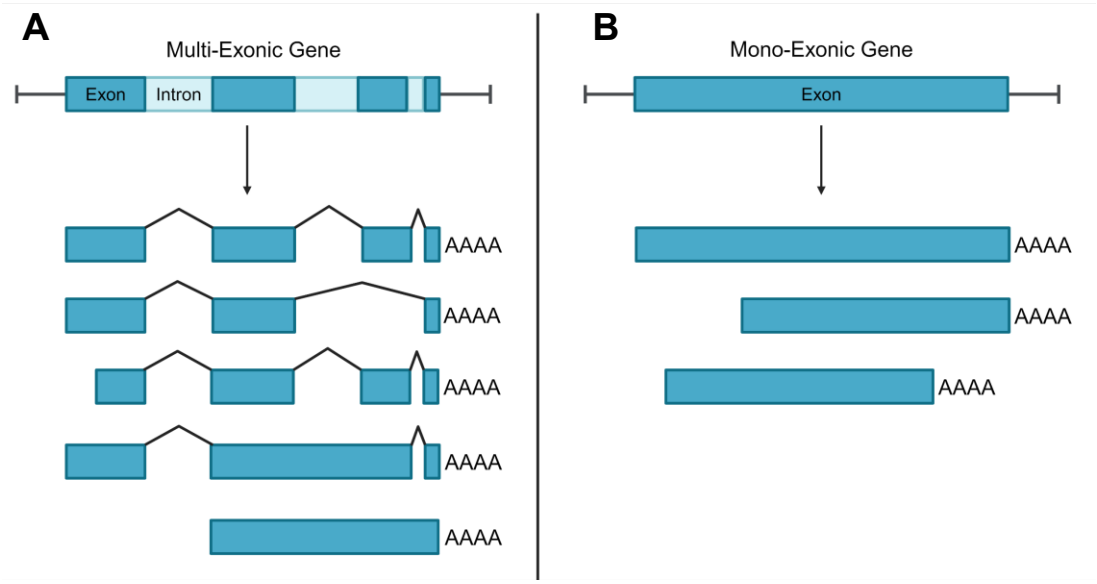
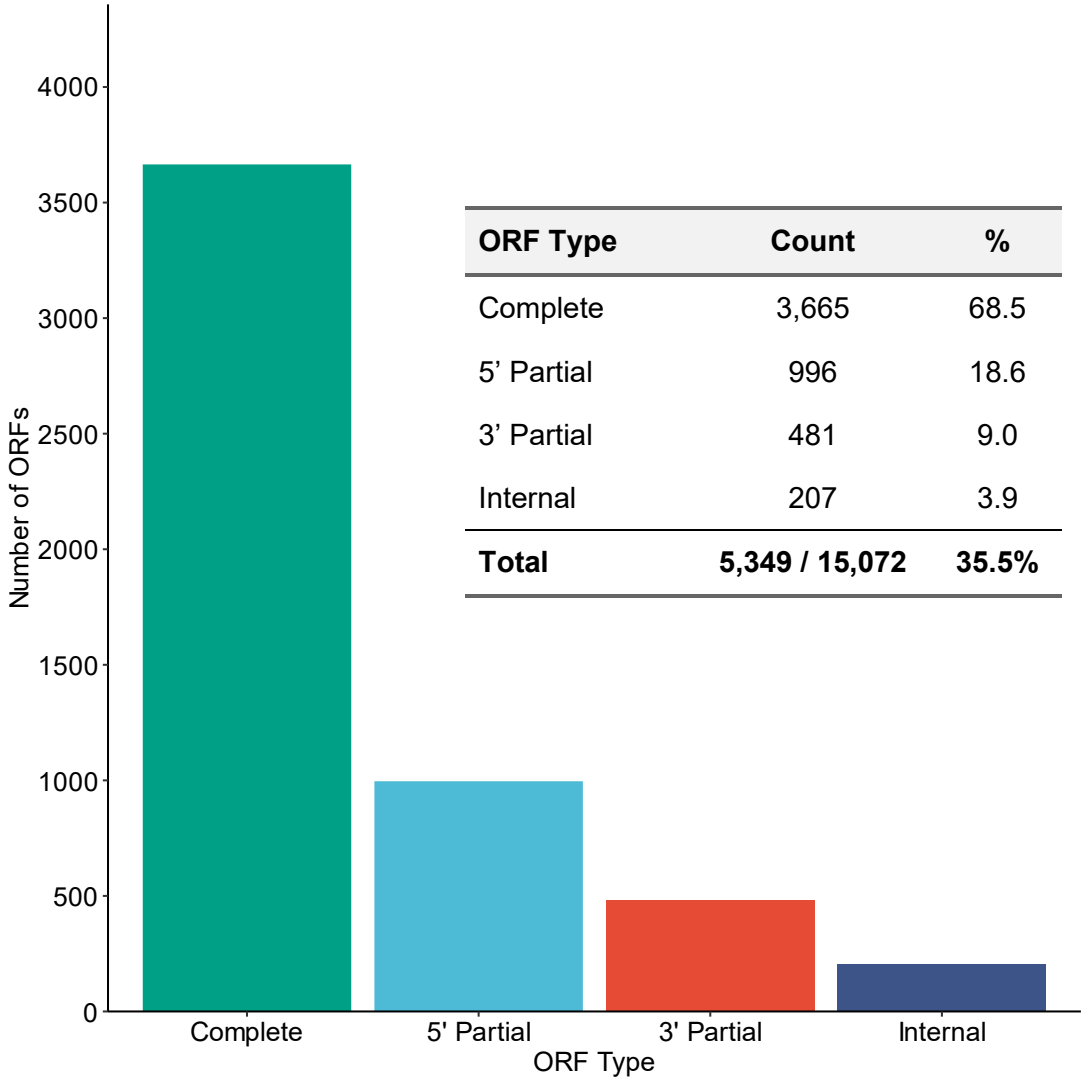


Figure 5.3: Schematic showing the difference between mono-exonic transcript models originating from (A) genes that produce multi-exonic and mono-exonic transcript variants, versus (B) genes that solely encode mono-exonic transcripts. Created with BioRender.com.



ORF Type	Count	%
Complete	3,665	68.5
5’ Partial	996	18.6
3’ Partial	481	9.0
Internal	207	3.9
Total	**5,349 / 15,072**	**35.5%**

Figure 5.4: Number of mono-exonic transcripts with ORFs categorised by TransDecoder; “Complete” = start and stop codon present, “5’ partial” = missing start codon but contains a stop codon, “3’ partial” = contains a start codon but no stop codon, and “Internal” = contains both a start and stop codon but the ORF is contained within a longer sequence which could potentially encode a longer protein. Inset table shows exact numbers of each category.

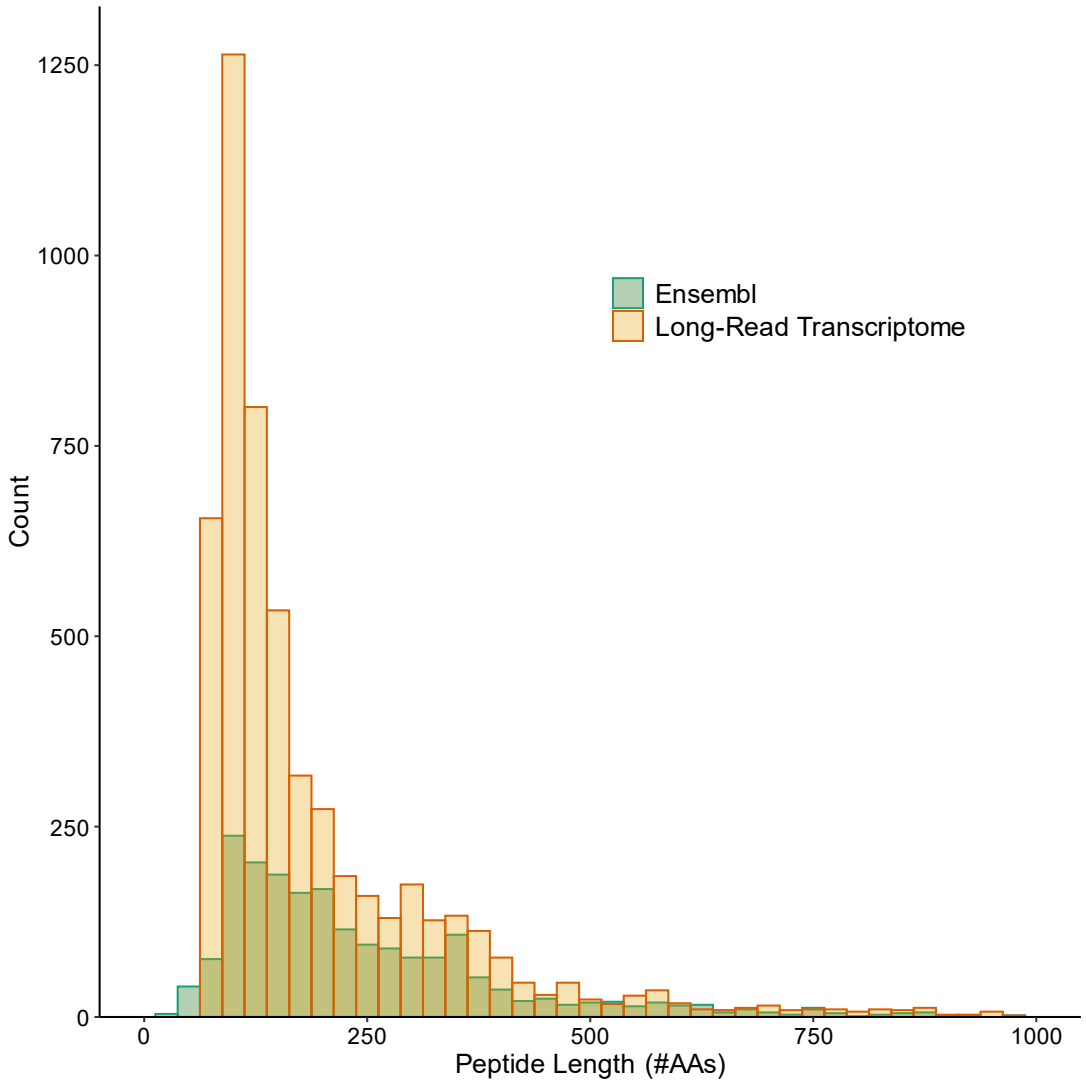


Figure 5.5: Lengths of the predicted peptides extracted from the long-read mono-exonic ORFs versus the peptide sequences predicted by Ensembl in the Ssal_v3.1 annotation.

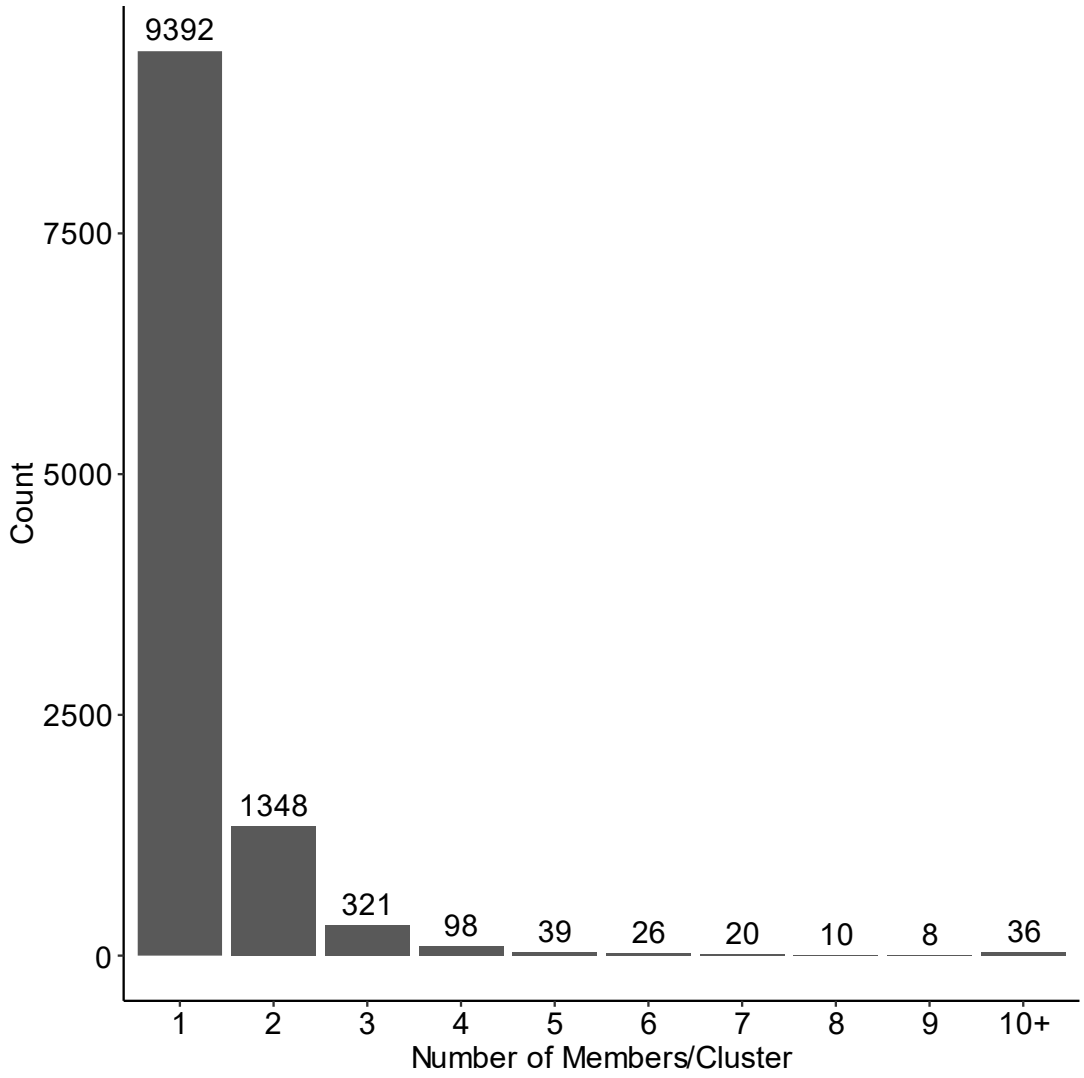


Figure 5.6: Barplot showing the number of mono-exonic transcript models assigned to CD-HIT clusters based on nucleotide sequence similarity >95%. Numbers above the bars indicate how many clusters are in each bin.

Table 5.1: Details of CD-HIT clusters containing 10 or more mono-exonic transcripts from the long-read transcriptome

CD-HIT Cluster ID	# Transcripts in Cluster	# Unique Genes in Cluster	Number of Chromosomes/Scaffolds
0	12	12	11
1	16	16	15
10	38	10	9
11	12	8	8
19	13	2	2
29	20	1	1
30	20	3	3
33	17	1	1
35	28	6	6
38	15	2	2
57	21	3	3
64	14	3	1
87	122	25	15
105	85	36	17
126	173	37	19
157	60	50	31
256	10	6	5
266	10	2	1
297	18	17	17
593	77	7	1
608	11	1	1
670	13	5	1
685	12	12	11
754	10	5	1
924	11	4	2
1210	10	7	6
1353	16	2	1
2287	11	2	2
3133	24	23	18
4328	13	3	2
5178	16	4	1
5793	12	2	1
6723	10	9	8
7388	11	2	2
8005	13	10	7
10123	12	5	1

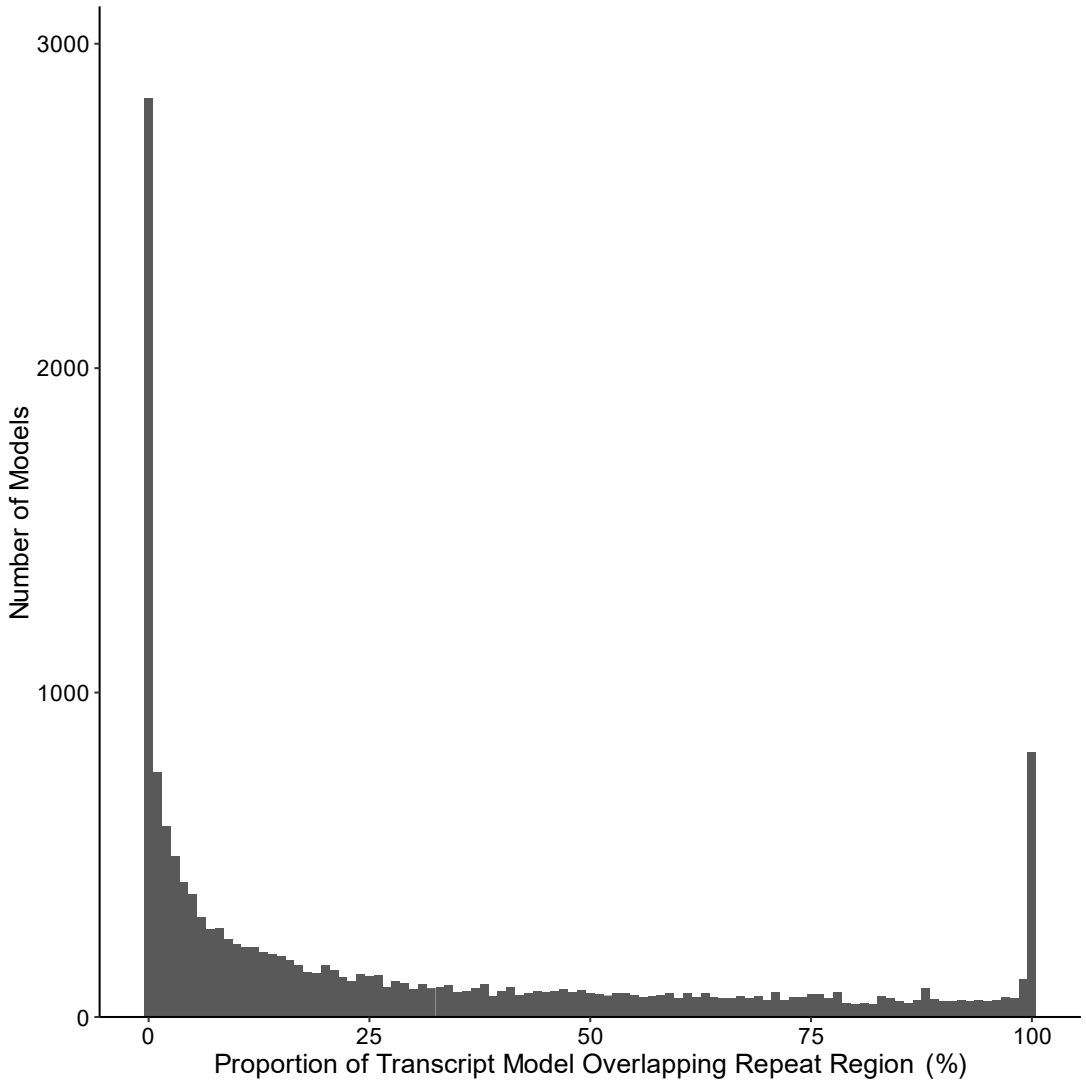


Figure 5.7: Proportion of the length of mono-exonic transcript sequences overlapping repeat regions in the Ensembl annotation of the Ssal_v3.1 assembly.

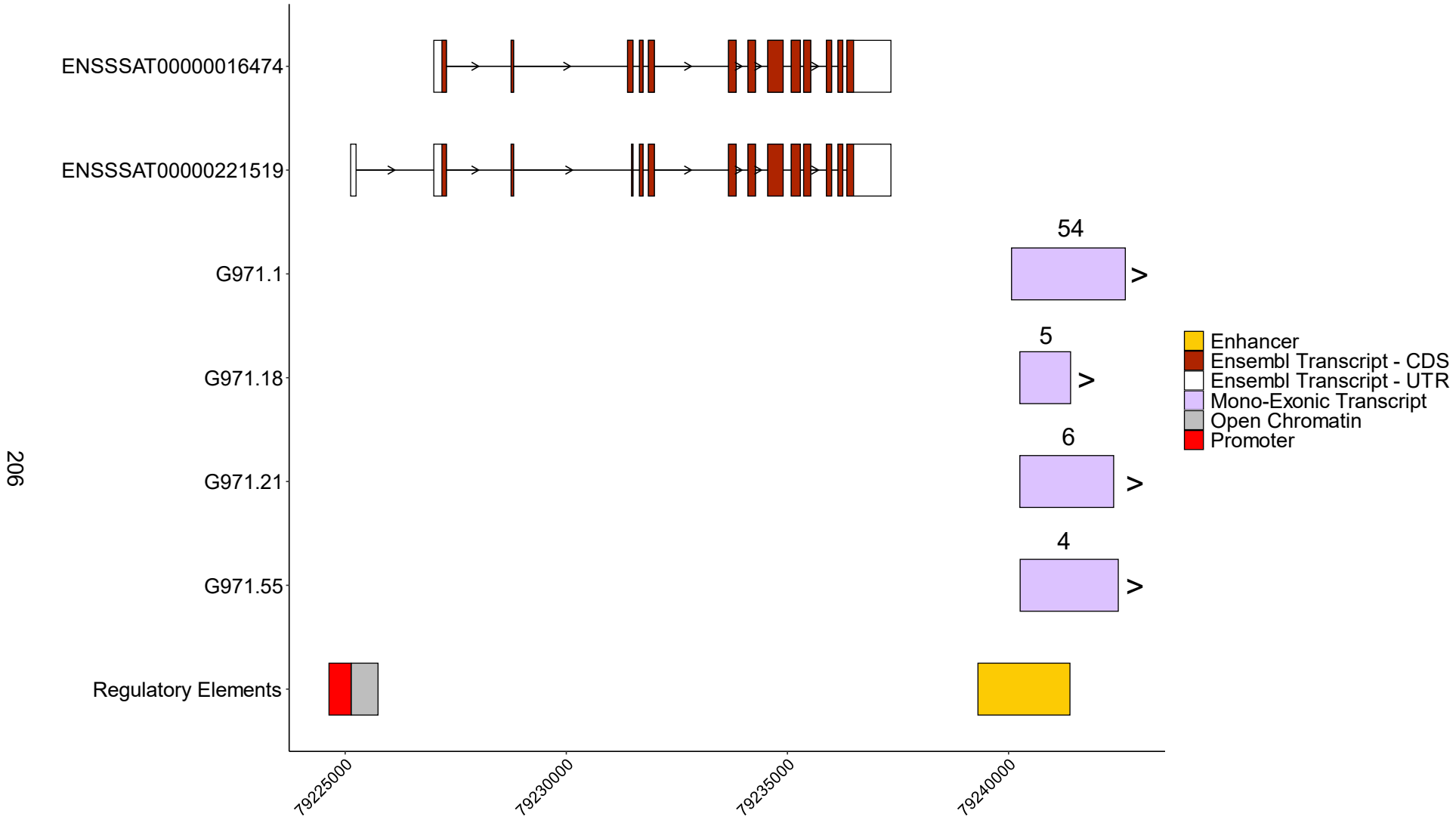


Figure 5.8: Overlap of mono-exonic transcripts G971.1, G971.18, G971.21 and G971.55 with annotated enhancers in the Atlantic salmon Ensembl regulatory build (Ensembl version 112)…(Legend continued on next page).

(Legend continued) Overlap of mono-exonic transcripts G971.1, G971.18, G971. 21 and G971.55 with annotated enhancers in the Atlantic salmon Ensembl regulatory build (Ensembl version 112). The surrounding area is shown to illustrate where the mono-exonic-enhancer overlap occurs in relation to other nearby features. The Ensembl gene shown is cilk1 (ENSSSAG00000007519). Regulatory element IDs are; promoter = ENSSSAR00000375493, open chromatin region = ENSSSAR00000375494, and enhancer = ENSSSAR00000375495. Region depicted is chr1:79,224,600-79,242,700. Arrows depict either forward or reverse strand where introns not marked. Numbers above the mono-exonic model show its read support.

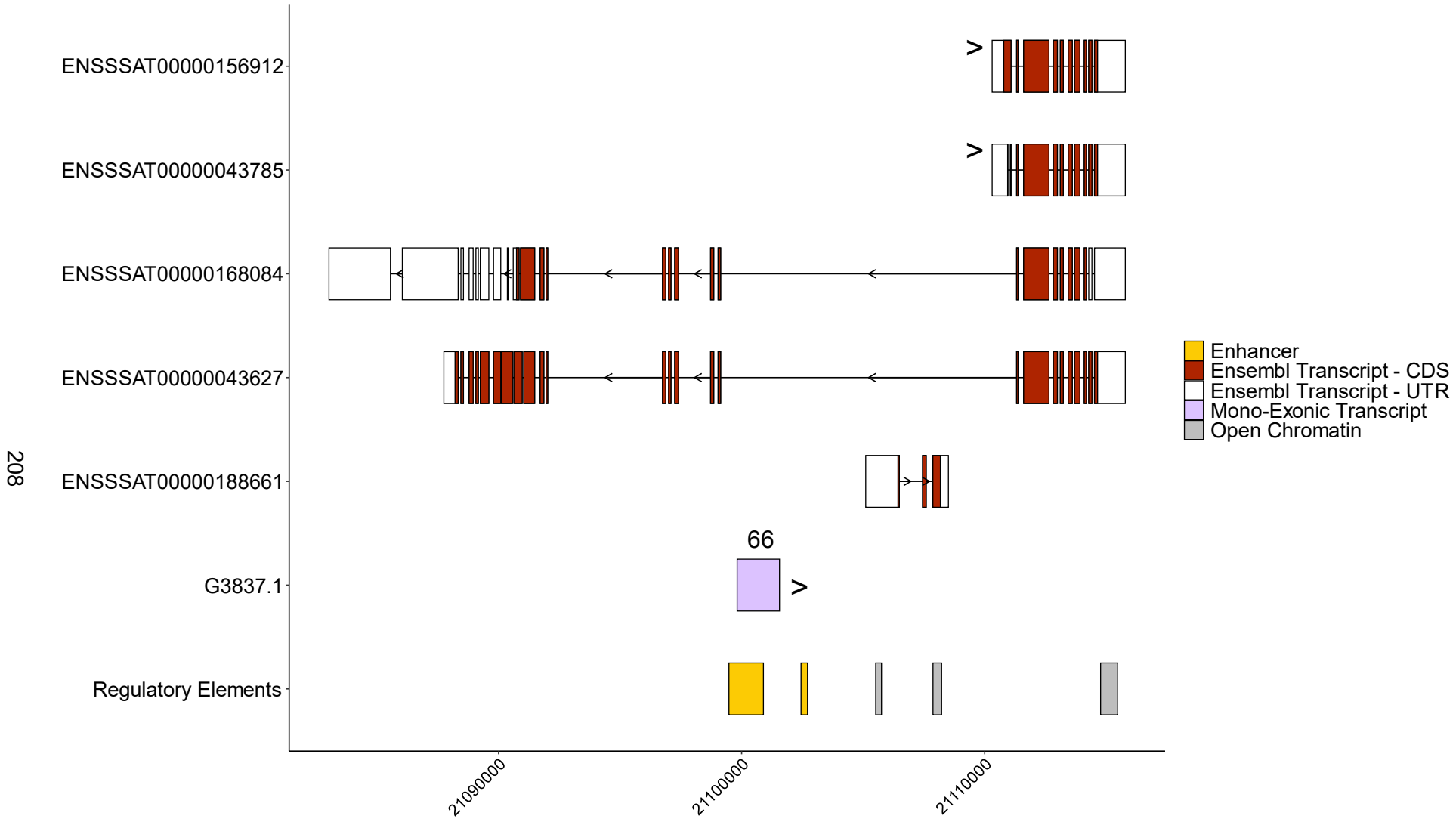


Figure 5.9: Overlap of mono-exonic transcript G3837.1 with annotated enhancers in the Atlantic salmon Ensembl regulatory build (Ensembl version 112)…(Legend continued on next page).

(Legend continued) Overlap of mono-exonic transcript G3837.1 with annotated enhancers in the Atlantic salmon Ensembl regulatory build (Ensembl version 112). The surrounding area is shown to illustrate where the mono-exonic-enhancer overlap occurs in relation with other nearby features. Ensembl genes shown are ENSSSAG00000029006 (reverse strand) and nrn1lb (ENSSSAG00000118447; forward strand). The regulatory elements IDs are as follows; enhancers a) ENSSSAR00000517371, b) ENSSSAR00000517372, and open chromatin regions c) ENSSSAR00000517373, d) ENSSSAR00000517374, e) ENSSSAR00000517375. Region depicted is chr11: 21,083,000-21,115,800. Arrows depict either forward or reverse strand where introns not marked. Numbers above the mono-exonic model show its read support

.

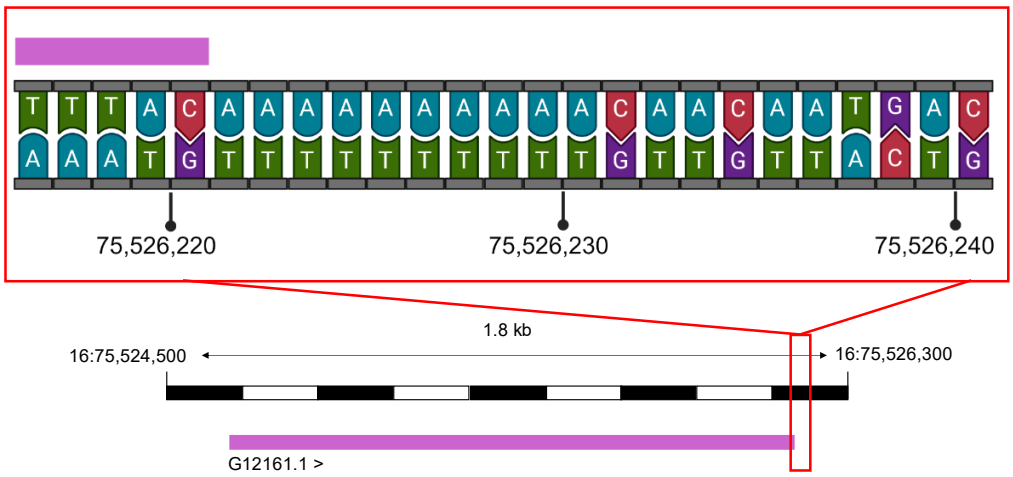


Figure 5.10: Diagram showing polyA motif located at TTS of predicted mono-exonic retrogene G12161.1. Sequence shown accurate to is G12161.1 showing 11/12 bases post-TSS to be As. Created with BioRender.com.

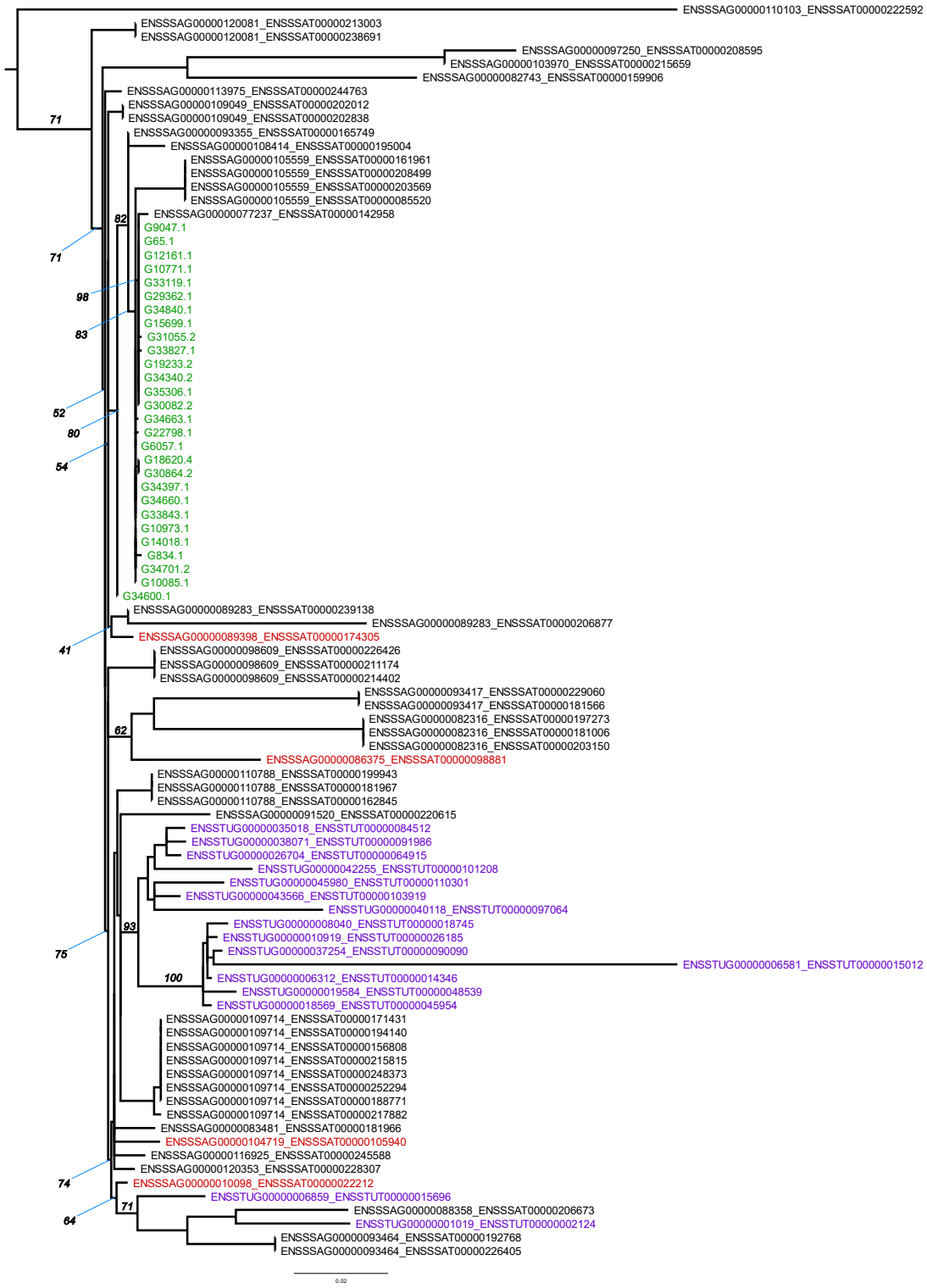


Figure 5.11: Maximum-likelihood phylogenetic tree showing long-read RNA-seq mono-exonic retrogene candidates (green), the transcripts of the Ensembl genes they hit against using BLAST (red), predicted paralogues of the reference BLAST hits (black) and predicted brown trout orthologues of the Atlantic salmon reference family (purple). Tree is midpoint rooted and branch support is annotated in italics for key nodes.

Chapter 6: General Discussion

6.1 Main Findings

This Thesis reports the first ONT-based transcriptome assembly for Atlantic salmon and develops techniques for conducting transcript-level expression analysis with long-read RNA-seq data. As such, my work represents a significant advancement in salmonid and non-model fish genomics, where the power of long-read RNA-seq for resolving gene and transcript expression dynamics for species with complex genomes is increasingly recognised.

In Chapter 2, I present a long-read transcriptome assembled using data from two distinct experiments; an embryonic development timecourse, and an immunostimulation study. The ONT transcriptome captured significant novel transcript diversity originating from both previously annotated and novel genes, highlighting the value of long-read RNA-seq for genome functional annotation in salmonids.

Using this long-read transcriptome, Chapter 3 describes the development of methods for DTE analysis, done using established short-read bioinformatic tools to examine transcript expression regulation in response to viral and bacterial immunostimulation. DTE remains an underexplored approach for long-read RNA-seq due to historic limitations in sequencing accuracy, throughput and the high costs of early TGS platforms (Van Dijk et al., 2018; Stark et al., 2019). However, by leveraging recent improvements in ONT RNA-seq per-base accuracy and throughput, I successfully captured the dynamics of alternative transcript expression during Atlantic salmon innate immune responses. Simply focussing on whole gene level expression overlooks potentially important changes in transcript usage, which are considered important to immune responses (Ergun et al., 2012; Liu et al., 2022b). Chapter 3 builds knowledge of the transcriptome underlying innate immunity in salmonids, contributing to the body of research dedicated to improving fish health in aquaculture settings.

Chapter 4 reports the use of SOM clustering to resolve transcript expression dynamics across embryonic development - from blastula to the eyed stage,

capturing many key transitions, including ZGA, gastrulation, segmentation and tissue patterning. The SOM approach deployed builds on past work in Atlantic salmon (Perojil-Morata, 2024) and zebrafish (Baranasic et al., 2022; White et al., 2017a) performed using bulk RNA-seq and other short-read 'omics data. With SOM clustering I was able to identify within-gene alternative transcript usage at different stages of development. For some genes, this indicated distinct transcript usage between maternally-derived mRNA and mRNAs transcribed by the embryo after ZGA. Focussing on salmonids, the SOM clustering approach employed in Chapter 4 provides a versatile method for studying embryogenesis, an important stage of ontogeny which forms the basis for adult traits essential to aquaculture production. The data reported in Chapter 4 provides a foundation for further investigations into both transcriptional regulation of development and the transcriptomic basis for developing traits in salmonid embryos. SOM (or other clustering methods) can further be readily adapted for long-read RNA-seq studies aiming to identify alternative transcript usage in complex datasets, including timecourses and other multifactorial experimental designs.

Chapter 5 explores the prevalence of mono-exonic transcripts, a class of RNA commonly discarded from RNA-seq analysis (Su et al., 2024). Previous studies have identified mono-exonic transcripts as largely representing non-coding RNAs, including lncRNAs and eRNAs (Laurent et al., 2015; Kubiak & Makałowska, 2017). Results from Chapter 5, including identification of a novel retrogene family and potential eRNAs, add to our knowledge of the expressed mono-exonic transcript repertoire in Atlantic salmon, forming a basis for further investigations into the role of these molecules in gene regulation in a variety of settings.

6.2 Applications and Benefits of this Research

Long-read RNA-seq is yet to be widely applied to aquaculture finfish species despite its proven ability to identify novel transcript structures and resolve transcript expression deriving from complex loci. The long-read approaches

developed in this Thesis fundamentally advance the field of long-read RNA-seq in finfish to include ONT-based RNA-seq methods in Atlantic salmon.

6.2.1 Functional Annotation of Immune Responses

The characterisation of transcript variants highly upregulated in the immune response provides a valuable resource to researchers studying salmonid immunity. DETs could be putative biomarkers for rapid disease diagnosis in aquaculture stocks. Early detection enables a faster response to disease outbreaks, better control measures and reduced transmission (Adams & Thompson, 2011). Alternative transcripts could also be used as targets for novel therapeutic treatments (Zhao, 2019) whilst long-read RNA-seq has applications for understanding transcriptional responses to aquaculture vaccines (Fu et al., 2022). Finally, the integration of long-read RNA-seq into quantitative trait locus (QTL) analysis may offer opportunities for describing the genetic basis behind transcript variant expression and its impacts on disease resistance.

6.2.2 Long-Read RNA-Seq-Guided QTL Analysis

QTLs are genomic regions where DNA variants (often single-nucleotide polymorphism [SNP]) are statistically associated with phenotypic traits (Ashton et al., 2016). QTLs are commonly incorporated into marker-assisted selection (MAS) practices in salmonid aquaculture breeding for traits such as disease resistance (Houston et al., 2010; Gonen et al., 2015; Fraslin et al., 2019; Marana et al., 2021), growth (Gutierrez et al., 2012; Wringe et al., 2010), harvest attributes like flesh colour and fat content (Houston et al., 2009; Baranski et al., 2010), and early maturation (Haidle et al., 2008).

QTLs typically contain many correlated variants and consequently it is a challenge to differentiate between causal variants or those in linkage disequilibrium with true casual variants. However, molecular QTLs, where the association is between genetic variants and molecular phenotypes including expression QTLs (eQTL) associated with changes to gene expression, and splicing QTLs (sQTL) associated with alterations to splicing activity, are promising for prioritising potential causal variants within QTLs. eQTLs traditionally interrogate the impact of DNA variants on total gene

expression. My long-read methods would allow transcript-specific eQTLs to associate variation with transcript-specific TPMs thus providing a more refined eQTL. sQTLs can identify genetic variants associated with individual variation in protein-coding isoform usage, which may be responsible for a trait of interest (Yamaguchi et al., 2022; Xiang et al., 2023; Qi et al., 2022). Nevertheless, sQTL analyses are often guided by incomplete or misannotated transcriptomes which limits their power (Bhattacharya et al., 2023). Thus, improvements to transcriptome assembly offered by long-read RNA-seq will significantly aid sQTL analysis by better prediction of protein isoforms produced by alternative splicing (Glinos et al., 2022).

Alternative splicing events and transcript diversity are clearly more accurately captured by long-read RNA-seq. When this approach becomes sufficiently cost effective to upscale to the large numbers of individuals required for statistical power in QTL analysis, it should become possible to identify the genetic basis for transcript-resolved expression and usage in a range of conditions (Castaldi et al., 2022), revealing variants causal for traits of importance in aquaculture.

6.2.3 Predicting Influence of Embryonic Rearing Conditions on Adult Traits

The importance of rearing conditions during teleost embryogenesis is well documented, for example, the impact of embryonic temperatures on long-term muscle development in zebrafish and Atlantic salmon (Johnston et al., 2009; Macqueen et al., 2008). Burgerhout et al. (2017) highlighted the effects of embryonic temperature on DNA methylation in Atlantic salmon, postulating that embryonic conditions have significant impacts on epigenetic variation with knock-on long-term effects for transcriptome expression, including transcript variants (Zhang et al., 2020). However, currently little is known about the role of alternative transcripts in early development of salmonids.

The description of embryonic transcript diversity provided by my project provides researchers with a baseline from which to evaluate the impacts of the environment on embryonic transcriptomes and their knock-on effects to the adult fish. Additionally, increased transcriptome characterisation may allow elucidation of the impact of alternative protein isoforms on embryonic

processes like stem cell pluripotency (Das et al., 2011) or impacts of splicing on transcription factors such as *nanog* and *cdx2* that influence embryonic differentiation (Revil et al., 2010).

6.2.4 Identification of Gene Editing Targets

The ability to alter genomes with CRISPR-Cas9 methods has revolutionised genome editing practices and facilitated assessment of functional impacts of knocking out or inserting target genes (Ran et al., 2013; Doudna & Charpentier, 2014). CRISPR-Cas9 genome editing has been used widely in aquaculture settings to target production traits including disease resistance, growth and sterility (reviewed in Gratacap et al., 2019). The work herein provides greater scope for more precise genome editing by annotating previously undescribed transcript features that could be targeted in future genome editing studies. In addition, the transcript-level resolution of my data may lead to improvements to the design of guide RNAs used in the delivery of the Cas9 protein to the genome, thus increasing precision of gene edits. Conversely, long-read RNA-seq has been used to validate the efficacy of gene edits in model species (Kim et al., 2021b), an approach which could be applied to aquaculture species in the future.

6.3. Limitations and Opportunities

6.3.1 Sample Diversity

Whilst this Thesis has provided valuable insights into transcriptomic diversity associated with Atlantic salmon immunity and embryogenesis, it is important to acknowledge limitations of the project. Considering the transcriptome assembly in Chapter 2, I obviously have not captured the full complement of transcripts expressed by Atlantic salmon on account of only sequencing whole embryos and adult head kidney tissue. Tissue, or even cell-specific transcripts have been identified as a key contributor to transcriptome diversity (Inamo et al., 2024). Consequently, incorporating additional tissue panels to the transcriptome assembly would improve the capture of transcript repertoire and hence diversity of future transcriptome assemblies. For instance, AQUA-FAANG conducted short-read RNA-seq on 8 tissues from both sexually immature and mature salmon individuals (brain, ovary, testes,

liver, muscle, gill, head kidney and distal intestine), a dataset which remains untouched by long-read RNA-seq. Carrying out long-read RNA-seq on this dataset would capture a host of tissue-specific transcript variation not expressed in head kidney or embryos thus increasing completeness of the transcriptome assembly.

Many other stressors have been shown to elicit transcriptomic responses. Alternative feed sources have impacts on immune function in Atlantic salmon (Tawfik et al., 2024) and can cause intestinal inflammation (Kiron et al., 2020). Understanding the transcripts driving these responses may offer the clues needed to resolve these issues and produce non-harmful alternative feeds. Another stressor becoming more pertinent to salmonid aquaculture is heat stress. Acute heat stress caused abnormal steroid synthesis in Atlantic salmon liver (Shi et al., 2019) whilst elevated temperatures in Chinook salmon were linked to increased apoptosis and induction of the inflammatory response (Tomalty et al., 2015). The expansion of salmonid aquaculture to warmer coastal waters such as China places fish under heat stress due to high summer water temperatures (Shi et al., 2019). Additionally, sea temperature increases caused by global warming will begin to place salmon under chronic stress in cold-water producers like Norway and Scotland, potentially having drastic economic ramifications to these industries (Ahmed et al., 2019). Thus, it is paramount to better understand the transcriptomic responses to stressors which drive physiological changes. Long-read RNA-seq studies will be able to address transcriptome responses to a range of potential stressors at greater resolution than short-read methods.

A constraint related to the immunostimulation study in Chapter 3 is the sole use of head kidney tissue. Whilst head kidney is the primary haematopoietic organ in teleosts (Martorell Ribera et al., 2020), other organs have immune activity such as spleen, thymus, kidney and interbrachial lymphoid tissue (reviewed in Bjørgen & Koppang, 2021), as well as gill (Emam et al., 2022), intestine (Kortner et al., 2024), and liver (Taylor et al., 2024). Thus, the findings in Chapter 3 fail to fully encapsulate transcriptional variability employed in the innate immune response. Future long-read RNA-seq studies in different immune tissues would capture tissue-specific transcript variants

(Inamo et al., 2024) and allow a more comprehensive analysis of the innate immune response in salmonids.

Furthermore, the use of poly I:C and inactivated *Vibrio* as mimics of natural infection do not represent a natural infection scenario so care should be taken when applying the findings of Chapter 3 to cases involving natural infection. Natural viruses can engage a broader range of immune receptors through recognition of proteins like viral envelope glycoproteins (Zhou et al., 2021) which may induce expression of specific transcripts not captured by the mimic stimulations in this project. Whilst inactivated *Vibrio* should present the same PAMPs to PRRs as a live *Vibrio* strain, live pathogens themselves also employ various tactics to evade host immune responses. For instance, intracellular bacteria affecting teleosts can evade phagocytic reactive oxygen responses by converting reactive oxygen anions to hydrogen peroxide and catalysing this further into non harmful O_2 and H_2O (Grayfer et al., 2014). Looking to viruses, the salmonid viral pathogens IPNV and ISAV possess the ability to suppress interferon pathways (Collet, 2014) whilst a non-virion gene expressed by viral hemorrhagic septicemia virus (VHSV) and infectious hematopoietic necrosis virus (IHNV) has been implicated in the suppression of host apoptosis in early innate immune responses (Ammayappan & Vakharia, 2011).

As discussed in Chapter 1, section 1.4.3, aquatic diseases represent a significant challenge for aquaculture sustainability. Given the ability of long-read RNA-seq to resolve transcript-level immune responses this, future studies should prioritise using long-read RNA-seq for understanding host pathogen interactions.

Six stages of embryonic development were sampled in Chapter 4, using three biological replicates. Aside from increasing replication to enhance statistical power of the DTE analysis, sampling more development stages would provide a higher resolution examination of transcriptional regulation of embryogenesis. The AQUA-FAANG project studied 14 development stages with short-read RNA-seq including a pre-basculation stage at the end of cleavage which contained greater signatures of maternally-derived mRNAs

than in my data (Perojil-Morata, 2024). Another study conducted by White and colleagues (2017) in zebrafish employed a similar clustering approach with short-read RNA-seq covering 18 developmental time points covering 4 pre-ZGA stages up to 5 days post-fertilisation, revealing a wealth of novel gene and transcript diversity.

In addition to alternative splicing, my results describe much novel transcript diversity generated via the use of alternative TSSs and TTSs. Long-read RNA-seq has already identified large numbers of unannotated TSSs and TTSs in humans (Byrne et al., 2017) and *Drosophila* (Alfonso-Gonzalez et al., 2023). Integrating other technologies into long-read RNA-seq may allow better characterisation of TSSs and TTSs. 5'-capture sequencing methods like CAGE-seq offer the ability to accurately identify the 5' ends of RNA transcripts and the integration of CAGE and long-read RNA-seq has proved effective for resolving full-length RNA structures, thus better characterising both TSSs and TTSs (Pardo-Palacios et al., 2024b). RAMPAGE methods which are effective for identifying promoter regions and quantifying their expression (Batut & Gingeras, 2015), are being used in conjunction with long-read RNA-seq to validate alternative TSS usage in humans (Reese et al., 2023). Adapting these methods for use in salmonids as well as incorporating the regulatory element information collected by AQUA-FAANG via ChIP-seq and ATAC-seq would constitute a comprehensive approach to both confidently annotate TSSs and quantify their usage under a variety of conditions.

6.3.2 Ohnologue Expression

Due to the recent Ss4R WGD (discussed in Chapter 1, section 1.4.2), the Atlantic salmon genome retains high levels of duplication and ohnologue retention (Lien et al., 2016). As such, salmonids offer a rare opportunity to study the potential fates of ohnologous gene pairs such as neofunctionalisation, subfunctionalisation and pseudogenisation (Gillard et al., 2021). It has been shown that human ohnologue pairs have a measurable loss of alternatively spliced transcript forms as time since WGD increases (Su et al., 2006) perhaps supporting a subfunctionalisation

situation where one ohnologue retains a portion of the ancestral transcript variants and its pair retains the remainder (Iñiguez & Hernández, 2017). These findings are consistent with a study in *Arabidopsis thaliana* which reveal extensive divergence of alternative splicing products between ohnologue pairs (Zhang et al., 2010). Currently, little is known about the regulation of alternative splicing and the extent of transcript partitioning between diverging ohnologues in the salmonids.

I have shown that long-read RNA-seq can differentiate highly similar ohnologues (Chapter 2, section 2.3.1) and results in Chapters 3 and 4 indicate similar transcript expression patterns in the few ohnologue pairs highlighted. However, questions regarding the conservation of ohnologue expression at the transcript level remain unanswered. Genetic redundancy following WGD allows for divergence of ohnologue function (Robertson et al., 2017), however, it is unknown to what extent this redundancy facilitates the evolution of transcript diversity as ohnologues diverge. Another unanswered question is how the complex rediploidisation process of salmonids (Gundappa et al. 2022) has impacted transcript expression and further, how it impacts alternative transcript usage between ohnologue pairs perhaps inducing subfunctionalisation. The data produced in this project could be used to resolve global transcript-level expression dynamics in ohnologues retained from Ss4R, providing novel insights into genome functional evolution after WGD. Finally, the evolution of ohnologue-specific transcript variants could be studied by conducting comparative long-read studies between salmonids, a teleost outgroup like northern pike*,* and other vertebrates.

6.3.3 Long-Read Single-Cell RNA-Seq

My long-read RNA-seq method is a form of bulk RNA-seq where cells or tissues are homogenised before sequencing. However, we know that tissue phenotypes are underwritten by cell-specific transcript expression (Inamo et al., 2024; Paik et al., 2020). As such, the contribution of gene and transcript expression from individual cell types was masked by the bulk RNA-seq nature of my method.

Single-cell and single-nuclei RNA-seq is an exciting recent development, which can elucidate the transcriptomic profiles of hundreds to millions of individual cells in single experiments (Papalexi & Satija, 2018; Kolodziejczyk et al., 2015; Aldridge & Teichmann, 2020). In Atlantic salmon, single-nuclei sequencing has already been used to describe the roles of different liver cell types for fighting bacterial infection (Taylor et al., 2022). Another study revealed significant T-cell and B-cell heterogeneity in healthy Atlantic salmon head kidney with matched single-cell and single-nuclei RNA-seq (Andresen et al., 2024). Single-cell RNA-seq is gaining increasing interest for aquaculture research with widespread applications including resolving the cell-specific basis for immune function and development (reviewed in Daniels et al., 2023). Long-read single-cell RNA-seq has been developed for mammal species (Tian et al., 2021; Lebrigand et al., 2020), revealing cell-type-specific mutations in human cancer (Shiau et al., 2023) and generating comprehensive annotations of T-cell receptor repertoires in human ovarian cancer (Byrne et al., 2024). However, long-read single-cell RNA-seq has yet to be applied to finfish. This represents an exciting opportunity to pair the transcript-level resolution provided by long-read RNA-seq with the cell-specific resolution of single-cell RNA-seq to further elucidate alternative transcript expression dynamics between individual cell types.

6.4 Closing Words

To conclude, this Thesis represents a major contribution towards understanding transcript diversity and expression dynamics of embryogenesis and the innate immune response in Atlantic salmon. It places long-read sequencing as a powerful emerging technology with potentially wide-reaching applications to advance fundamental knowledge in fish biology, and the sustainability of aquaculture. The data, bioinformatic pipelines and code generated in this project will be made publicly available to contribute to improved Atlantic salmon reference annotations which will undoubtedly assist further transcriptomic study of the salmonids and beyond.

References

Aaen, S.M., Helgesen, K.O., Bakke, M.J., Kaur, K. and Horsberg, T.E. (2015) 'Drug resistance in sea lice: a threat to salmonid aquaculture'. *Trends in Parasitology*, **31**(2), pp. 72-81.

Aanes, H., Winata, C.L., Lin, C.H., Chen, J.P., Srinivasan, K.G., Lee, S.G., Lim, A.Y., Hajan, H.S., Collas, P., Bourque, G. and Gong, Z. (2011) 'Zebrafish mRNA sequencing deciphers novelties in transcriptome dynamics during maternal to zygotic transition'. *Genome Research*, **21**(8), pp. 1328-1338.

Abdellaoui, N. and Kim, M.S. (2024) 'Transcriptome Profiling of Gene Expression in Atlantic Salmon (*Salmo salar*) at Early Stage of Development'. *Marine Biotechnology*, pp. 1-11.

Abdelrahman, H., ElHady, M., Alcivar-Warren, A., Allen, S., Al-Tobasei, R., Bao, L., Beck, B., Blackburn, H., Bosworth, B. and Buchanan, J. (2017) 'Aquaculture genomics, genetics and breeding in the United States: current status, challenges, and priorities for future research'. *BMC Genomics*, **18**(191), pp. 1-23.

Abebe, J.S., Alwie, Y., Fuhrmann, E., Leins, J., Mai, J., Verstraten, R., Schreiner, S., Wilson, A.C. and Depledge, D.P. (2024) 'Nanopore Guided Annotation of Transcriptome Architectures'. *bioRxiv*, [preprint], doi: 10.1101/2024.04.02.587744.

Adamek, M., Davies, J., Beck, A., Jordan, L., Becker, A.M., Mojzesz, M., Rakus, K., Rumiac, T., Collet, B., Brogden, G. and Way, K. (2021) 'Antiviral actions of 25-hydroxycholesterol in fish vary with the virus-host combination'. *Frontiers in Immunology*, **12**, 581786.

Adams, A. and Thompson, K.D. (2011) 'Development of diagnostics for aquaculture: challenges and opportunities'. *Aquaculture Research*, **42**, pp. 93-102.

Adams, M.D., Dubnick, M., Kerlavage, A.R., Moreno, R., Kelley, J.M., Utterback, T.R., Nagle, J.W., Fields, C. and Venter, J.C. (1992) 'Sequence identification of 2,375 human brain genes'. *Nature*, **355**(6361), pp. 632-634.

Adenaya, A., Berger, M., Brinkhoff, T., Ribas-Ribas, M. and Wurl, O. (2023) 'Usage of antibiotics in aquaculture and the impact on coastal waters'. *Marine Pollution Bulletin*, **188**, 114645.

Aderem, A. and Ulevitch, R.J. (2000) 'Toll-like receptors in the induction of the innate immune response'. *Nature*, **406**(6797), pp. 782-787.

Adiconis, X., Haber, A.L., Simmons, S.K., Levy Moonshine, A., Ji, Z., Busby, M.A., Shi, X., Jacques, J., Lancaster, M.A., Pan, J.Q. and Regev, A. (2018) 'Comprehensive comparative analysis of 5′-end RNA-sequencing methods'. *Nature Methods*, **15**(7), pp. 505-511.

Ahmed, N., Thompson, S. and Glaser, M. (2019) 'Global aquaculture productivity, environmental sustainability, and climate change adaptability'. *Environmental Management*, **63**, pp. 159-172.

Aird, D., Ross, M.G., Chen, W.S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C. and Gnirke, A. (2011) 'Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries'. *Genome Biology*, **12**(R18), pp. 1-14.

Aldridge, S. and Teichmann, S.A. (2020) 'Single cell transcriptomics comes of age'. *Nature Communications*, **11**(1), 4307.

Alexander, R.P., Fang, G., Rozowsky, J., Snyder, M. and Gerstein, M.B. (2010) 'Annotating non-coding regions of the genome'. *Nature Reviews Genetics*, **11**(8), pp. 559-571.

Alexandratos, N. and Bruinsma, J. (2012). *World agriculture towards 2030/2050: the 2012 revision*. ESA Working Paper No.12-03. Rome, FAO.

Alfonso-Gonzalez, C., Legnini, I., Holec, S., Arrigoni, L., Ozbulut, H.C., Mateos, F., Koppstein, D., Rybak-Wolf, A., Bönisch, U., Rajewsky, N. and Hilgers, V. (2023) 'Sites of transcription initiation drive mRNA isoform selection'. *Cell*, **186**(11), pp. 2438-2455.

Ali, A., Thorgaard, G.H. and Salem, M. (2021) 'PacBio Iso-Seq improves the rainbow trout genome annotation and identifies alternative splicing associated with economically important phenotypes'. *Frontiers in Genetics*, **12**, 683408.

Álvarez, C.A., Guzmán, F., Cárdenas, C., Marshall, S.H. and Mercado, L. (2014) 'Antimicrobial activity of trout hepcidin'. *Fish & Shellfish Immunology*, **41**(1), pp. 93-101.

Alzaid, A., Castro, R., Wang, T., Secombes, C.J., Boudinot, P., Macqueen, D.J. and Martin, S.A. (2016) 'Cross talk between growth and immunity: coupling of the IGF axis to conserved cytokine pathways in rainbow trout'. *Endocrinology*, **157**(5), pp. 1942-1955.

Amarasinghe, S.L., Su, S., Dong, X., Zappia, L., Ritchie, M.E. and Gouil, Q. (2020) 'Opportunities and challenges in long-read sequencing data analysis'. *Genome Biology*, **21**(30).

Ammayappan, A. and Vakharia, V.N. (2011) ‘Nonvirion protein of novirhabdovirus suppresses apoptosis at the early stage of virus infection’. *Journal of Virology*, **85**(16), pp. 8393-8402.

Anders, S. and Huber, W. (2010) ‘Differential expression analysis for sequence count data’. *Nature Precedings*.

Anderson, J.L., Asche, F., Garlock, T. and Chu, J. (2017) ‘Aquaculture: Its role in the future of food’, in *World Agricultural Resources and Food Security: International Food Security (Frontiers of Economics and Globalization, Vol.17)*. Emerald Publishing Limited, Leeds, pp. 159-173.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T. and Ntini, E. (2014) ‘An atlas of active enhancers across human cell types and tissues’. *Nature*, **507**(7493), pp. 455-461.

Andersson, R. and Sandelin, A. (2020) ‘Determinants of enhancer and promoter activities of regulatory elements’. *Nature Reviews Genetics*, **21**(2), pp. 71-87.

Andresen, A.M.S., Boudinot, P. and Gjøen, T. (2020) ‘Kinetics of transcriptional response against poly (I: C) and infectious salmon anemia virus (ISAV) in Atlantic salmon kidney (ASK) cell line’. *Developmental & Comparative Immunology*, **110**, 103716.

Andresen, A.M., Taylor, R.S., Grimholt, U., Daniels, R.R., Sun, J., Dobie, R., Henderson, N.C., Martin, S.A., Macqueen, D.J. and Fosse, J.H. (2024) ‘Mapping the cellular landscape of Atlantic salmon head kidney by single cell and single nucleus transcriptomics’. *Fish & Shellfish Immunology*, **146**, 109357.

Archibald, A.L., Bottema, C.D., Brauning, R., Burgess, S.C., Burt, D.W., Casas, E., Cheng, H.H., Clarke, L., Couldrey, C., Dalrymple, B.P. and Elsik, C.G. (2015) ‘Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project: open letter’. *Genome Biology*, **16**(57).

Arner, E., Daub, C.O., Vitting-Seerup, K., Andersson, R., Lilje, B., Drabløs, F., Lennartsson, A., Rönnerblad, M., Hrydziuszko, O., Vitezic, M. and Freeman, T.C. (2015) ‘Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells’. *Science*, **347**(6225), pp. 1010-1014.

Arnold, P.R., Wells, A.D. and Li, X.C. (2020) ‘Diversity and emerging roles of enhancer RNA in regulation of gene expression and cell fate’. *Frontiers in Cell and Developmental Biology*, **7**, 377.

Ashton, D.T., Ritchie, P.A. and Wellenreuther, M. (2017) ‘Fifteen years of quantitative trait loci studies in fish: challenges and future directions’. *Molecular Ecology*, **26**(6), pp. 1465-1476.

Aunsmo, A., Valle, P.S., Sandberg, M., Midtlyng, P.J. and Bruheim, T. (2010) ‘Stochastic modelling of direct costs of pancreas disease (PD) in Norwegian farmed Atlantic salmon (*Salmo salar* L.)’. *Preventive Veterinary Medicine*, **93**(2-3), pp. 233-241.

Aunsmo, A., Persson, D., Stormoen, M., Romstad, S., Jamtøy, O. and Midtlyng, P.J. (2023) ‘Real-time monitoring of cause-specific mortality-and losses in industrial salmon farming’. *Aquaculture*, **563**, 738969.

Aviña-Padilla, K., Ramírez-Rafael, J.A., Herrera-Oropeza, G.E., Muley, V.Y., Valdivia, D.I., Díaz-Valenzuela, E., García-García, A., Varela-Echavarría, A. and Hernández-Rosales, M. (2021) ‘Evolutionary perspective and expression analysis of intronless genes highlight the conservation of their regulatory role’. *Frontiers in Genetics*, **12**, 654256.

Baralle, F.E. and Giudice, J. (2017) ‘Alternative splicing as a regulator of development and tissue identity’. *Nature Reviews Molecular Cell Biology*, **18**(7), pp. 437-451.

Baranasic, D., Hörtenhuber, M., Balwierz, P.J., Zehnder, T., Mukarram, A.K., Nepal, C., Várnai, C., Hadzhiev, Y., Jimenez-Gonzalez, A., Li, N. and Wragg, J. (2022) ‘Multiomic atlas with functional stratification and developmental dynamics of zebrafish cis-regulatory elements’. *Nature Genetics*, **54**(7), pp. 1037-1050.

Baranski, M., Moen, T. and Våge, D.I. (2010) ‘Mapping of quantitative trait loci for flesh colour and growth traits in Atlantic salmon (*Salmo salar*)’. *Genetics Selection Evolution*, **42**(17).

Barría, A., Trịnh, T.Q., Mahmuddin, M., Peñaloza, C., Papadopoulou, A., Gervais, O., Chadag, V.M., Benzie, J.A. and Houston, R.D. (2021) ‘A major quantitative trait locus affecting resistance to Tilapia lake virus in farmed Nile tilapia (*Oreochromis niloticus*)’. *Heredity*, **127**(3), pp. 334-343.

Batut, P. and Gingeras, T.R. (2013) ‘RAMPAGE: Promoter Activity Profiling by Paired-End Sequencing of 5′-Complete cDNAs’. *Current Protocols in Molecular Biology*, **104**(1), pp. 25B.11.1-25B.11.16.

Bayega, A., Oikonomopoulos, S., Gregoriou, M.E., Tsoumani, K.T., Giakountis, A., Wang, Y.C., Mathiopoulos, K.D. and Ragoussis, J. (2021) ‘Nanopore long-read RNA-seq and absolute quantification delineate transcription dynamics in early embryo development of an insect pest’. *Scientific Reports*, **11**(1), 7878.

Bayne, C.J. and Gerwick, L. (2001) 'The acute phase response and innate immunity of fish'. *Developmental & Comparative Immunology*, **25**(8-9), pp. 725-743.

Begik, O., Diensthuber, G., Liu, H., Delgado-Tejedor, A., Kontur, C., Niazi, A.M., Valen, E., Giraldez, A.J., Beaudoin, J.D., Mattick, J.S. and Novoa, E.M. (2023) 'Nano3P-seq: transcriptome-wide analysis of gene expression and tail dynamics using end-capture nanopore cDNA sequencing'. *Nature Methods,* **20**(1), pp. 75-85.

Beiki, H., Liu, H., Huang, J., Manchanda, N., Nonneman, D., Smith, T.P.L., Reecy, J.M. and Tuggle, C.K. (2019) 'Improved annotation of the domestic pig genome through integration of Iso-Seq and RNA-seq data'. *BMC Genomics,* **20**(344).

Belton, J.M., McCord, R.P., Gibcus, J.H., Naumova, N., Zhan, Y. and Dekker, J. (2012) 'Hi–C: a comprehensive technique to capture the conformation of genomes'. *Methods*, **58**(3), pp. 268-276.

Berg, M.D. and Brandl, C.J. (2021) 'Transfer RNAs: diversity in form and function'. *RNA Biology*, **18**(3), pp. 316-339.

Bernatchez, L., Wellenreuther, M., Araneda, C., Ashton, D.T., Barth, J.M., Beacham, T.D., Maes, G.E., Martinsohn, J.T., Miller, K.M., Naish, K.A. and Ovenden, J.R. (2017) 'Harnessing the power of genomics to secure the future of seafood'. *Trends in Ecology & Evolution*, **32**(9), pp. 665-680.

Berthelot, C., Brunet, F., Chalopin, D., Juanchich, A., Bernard, M., Noël, B., Bento, P., Da Silva, C., Labadie, K., Alberti, A. and Aury, J.M. (2014) 'The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates'. *Nature Communications*, **5**(1), pp. 1-10.

Bhattacharya, A., Vo, D.D., Jops, C., Kim, M., Wen, C., Hervoso, J.L., Pasaniuc, B. and Gandal, M.J. (2023) 'Isoform-level transcriptome-wide association uncovers genetic risk mechanisms for neuropsychiatric disorders in the human brain'. *Nature Genetics*, **55**(12), pp. 2117-2128.

Bjørgen, H., Løken, O.M., Aas, I.B., Fjelldal, P.G., Hansen, T., Austbø, L. and Koppang, E.O. (2019) 'Visualization of CCL19-like transcripts in the ILT, thymus and head kidney of Atlantic salmon (*Salmo salar* L.)'. *Fish & Shellfish Immunology*, **93**, pp. 763-765.

Bjørgen, H. and Koppang, E.O. (2022) 'Anatomy of teleost fish immune structures and organs', in Buchmann, K. and Secombes, C.J. (eds) *Principles of Fish Immunology*. Springer, Cham.

Blix, T.B., Dalmo, R.A., Wargelius, A. and Myhr, A.I. (2021) 'Genome editing on finfish: current status and implications for sustainability'. *Reviews in Aquaculture*, **13**(4), pp. 2344-2363.

Boerlage, A.S., Ashby, A., Herrero, A., Reeves, A., Gunn, G.J. and Rodger, H.D. (2020) 'Epidemiology of marine gill diseases in Atlantic salmon (*Salmo salar*) aquaculture: a review'. *Reviews in Aquaculture*, **12**(4), pp. 2140-2159.

Boltaña, S., Valenzuela-Miranda, D., Aguilar, A., Mackenzie, S. and Gallardo-Escárate, C. (2016) 'Long noncoding RNAs (lncRNAs) dynamics evidence immunomodulation during ISAV-Infected Atlantic salmon (*Salmo salar*)'. *Scientific Reports*, **6**(1), 22698.

Bondad-Reantaso, M.G., MacKinnon, B., Karunasagar, I., Fridman, S., Alday-Sanz, V., Brun, E., Le Groumellec, M., Li, A., Surachetpong, W., Karunasagar, I. and Hao, B. (2023) 'Review of alternatives to antibiotic use in aquaculture'. *Reviews in Aquaculture*, **15**(4), pp. 1421-1451.

Boulet, A., Vest, K.E., Maynard, M.K., Gammon, M.G., Russell, A.C., Mathews, A.T., Cole, S.E., Zhu, X., Phillips, C.B., Kwong, J.Q. and Dodani, S.C. (2018) 'The mammalian phosphate carrier SLC25A3 is a mitochondrial copper transporter required for cytochrome c oxidase biogenesis'. *Journal of Biological Chemistry*, **293**(6), pp. 1887-1896.

Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S. and Crawford, G.E. (2008) 'High-resolution mapping and characterization of open chromatin across the genome'. *Cell*, **132**(2), pp. 311-322.

Breznak, S.M., Kotb, N.M. and Rangan, P. (2023) 'Dynamic regulation of ribosome levels and translation during development', in *Seminars in Cell & Developmental Biology.* Academic Press, vol. 136, pp. 27-37.

Briolat, V., Jouneau, L., Carvalho, R., Palha, N., Langevin, C., Herbomel, P., Schwartz, O., Spaink, H.P., Levraud, J.P. and Boudinot, P. (2014) 'Contrasted innate responses to two viruses in zebrafish: insights into the ancestral repertoire of vertebrate IFN-stimulated genes. *The Journal of Immunology*, **192**(9), pp. 4328-4341.

Brophy, J.A. and Voigt, C.A. (2016) 'Antisense transcription as a tool to tune gene expression'. *Molecular Systems Biology*, **12**(1), 854.

Brosseau, C., Colas, L., Magnan, A. and Brouard, S. (2018) 'CD9 tetraspanin: a new pathway for the regulation of inflammation?'. *Frontiers in Immunology*, **9**, 2316.

Brown, C.G. and Clarke, J. (2016) 'Nanopore development at Oxford nanopore'. *Nature Biotechnology*, **34**(8), pp. 810-811.

Bruce, T.J. and Brown, M.L. (2017) ‘A Review of Immune System Components, Cytokines, and Immunostimulants in Cultured Finfish Species’. *Open Journal of Animal Sciences*, **7**(3), pp. 267-288.

Brudeseth, B.E., Wiulsrød, R., Fredriksen, B.N., Lindmo, K., Løkling, K.E., Bordevik, M., Steine, N., Klevan, A. and Gravningen, K. (2013) ‘Status and future perspectives of vaccines for industrialised fin-fish farming’. *Fish & Shellfish Immunology*, **35**(6), pp. 1759-1768.

Brunet, F.G., Crollius, H.R., Paris, M., Aury, J.M., Gibert, P., Jaillon, O., Laudet, V. and Robinson-Rechavi, M. (2006) ‘Gene loss and evolutionary rates following whole-genome duplication in teleost fishes’. *Molecular Biology and Evolution*, **23**(9), pp. 1808-1816.

Brunner, S.R., Varga, J.F. and Dixon, B. (2020) ‘Antimicrobial peptides of salmonid fish: from form to function’. *Biology*, **9**(8), 233.

Bryzghalov, O., Szcześniak, M. and Makałowska, I. (2016) ‘Retroposition as a source of antisense long non-coding RNAs with possible regulatory functions’. *Acta Biochimica Polonica*, **63**(4), pp. 825-833.

Buchfink, B., Reuter, K. and Drost, H.G. (2021) ‘Sensitive protein alignments at tree-of-life scale using DIAMOND’. *Nature Methods*, **18**(4), pp. 366-368.

Buchmann, K. (2022). ‘Antibacterial Immune Responses’, in Buchmann, K. and Secombes, C.J. (eds) *Principles of Fish Immunology*. Springer, Cham. pp. 511-533.

Buenrostro, J.D., Wu, B., Chang, H.Y. and Greenleaf, W.J. (2015) ‘ATAC-seq: a method for assaying chromatin accessibility genome-wide’. *Current Protocols in Molecular Biology*, **109**(1), pp. 21-29.

Bulger, M. and Groudine, M. (2011) ‘Functional and mechanistic diversity of distal transcription enhancers’. *Cell*, **144**(3), pp. 327-339.

Burgerhout, E., Mommens, M., Johnsen, H., Aunsmo, A., Santi, N. and Andersen, Ø. (2017) ‘Genetic background and embryonic temperature affect DNA methylation and expression of myogenin and muscle development in Atlantic salmon (*Salmo salar*)’. *PLoS ONE*, **12**(6), e0179918.

Byrne, A., Beaudin, A.E., Olsen, H.E., Jain, M., Cole, C., Palmer, T., DuBois, R.M., Forsberg, E.C., Akeson, M. and Vollmers, C. (2017) ‘Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells’. *Nature Communications*, **8**(1), 16027.

Byrne, A., Cole, C., Volden, R. and Vollmers, C. (2019) 'Realizing the potential of full-length transcriptome sequencing'. *Philosophical Transactions of the Royal Society B*, **374**(1786), 20190097.

Byrne, A., Le, D., Sereti, K., Menon, H., Vaidya, S., Patel, N., Lund, J., Xavier-Magalhães, A., Shi, M., Liang, Y. and Sterne-Weiler, T. (2024) 'Single-cell long-read targeted sequencing reveals transcriptional variation in ovarian cancer'. *Nature Communications*, **15**(1), 6916.

Cai, W., Kumar, S., Navaneethaiyer, U., Caballero-Solares, A., Carvalho, L.A., Whyte, S.K., Purcell, S.L., Gagne, N., Hori, T.S., Allen, M. and Taylor, R.G. (2022) 'Transcriptome analysis of Atlantic salmon (*Salmo salar*) skin in response to sea lice and infectious Salmon Anemia Virus co-infection under different experimental functional diets'. *Frontiers in Immunology*, **12**, 787033.

Cain, K. and Swan, C. (2010) 'Barrier function and immunology', in Grosell, M., Farrell, A.P. and Brauner, C.J. (eds) *Fish Physiology*. Academic Press, vol.30, pp. 111-134.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2009) 'BLAST+: architecture and applications'. *BMC Bioinformatics*, **10**(421), pp.1-9.

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J. and Trapnell, C. (2019) 'The single-cell transcriptional landscape of mammalian organogenesis'. *Nature*, **566**(7745), pp. 496-502.

Cao, M., Zhang, M., Yang, N., Fu, Q., Su, B., Zhang, X., Li, Q., Yan, X., Thongda, W. and Li, C. (2020) 'Full length transcriptome profiling reveals novel immune-related genes in black rockfish (*Sebastes schlegelii*)'. *Fish & Shellfish Immunology*, **106**, pp. 1078-1086.

Carlson, M. and Pagès, H. (2024). 'AnnotationForge: Tools for building SQLite-based annotation data packages'. R package version 1.46.0, https://bioconductor.org/packages/AnnotationForge

Carow, B., Hauling, T., Qian, X., Kramnik, I., Nilsson, M. and Rottenberg, M.E. (2019) 'Spatial and temporal localization of immune transcripts defines hallmarks and diversity in the tuberculosis granuloma'. *Nature Communications*, **10**(1), 1823.

Carpenter, S., Ricci, E.P., Mercier, B.C., Moore, M.J. and Fitzgerald, K.A. (2014) 'Post-transcriptional regulation of gene expression in innate immunity'. *Nature Reviews Immunology*, **14**(6), pp. 361-376.

Casola, C. and Betrán, E. (2017) 'The genomic impact of gene retrocopies: what have we learned from comparative genomics, population genomics,

and transcriptomic analyses?'. *Genome Biology and Evolution*, **9**(6), pp. 1351-1373.

Castaldi, P.J., Abood, A., Farber, C.R. and Sheynkman, G.M. (2022) 'Bridging the splicing gap in human genetics with long-read RNA sequencing: finding the protein isoform drivers of disease'. *Human Molecular Genetics*, **31**(R1), pp. R123-R136.

Castro, R., Abós, B., Pignatelli, J., von Gersdorff Jørgensen, L., González Granja, A., Buchmann, K. and Tafalla, C. (2014) 'Early Immune Responses in Rainbow Trout Liver upon Viral Hemorrhagic Septicemia Virus (VHSV) Infection'. *PLoS ONE*, **9**(10), e111084.

Casuso, A., Valenzuela-Muñoz, V. and Gallardo-Escárate, C. (2022) 'Dual RNA-Seq Analysis Reveals Transcriptome Effects during the Salmon–Louse Interaction in Fish Immunized with Three Lice Vaccines'. *Vaccines*, **10**(11), 1875.

Cenik, E.S., Meng, X., Tang, N.H., Hall, R.N., Arribere, J.A., Cenik, C., Jin, Y. and Fire, A. (2019) 'Maternal ribosomes are sufficient for tissue diversification during embryonic development in *C. elegans'*. *Developmental Cell*, **48**(6), pp. 811-826.

Cheetham, S.W., Faulkner, G.J. and Dinger, M.E. (2020) 'Overcoming challenges and dogmas to understand the functions of pseudogenes'. *Nature Reviews Genetics*, **21**(3), pp. 191-201.

Chen, S.N., Zou, P.F. and Nie, P. (2017) 'Retinoic acid-inducible gene I (RIG-I)-like receptors (RLR s) in fish: current knowledge and future perspectives'. *Immunology*, **151**(1), pp. 16-25.

Chen, Y., Chen, L., Lun, A.T., Baldoni, P.L. and Smyth, G.K. (2024) 'edgeR 4.0: powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets'. *bioRxiv*, doi: 10.1101/2024.01.21.576131.

Cheng, G.F., Kong, W.G., Zhai, X., Mu, Q.J., Dong, Z.R., Zhan, M.T. and Xu, Z. (2021) 'Molecular cloning and expression analysis of CD79a and CD79b in rainbow trout (*Oncorhynchus mykiss*) after bacterial, parasitic, and viral infection'. *Fish & Shellfish Immunology*, **118**, pp. 385-395.

Ciomborowska-Basheer, J., Staszak, K., Kubiak, M.R. and Makałowska, I. (2021) 'Not so dead genes—retrocopies as regulators of their disease-related progenitors and hosts'. *Cells*, **10**(4), 912.

Clark, E.L., Archibald, A.L., Daetwyler, H.D., Groenen, M.A., Harrison, P.W., Houston, R.D., Kühn, C., Lien, S., Macqueen, D.J., Reecy, J.M. and Robledo, D. (2020) 'From FAANG to fork: application of highly annotated

genomes to improve farmed animal production'. *Genome Biology*, **21**(285).

Clark, T.C., Tinsley, J., Macqueen, D.J. and Martin, S.A.M. (2019) 'Rainbow trout (*Oncorhynchus mykiss*) urea cycle and polyamine synthesis gene families show dynamic expression responses to inflammation'. *Fish & Shellfish Immunology*, **89**, pp. 290-300.

Clark, T.C., Naseer, S., Gundappa, M.K., Laurent, A., Perquis, A., Collet, B., Macqueen, D.J., Martin, S.A. and Boudinot, P. (2023) 'Conserved and divergent arms of the antiviral response in the duplicated genomes of salmonid fishes'. *Genomics*, **115**(4), 110663.

Cleveland, B.M., Yamaguchi, G., Radler, L.M. and Shimizu, M. (2018) 'Editing the duplicated insulin-like growth factor binding protein-2b gene in rainbow trout (*Oncorhynchus mykiss*)'. *Scientific Reports*, **8**(1), 16054.

Cobb, M. (2017) '60 years ago, Francis Crick changed the logic of biology'. PLoS Biology, **15**(9), e2003243.

Colgan, T.J., Moran, P.A., Archer, L.C., Wynne, R., Hutton, S.A., McGinnity, P. and Reed, T.E. (2021) 'Evolution and Expression of the Immune System of a Facultatively Anadromous Salmonid'. *Frontiers in Immunology*, **12**, 568729.

Collet, B. (2014) 'Innate immune responses of salmonid fish to viral infections'. *Developmental & Comparative Immunology*, **43**(2), pp. 160-173.

Collins, C., Lester, K., Del-Pozo, J. and Collet, B. (2021) 'Non-Lethal Sequential Individual Monitoring of Viremia in Relation to DNA Vaccination in Fish–Example Using a Salmon Alphavirus DNA Vaccine in Atlantic Salmon *Salmo salar'. Vaccines*, **9**(2), 163.

Collins, C.M., Olstad, K., Sterud, E., Jones, C.S., Noble, L.R., Mo, T.A. and Cunningham, C.O. (2007) 'Isolation of a novel fish thymidylate kinase gene, upregulated in Atlantic salmon (*Salmo salar* L.) following infection with the monogenean parasite *Gyrodactylus salaris'*. *Fish & Shellfish Immunology*, **23**(4), pp. 793-807.

Conte, M.A., Gammerdinger, W.J., Bartie, K.L., Penman, D.J. and Kocher, T.D. (2017) 'A high quality assembly of the Nile Tilapia (*Oreochromis niloticus)* genome reveals the structure of two sex determination regions'. *BMC Genomics*, **18**(341).

Conway, J.R., Lex, A. and Gehlenborg, N, (2017) 'UpSetR: an R package for the visualization of intersecting sets and their properties'. *Bioinformatics*, **33**(18), pp. 2938-2940.

Corley, S.M., Troy, N.M., Bosco, A. and Wilkins, M.R. (2019) 'QuantSeq. 3′ Sequencing combined with Salmon provides a fast, reliable approach for high throughput RNA expression analysis'. *Scientific Reports*, **9**(1), 18895.

Corrigall, V.M., Bodman-Smith, M.D., Brunst, M., Cornell, H. and Panayi, G.S. (2004) 'Inhibition of antigen-presenting cell function and stimulation of human peripheral blood mononuclear cells to express an antiinflammatory cytokine profile by the stress protein BiP: relevance to the treatment of inflammatory arthritis'. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, **50**(4), pp. 1164-1171.

Crick, F. (1970) 'Central Dogma of Molecular Biology'. *Nature*, **227**(5258), pp. 561-563.

da Fonseca, R.R., Albrechtsen, A., Themudo, G.E., Ramos-Madrigal, J., Sibbesen, J.A., Maretty, L., Zepeda-Mendoza, M.L., Campos, P.F., Heller, R. and Pereira, R.J. (2016) 'Next-generation biology: sequencing and data analysis approaches for non-model organisms'. *Marine Genomics*, **30**, pp. 3-13.

Dai, Z., Li, J., Hu, C., Wang, F., Wang, B., Shi, X., Hou, Q., Huang, W. and Lin, G. (2017) 'Transcriptome data analysis of grass carp (*Ctenopharyngodon idella*) infected by reovirus provides insights into two immune-related genes'. *Fish & Shellfish Immunology*, **64**, pp. 68-77.

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M. and Li, H. (2021) 'Twelve years of SAMtools and BCFtools'. *Gigascience*, **10**(2), giab008.

Daniels, R.R., Taylor, R.S., Robledo, D. and Macqueen, D.J. (2023) 'Single cell genomics as a transformative approach for aquaculture research and innovation'. *Reviews in Aquaculture*, **15**(4), pp. 1618-1637.

Darwin Tree of Life Project Consortium (2022) 'Sequence locally, think globally: the Darwin Tree of Life Project'. *Proceedings of the National Academy of Sciences*, **119**(4), e2115642118.

Das, S., Jena, S. and Levasseur, D.N. (2011) 'Alternative splicing produces Nanog protein variants with different capacities for self-renewal and pluripotency in embryonic stem cells'. *Journal of Biological Chemistry*, **286**(49), pp. 42690-42703.

Davidson, W.S., Koop, B.F., Jones, S.J., Iturra, P., Vidal, R., Maass, A., Jonassen, I., Lien, S. and Omholt, S.W. (2010) 'Sequencing the genome of the Atlantic salmon (*Salmo salar*)'. *Genome Biology*, **11**(403).

De Coster, W., D'hert, S., Schultz, D.T., Cruts, M. and Van Broeckhoven, C. (2018) 'NanoPack: visualizing and processing long-read sequencing data'. *Bioinformatics*, **34**(15), pp. 2666-2669.

de Klerk, E. and AC't Hoen, P. (2015) 'Alternative mRNA transcription, processing, and translation: insights from RNA sequencing'. *Trends in Genetics*, **31**(3), pp. 128-139.

Dehal, P. and Boore, J.L. (2005) 'Two rounds of whole genome duplication in the ancestral vertebrate'. *PLoS Biology*, **3**(10), e314.

Dehler, C.E., Lester, K., Della Pelle, G., Jouneau, L., Houel, A., Collins, C., Dovgan, T., Machat, R., Zou, J., Boudinot, P. and Martin, S.A. (2019) 'Viral resistance and IFN signaling in STAT2 knockout fish cells'. *The Journal of Immunology*, **203**(2), pp.465-475.

Dehler, C.E., Boudinot, P., Collet, B. and Martin, S.M. (2023) 'Phylogeny and expression of tetraspanin CD9 paralogues in rainbow trout (*Oncorhynchus mykiss*)'. *Developmental & Comparative Immunology*, **146**, 104735.

Dennis, M.Y., Nuttle, X., Sudmant, P.H., Antonacci, F., Graves, T.A., Nefedov, M., Rosenfeld, J.A., Sajjadian, S., Malig, M., Kotkiewicz, H. and Curry, C.J. (2012) 'Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication'. *Cell*, **149**(4), pp. 912-922.

Depledge, D.P., Srinivas, K.P., Sadaoka, T., Bready, D., Mori, Y., Placantonakis, D.G., Mohr, I. and Wilson, A.C. (2019) 'Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen'. *Nature Communications*, **10**(1), p. 754.

Dequéant, M.L. and Pourquié, O. (2008) 'Segmental patterning of the vertebrate embryonic axis'. *Nature Reviews Genetics*, **9**(5), pp. 370-382.

Deschamps-Francoeur, G., Simoneau, J. and Scott, M.S. (2020) 'Handling multi-mapped reads in RNA-seq'. *Computational and Structural Biotechnology Journal*, **18**, pp. 1569-1576.

Des Marais, D.L. and Rausher, M.D. (2008) 'Escape from adaptive conflict after duplication in an anthocyanin pathway gene'. *Nature*, **454**(7205), pp. 762-765.

Di Cara, F., Savary, S., Kovacs, W.J., Kim, P. and Rachubinski, R.A. (2023) 'The peroxisome: an up-and-coming organelle in immunometabolism'. *Trends in Cell Biology*, **33**(1), pp. 70-86.

Dohm, J.C., Peters, P., Stralis-Pavese, N. and Himmelbauer, H. (2020) 'Benchmarking of long-read correction methods'. *NAR Genomics and Bioinformatics*, **2**(2), p.lqaa037.

Dong, X., Tian, L., Gouil, Q., Kariyawasam, H., Su, S., De Paoli-Iseppi, R., Prawer, Y.D.J., Clark, M.B., Breslin, K., Iminitoff, M. and Blewitt, M.E. (2021) 'The long and the short of it: unlocking nanopore long-read RNA sequencing data with short-read differential expression analysis tools'. *NAR Genomics and Bioinformatics*, **3**(2), lqab028.

Dong, X., Du, M.R., Gouil, Q., Tian, L., Jabbari, J.S., Bowden, R., Baldoni, P.L., Chen, Y., Smyth, G.K., Amarasinghe, S.L. and Law, C.W. (2023) 'Benchmarking long-read RNA-sequencing analysis tools using in silico mixtures'. *Nature Methods*, **20**(11), pp. 1810-1821.

Doudna, J.A. and Charpentier, E. (2014) 'The new frontier of genome engineering with CRISPR-Cas9'. *Science*, **346**(6213), 1258096.

Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G. and Dahl, F. (2010) 'Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays'. *Science*, **327**(5961), pp. 78-81.

Drouin, M., Saenz, J. and Chiffoleau, E. (2020) 'C-type lectin-like receptors: head or tail in cell death immunity'. *Frontiers in Immunology*, **11**, 251.

Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C., Shamim, M.S., Machol, I., Lander, E.S., Aiden, A.P. and Aiden, E.L. (2017) 'De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds'. *Science*, **356**(6333), pp. 92-95.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A. and Huber, W. (2005) 'BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis'. *Bioinformatics*, **21**(16), pp. 3439-3440.

Durinck, S., Spellman, P.T., Birney, E. and Huber, W. (2009) 'Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt'. *Nature Protocols*, **4**(8), pp. 1184-1191.

Elbarbary, R.A., Lucas, B.A. and Maquat, L.E. (2016) 'Retrotransposons as regulators of gene expression'. *Science*, **351**(6274), aac7247.

Elsafadi, M., Manikandan, M., Dawud, R.A., Alajez, N.M., Hamam, R., Alfayez, M., Kassem, M., Aldahmash, A. and Mahmood, A. (2016) 'Transgelin is a TGFβ-inducible gene that regulates osteoblastic and adipogenic differentiation of human skeletal stem cells through actin cytoskeleston organization'. *Cell Death & Disease*, **7**(8), e2321.

Emam, M., Caballero-Solares, A., Xue, X., Umasuthan, N., Milligan, B., Taylor, R.G., Balder, R. and Rise, M.L. (2022) 'Gill and liver transcript expression changes associated with gill damage in Atlantic salmon (*Salmo salar*)'. *Frontiers in Immunology*, **13**, 806484.

The ENCODE Project Consortium. (2007). ‘Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project’. *Nature,* **447**, pp. 799–816.

The ENCODE Project Consortium., Snyder, M.P., Gingeras, T.R., Moore, J.E, Weng, Z., Gerstein, M.B., Ren, B., Hardison, R.C., Stamatoyannopoulos, J.A., Graveley, B.R., Feingold, E.A., Pazin, M., Pagan, M., Gilchrist, D.A., Hitz, B.C., Cherry, J.M., Bernstein, B.E., Mendenhall, E.M., Zerbino, D.R., Frankish, A., Flicek, P. and Myers, R.M. (2020) ‘Perspectives on ENCODE’. *Nature,* **583**, pp. 693-698

The ENCODE Project Consortium. (2025) *Long Read RNA-seq Data Standards and Processing Pipeline.* Available at: https://www.encodeproject.org/rna-seq/long-read-rna-seq/ (Accessed: 26 January 2025).

Ergun, A., Doran, G., Costello, J.C., Paik, H.H., Collins, J.J., Mathis, D., Benoist, C., ImmGen Consortium, Blair, D.A., Dustin, M.L. and Shinton, S.A. (2013) ‘Differential splicing across immune system lineages’. *Proceedings of the National Academy of Sciences*, **110**(35), pp. 14324-14329.

Eslamloo, K., Caballero-Solares, A., Inkpen, S.M., Emam, M., Kumar, S., Bouniot, C., Avendaño-Herrera, R., Jakob, E. and Rise, M.L. (2020) ‘Transcriptomic profiling of the adaptive and innate immune responses of Atlantic salmon to *Renibacterium salmoninarum* infection’. *Frontiers in Immunology*, **11**, 567838.

Evans, C., Hardin, J. and Stoebel, D.M. (2018) ‘Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions’. *Briefings in Bioinformatics*, **19**(5), pp.776-792.

Ewart, K.V., Johnson, S.C. and Ross, N.W. (1999) ‘Identification of a pathogen-binding lectin in salmon serum’. *Comparative Biochemistry and Physiology Part C: Pharmacology, Toxicology and Endocrinology*, **123**(1), pp. 9-15.

FAO. (2024a). *The State of World Fisheries and Aquaculture 2024 - Blue Transformation in Action*. Rome.

FAO. (2024b) *‘FishStatJ - Software for Fishery and Aquaculture Statistical Time Series’*. FAO Fisheries Division, Rome. [online] Software v4.04.00 (April 2024). Available at: http://www.fao.org/fishery/statistics/software/fishstatj/en (Accessed: 23 May 2024).

Farrell, J.A., Wang, Y., Riesenfeld, S.J., Shekhar, K., Regev, A. and Schier, A.F. (2018) 'Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis'. *Science*, **360**(6392), eaar3131.

Fast, M.D., Sims, D.E., Burka, J.F., Mustafa, A. and Ross, N.W. (2002) 'Skin morphology and humoral non-specific defence parameters of mucus and plasma in rainbow trout, coho and Atlantic salmon'. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*, **132**(3), pp. 645-657.

Ferrari, K.J., Scelfo, A., Jammula, S., Cuomo, A., Barozzi, I., Stützer, A., Fischle, W., Bonaldi, T. and Pasini, D. (2014) 'Polycomb-dependent H3K27me1 and H3K27me2 regulate active transcription and enhancer fidelity'. *Molecular Cell*, **53**(1), pp. 49-62.

Flajnik, M.F. (2018) 'A cold-blooded view of adaptive immunity'. *Nature Reviews Immunology*, **18**(7), pp. 438-453.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L. and Postlethwait, J. (1999) 'Preservation of duplicate genes by complementary, degenerative mutations'. *Genetics*, **151**(4), pp. 1531-1545.

Fortier, M.E., Kent, S., Ashdown, H., Poole, S., Boksa, P. and Luheshi, G.N. (2004) 'The viral mimic, polyinosinic: polycytidylic acid, induces fever in rats via an interleukin-1-dependent mechanism'. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, **287**(4), pp. R759-R766.

Frankiw, L., Baltimore, D. and Li, G. (2019) 'Alternative mRNA splicing in cancer immunotherapy'. *Nature Reviews Immunology*, **19**(11), pp. 675-687.

Frans, I., Michiels, C.W., Bossier, P., Willems, K.A., Lievens, B. and Rediers, H. (2011) '*Vibrio anguillarum* as a fish pathogen: virulence factors, diagnosis and prevention'. *Journal of Fish Diseases*, **34**(9), pp. 643-661.

Fraslin, C., Brard-Fudulea, S., d'Ambrosio, J., Bestin, A., Charles, M., Haffray, P., Quillet, E. and Phocas, F. (2019) 'Rainbow trout resistance to bacterial cold water disease: two new quantitative trait loci identified after a natural disease outbreak on a French farm'. *Animal Genetics*, **50**(3), pp. 293-297.

Fu, B., Xiong, Y., Sha, Z., Xue, W., Xu, B., Tan, S., Guo, D., Lin, F., Wang, L., Ji, J. and Luo, Y. (2023) 'SEPTIN2 suppresses an IFN-γ-independent, proinflammatory macrophage activation pathway'. *Nature Communications*, **14**(1), 7441.

Fu, L., Niu, B., Zhu, Z., Wu, S. and Li, W. (2012) 'CD-HIT: accelerated for clustering the next-generation sequencing data'. *Bioinformatics*, **28**(23), pp. 3150-3152.

Fu, Q., Li, Y., Zhang, H., Cao, M., Zhang, L., Gao, C., Cai, X., Chen, D., Yang, Z., Li, J. and Yang, N. (2022) 'Comparative transcriptome analysis of spleen reveals potential regulation of genes and immune pathways following administration of *Aeromonas salmonicida* subsp. masoucida vaccine in Atlantic Salmon (*Salmo salar*)'. *Marine Biotechnology*, **24**(1), pp. 97-115.

Fu, Y., Dominissini, D., Rechavi, G. and He, C. (2014) 'Gene expression regulation mediated through reversible m6A RNA methylation'. *Nature Reviews Genetics*, **15**(5), pp. 293-306.

Gallardo-Escarate, C., Valenzuela-Munoz, V. and Nunez-Acuna, G. (2014) 'RNA-Seq analysis using de novo transcriptome assembly as a reference for the salmon louse *Caligus rogercresseyi*'. *PLoS ONE*, **9**(4), e92239.

Gan, Z., Chen, S.N., Huang, B., Zou, J. and Nie, P. (2020) 'Fish type I and type II interferons: composition, receptor usage, production and function'. *Reviews in Aquaculture*, **12**(2), pp. 773-804.

Garalde, D.R., Snell, E.A., Jachimowicz, D., Sipos, B., Lloyd, J.H., Bruce, M., Pantic, N., Admassu, T., James, P., Warland, A. and Jordan, M. (2018) 'Highly parallel direct RNA sequencing on an array of nanopores'. *Nature Methods*, **15**(3), pp. 201-206.

Garcia de la Serrana, D. and Macqueen, D.J. (2018) 'Insulin-like growth factor-binding proteins of teleost fishes'. *Frontiers in Endocrinology*, **9**, 80.

Garnier, S., Ross, N., Rudis, R., Camargo, P.A., Sciaini, M. and Scherer, C. (2024) 'viridis(Lite) - Colorblind-Friendly Color Maps for R'. R package version 0.6.5, https://sjmgarnier.github.io/viridis/, https://cran.r-project.org/package=viridisLite.

Garseth, Å.H., Fritsvold, C., Svendsen, J.C., Bang Jensen, B. and Mikalsen, A.B. (2018) 'Cardiomyopathy syndrome in Atlantic salmon *Salmo salar* L.: A review of the current state of knowledge'. *Journal of Fish Diseases*, **41**(1), pp. 11-26.

Gehlenborg, N. (2019) 'UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets'. R package version 1.4.0, https://CRAN.R-project.org/package=UpSetR

Gervais, O., Barria, A., Papadopoulou, A., Gratacap, R.L., Hillestad, B., Tinch, A.E., Martin, S.A.M., Robledo, D. and Houston, R.D. (2021) 'Exploring genetic resistance to infectious salmon anaemia virus in

Atlantic salmon by genome-wide association and RNA sequencing'. *BMC Genomics*, **22**(345).

Gharbi, K., Matthews, L., Bron, J., Roberts, R., Tinch, A. and Stear, M. (2015) 'The control of sea lice in Atlantic salmon by selective breeding'. *Journal of the Royal Society Interface*, **12**(110), 20150574.

Ghosh, S. and Marsh, E.N.G. (2020). 'Viperin: An ancient radical SAM enzyme finds its place in modern cellular metabolism and innate immunity'. *Journal of Biological Chemistry*, **295**(33), pp. 11513-11528.

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. and Rohlfs, R.V. (2021) 'Comparative regulomics supports pervasive selection on gene dosage following whole genome duplication'. *Genome Biology*, **22**(103).

Gilmore, T.D. and Wolenski, F.S. (2012) 'NF-κB: where did it come from and why?'. *Immunological Reviews*, **246**(1), pp. 14-35.

Giuffra, E., Tuggle, C.K. and FAANG Consortium. (2019) 'Functional annotation of animal genomes (FAANG): current achievements and roadmap'. *Annual Review of Animal Biosciences*, **7**, pp. 65-88.

Gjedrem, T., Robinson, N. and Rye, M. (2012) 'The importance of selective breeding in aquaculture to meet future demands for animal protein: a review'. *Aquaculture*, **350**, pp. 117-129.

Gjedrem, T. and Rye, M. (2018) 'Selection response in fish and shellfish: a review'. Reviews in Aquaculture, **10**(1), pp. 168-179.

Glasauer, S.M. and Neuhauss, S.C. (2014) 'Whole-genome duplication in teleost fishes and its evolutionary consequences'. *Molecular Genetics and Genomics*, **289**, pp. 1045-1060.

Glinos, D.A., Garborcauskas, G., Hoffman, P., Ehsan, N., Jiang, L., Gokden, A., Dai, X., Aguet, F., Brown, K.L., Garimella, K. and Bowers, T. (2022) 'Transcriptome variation in human tissues revealed by long-read sequencing'. *Nature*, **608**(7922), pp. 353-359.

Gohel, D. and Skintzos, P. (2024) 'flextable: Functions for Tabular Reporting'. R package version 0.9.6, https://davidgohel.github.io/flextable/, https://ardata-fr.github.io/flextable-book/

Gonen, S., Baranski, M., Thorland, I., Norris, A., Grove, H., Arnesen, P., Bakke, H., Lien, S., Bishop, S.C. and Houston, R.D. (2015) 'Mapping and validation of a major QTL affecting resistance to pancreas disease

(salmonid alphavirus) in Atlantic salmon (*Salmo salar*)'. *Heredity*, **115**(5), pp. 405-414.

Gorodilov, Y.N. (1996) 'Description of the early ontogeny of the Atlantic salmon, *Salmo salar*, with a novel system of interval (state) identification'. *Environmental Biology of Fishes*, **47**, pp. 109-127.

Gratacap, R.L., Wargelius, A., Edvardsen, R.B. and Houston, R.D. (2019) 'Potential of genome editing to improve aquaculture breeding and production'. *Trends in Genetics*, **35**(9), pp. 672-684.

Graveley, B.R., Brooks, A.N., Carlson, J.W., Duff, M.O., Landolin, J.M., Yang, L., Artieri, C.G., Van Baren, M.J., Boley, N., Booth, B.W. and Brown, J.B. (2011) 'The developmental transcriptome of *Drosophila melanogaster'. Nature*, **471**(7339), pp. 473-479.

Grayfer, L., Hodgkinson, J.W. and Belosevic, M. (2014) 'Antimicrobial responses of teleost phagocytes and innate immune evasion strategies of intracellular bacteria'. *Developmental & Comparative Immunology*, **43**(2), pp. 223-242.

Grayfer, L., Kerimoglu, B., Yaparla, A., Hodgkinson, J.W., Xie, J. and Belosevic, M. (2018) 'Mechanisms of fish macrophage antimicrobial immunity'. *Frontiers in Immunology*, **9**, 1105.

Grimholt, U., Hauge, H., Hauge, A.G., Leong, J. and Koop, B.F. (2015) 'Chemokine receptors in Atlantic salmon'. *Developmental & Comparative Immunology*, **49**(1), pp. 79-95.

Gu, Z. (2022) 'Complex heatmap visualization'. *Imeta*, **1**(3), e43.

Guan, D., Halstead, M.M., Islas-Trejo, A.D., Goszczynski, D.E., Cheng, H.H., Ross, P.J. and Zhou, H. (2022) 'Prediction of transcript isoforms in 19 chicken tissues by Oxford Nanopore long-read sequencing'. *Frontiers in Genetics*, **13**, 997460.

Gundappa, M.K., To, T.H., Grønvold, L., Martin, S.A., Lien, S., Geist, J., Hazlerigg, D., Sandve, S.R. and Macqueen, D.J. (2022) 'Genome-wide reconstruction of rediploidization following autopolyploidization across one hundred million years of salmonid evolution'. *Molecular Biology and Evolution*, **39**(1), msab310.

Gunter, H.M., Idrisoglu, S., Singh, S., Han, D.J., Ariens, E., Peters, J.R., Wong, T., Cheetham, S.W., Xu, J., Rai, S.K. and Feldman, R. (2023) 'mRNA vaccine quality analysis using RNA sequencing'. *Nature Communications*, **14**(1), 5663.

Gustavsson, E.K., Zhang, D., Reynolds, R.H., Garcia-Ruiz, S. and Ryten, M. (2022) 'ggtranscript: an R package for the visualization and interpretation

of transcript isoforms using ggplot2'. *Bioinformatics*, **38**(15), pp. 3844-3846.

Gutierrez, A.P., Lubieniecki, K.P., Davidson, E.A., Lien, S., Kent, M.P., Fukui, S., Withler, R.E., Swift, B. and Davidson, W.S. (2012) 'Genetic mapping of quantitative trait loci (QTL) for body-weight in Atlantic salmon (*Salmo salar*) using a 6.5 K SNP array'. *Aquaculture*, **358**, pp. 61-70.

Haas, B.J. (2023) *TransDecoder* (Version 5.7.1) [Computer program]. Available at: https://github.com/TransDecoder/TransDecoder.

Haberle, V. and Stark, A. (2018) 'Eukaryotic core promoters and the functional basis of transcription initiation'. *Nature Reviews Molecular Cell Biology*, **19**(10), pp. 621-637.

Haidle, L., Janssen, J.E., Gharbi, K., Moghadam, H.K., Ferguson, M.M. and Danzmann, R.G. (2008) 'Determination of quantitative trait loci (QTL) for early maturation in rainbow trout (*Oncorhynchus mykiss*)'. *Marine Biotechnology*, **10**, pp. 579-592.

Hall, T.A. (1999) 'BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT'. *Nucleic Acids Symposium Series*, **41**, pp. 95-98.

Halstead, M.M., Islas-Trejo, A., Goszczynski, D.E., Medrano, J.F., Zhou, H. and Ross, P.J. (2021) 'Large-scale multiplexing permits full-length transcriptome annotation of 32 bovine tissues from a single nanopore flow cell'. *Frontiers in Genetics*, **12**, 664260.

Handeland, S.O., Berge, Å., Björnsson, B.T. and Stefansson, S.O. (1998) 'Effects of temperature and salinity on osmoregulation and growth of Atlantic salmon (*Salmo salar* L.) smolts in seawater'. *Aquaculture*, **168**(1-4), pp. 289-302.

Hansen, S.A.H., Ramberg, S., Lekanova, N., Høyheim, B., Horsberg, T.E., Andreassen, R. and Bakke, M.J. (2023) 'De novo high-accuracy transcriptomes from long-read sequencing reveals a wide variety of novel splice variants in copepodids and adult female salmon lice (*Lepeophtheirus salmonis*)'. *Frontiers in Marine Science*, **10**, 1167402.

Happold, J., Sadler, R., Meyer, A., Hillman, A., Cowled, B., Mackenzie, C., Lagno, A.L.G. and Cameron, A. (2020) 'Effectiveness of vaccination for the control of salmonid rickettsial septicaemia in commercial salmon and trout farms in Chile'. *Aquaculture*, **520**, 734968.

Hardwick, S.A., Joglekar, A., Flicek, P., Frankish, A. and Tilgner, H.U. (2019) 'Getting the entire message: progress in isoform sequencing'. *Frontiers in Genetics*, **10**, 709.

Harte, A., Tian, G., Xu, Q., Secombes, C.J. and Wang, T. (2020) 'Five subfamilies of β-defensin genes are present in salmonids: Evolutionary insights and expression analysis in Atlantic salmon *Salmo salar'. Developmental & Comparative Immunology*, **104**, 103560.

Harvey, T.N., Gillard, G.B., Røsæg, L.L., Grammes, F., Monsen, Ø., Vik, J.O., Hvidsten, T.R. and Sandve, S.R. (2024) 'The genome regulatory landscape of Atlantic salmon liver through smoltification'. *PLoS ONE*, **19**(4), e0302388.

He, Y., Carrillo, J.A., Luo, J., Ding, Y., Tian, F., Davidson, I. and Song, J. (2014) 'Genome-wide mapping of DNase I hypersensitive sites and association analysis with gene expression in MSB1 cells'. *Frontiers in Genetics*, **5**, 308.

Heinz, L.X., Lee, J., Kapoor, U., Kartnig, F., Sedlyarov, V., Papakostas, K., César-Razquin, A., Essletzbichler, P., Goldmann, U., Stefanovic, A. and Bigenzahn, J.W. (2020) 'TASL is the SLC15A4-associated adaptor for IRF5 activation by TLR7–9'. *Nature*, **581**(7808), pp. 316-322.

Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A.J., Searle, S.M., Amode, R., Brent, S. and Spooner, W. (2016) 'Ensembl comparative genomics resources'. *Database*, bav096.

Heyn, P., Kircher, M., Dahl, A., Kelso, J., Tomancak, P., Kalinka, A.T. and Neugebauer, K.M. (2014) 'The earliest transcribed zygotic genes are short, newly evolved, and different across species'. *Cell Reports*, **6**(2), pp. 285-292.

Hjelmen, C.E. (2024) 'Genome size and chromosome number are critical metrics for accurate genome assembly assessment in Eukaryota'. *Genetics*, **227**(4).

Holland, L.Z. and Ocampo Daza, D. (2018) 'A new look at an old question: when did the second whole genome duplication occur in vertebrate evolution?'. *Genome Biology*, **19**(209).

Hori, Y., Engel, C. and Kobayashi, T. (2023) 'Regulation of ribosomal RNA gene copy number, transcription and nucleolus organization in eukaryotes'. *Nature Reviews Molecular Cell Biology*, **24**(6), pp. 414-429.

Houston, R.D., Bishop, S.C., Hamilton, A., Guy, D.R., Tinch, A.E., Taggart, J.B., Derayat, A., McAndrew, B.J. and Haley, C.S. (2009) 'Detection of QTL affecting harvest traits in a commercial Atlantic salmon population'. *Animal Genetics*, **40**(5), pp. 753-755.

Houston, R.D., Haley, C.S., Hamilton, A., Guy, D.R., Mota-Velasco, J.C., Gheyas, A.A., Tinch, A.E., Taggart, J.B., Bron, J.E., Starkey, W.G. and McAndrew, B.J. (2010) 'The susceptibility of Atlantic salmon fry to

freshwater infectious pancreatic necrosis is largely explained by a major QTL'. *Heredity*, **105**(3), pp. 318-327.

Houston, R.D. (2017) 'Future directions in breeding for disease resistance in aquaculture species'. *Revista Brasileira de Zootecnia*, **46**, pp. 545-551.

Houston, R.D. and Macqueen, D.J. (2019) 'Atlantic salmon (*Salmo salar* L.) genetics in the 21st century: taking leaps forward in aquaculture and biological understanding'. *Animal Genetics*, **50**(1), pp. 3-14.

Houston, R.D., Bean, T.P., Macqueen, D.J., Gundappa, M.K., Jin, Y.H., Jenkins, T.L., Selly, S.L.C., Martin, S.A., Stevens, J.R., Santos, E.M. and Davie, A. (2020) 'Harnessing genomics to fast-track genetic improvement in aquaculture'. *Nature Reviews Genetics*, **21**(7), pp. 389-409.

Hrdlickova, R., Toloue, M. and Tian, B. (2017) 'RNA-Seq methods for transcriptome analysis'. *Wiley Interdisciplinary Reviews: RNA*, **8**(1), e1364.

Hsieh, T.B. and Jin, J.P. (2023) 'Evolution and function of calponin and transgelin'. *Frontiers in Cell and Developmental Biology*, **11**, 1206147.

Hu, J., Fan, J., Sun, Z. and Liu, S. (2020) 'NextPolish: a fast and efficient genome polishing tool for long-read assembly'. *Bioinformatics*, **36**(7), pp.2253-2255.

Hu, Y., Fang, L., Chen, X., Zhong, J.F., Li, M. and Wang, K. (2021) 'LIQA: long-read isoform quantification and analysis'. *Genome Biology*, **22**(182).

Huang, S. (2009) 'Non-genetic heterogeneity of cells in development: more than just noise'. *Development*, **136**(23), pp. 3853-3862.

Huang, J., Chen, W., Wang, Q., Zhang, Y., Liu, Q. and Yang, D. (2022) 'Iso-Seq assembly and functional annotation of full-length transcriptome of turbot (*Scophthalmus maximus*) during bacterial infection'. *Marine Genomics*, **63**, 100954.

Huang, Y.T., Liu, P.Y. and Shih, P.W. (2021) 'Homopolish: a method for the removal of systematic errors in nanopore sequencing by homologous polishing'. *Genome Biology*, **22**(95).

Hubbard, T., Andrews, D., Cáccamo, M., Cameron, G., Chen, Y., Clamp, M., Clarke, L., Coates, G., Cox, T., Cunningham, F. and Curwen, V. (2005) 'Ensembl 2005'. *Nucleic Acids Research*, **33**(suppl_1), pp. D447-D453.

Hurley, I.A., Mueller, R.L., Dunn, K.A., Schmidt, E.J., Friedman, M., Ho, R.K., Prince, V.E., Yang, Z., Thomas, M.G. and Coates, M.I. (2007) 'A new time-scale for ray-finned fish evolution'. *Proceedings of the Royal Society B: Biological Sciences*, **274**(1609), pp. 489-498.

Hwang, B., Lee, J.H. and Bang, D. (2018) 'Single-cell RNA sequencing technologies and bioinformatics pipelines'. *Experimental & Molecular Medicine*, **50**(8), pp. 1-14.

Illumina. (2024) *NovaSeq X Series.* Available at: https://emea.illumina.com/systems/sequencing-platforms/novaseq-x-plus.html (Accessed: 13 September 2024).

Inamo, J., Suzuki, A., Ueda, M.T., Yamaguchi, K., Nishida, H., Suzuki, K., Kaneko, Y., Takeuchi, T., Hatano, H., Ishigaki, K. and Ishihama, Y. (2024) 'Long-read sequencing for 29 immune cell subsets reveals disease-linked isoforms'. *Nature Communications*, **15**(1), 4285.

Iñiguez, L.P. and Hernández, G. (2017) 'The evolutionary relationship between alternative splicing and gene duplication'. *Frontiers in Genetics*, **8**, 14.

International Human Genome Sequencing Consortium. (2004) 'Finishing the euchromatic sequence of the human genome'. *Nature*, **431**(7011), pp. 931-945.

Irie, N. and Kuratani, S. (2011) 'Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis'. *Nature Communications*, **2**(1), p. 248.

Jain, M., Fiddes, I.T., Miga, K.H., Olsen, H.E., Paten, B. and Akeson, M. (2015) 'Improved data analysis for the MinION nanopore sequencer'. *Nature Methods*, **12**(4), pp. 351-356.

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. and Malla, S. (2018) 'Nanopore sequencing and assembly of a human genome with ultra-long reads'. *Nature Biotechnology*, **36**(4), pp. 338-345.

Jansen, M.D., Bang Jensen, B., McLoughlin, M.F., Rodger, H.D., Taksdal, T., Sindre, H., Graham, D.A. and Lillehaug, A. (2017) 'The epidemiology of pancreas disease in salmonid aquaculture: a summary of the current state of knowledge'. *Journal of Fish Diseases*, **40**(1), pp. 141-155.

Jathar, S., Kumar, V., Srivastava, J., Tripathi, V. (2017) 'Technological Developments in lncRNA Biology', in Rao, M. (ed) *Long Non Coding RNA Biology. Advances in Experimental Medicine and Biology*. Singapore: Springer, Singapore, vol 1008

Jiang, Y.J., Aerne, B.L., Smithers, L., Haddon, C., Ish-Horowicz, D. and Lewis, J. (2000) 'Notch signalling and the synchronization of the somite segmentation clock'. *Nature*, **408**(6811), pp. 475-479.

Jiao, W., Chen, Y., Song, H., Li, D., Mei, H., Yang, F., Fang, E., Wang, X., Huang, K., Zheng, L. and Tong, Q. (2018) 'HPSE enhancer RNA promotes cancer progression through driving chromatin looping and regulating hnRNPU/p300/EGR1/HPSE axis'. *Oncogene*, **37**(20), pp. 2728-2745.

Johnston, I.A., Lee, H.T., Macqueen, D.J., Paranthaman, K., Kawashima, C., Anwar, A., Kinghorn, J.R. and Dalmay, T. (2009) 'Embryonic temperature affects muscle fibre recruitment in adult zebrafish: genome-wide changes in gene and microRNA expression associated with the transition from hyperplastic to hypertrophic growth phenotypes'. *Journal of Experimental Biology*, **212**(12), pp. 1781-1793.

Johnston, I.A., Kent, M.P., Boudinot, P., Looseley, M., Bargelloni, L., Faggion, S., Merino, G.A., Ilsley, G.R., Bobe, J., Tsigenopoulos, C.S. and Robertson, J. (2024) 'Advancing fish breeding in aquaculture through genome functional annotation'. *Aquaculture*, **583**, 740589.

Jorquera, R., González, C., Clausen, P.T.L.C., Petersen, B. and Holmes, D.S. (2021) 'SinEx DB 2.0 update 2020: database for eukaryotic single-exon coding sequences'. *Database*, **2021**, p.baab002.

Jukam, D., Shariati, S.A.M. and Skotheim, J.M. (2017) 'Zygotic genome activation in vertebrates'. *Developmental Cell*, **42**(4), pp. 316-332.

Junker, J.P., Noel, E.S., Guryev, V., Peterson, K.A., Shah, G., Huisken, J., McMahon, A.P., Berezikov, E., Bakkers, J. and van Oudenaarden, A. (2014) 'Genome-wide RNA tomography in the zebrafish embryo'. *Cell*, **159**(3), pp. 662-675.

Kadobianskyi, M., Schulze, L., Schuelke, M. and Judkewitz, B. (2019) 'Hybrid genome assembly and annotation of *Danionella translucida*'. *Scientific Data*, **6**(1), 156.

Kaessmann, H., Vinckenbosch, N. and Long, M. (2009) 'RNA-based gene duplication: mechanistic and evolutionary insights'. *Nature Reviews Genetics*, **10**(1), pp. 19-31.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K., Von Haeseler, A. and Jermiin, L.S. (2017) 'ModelFinder: fast model selection for accurate phylogenetic estimates'. *Nature Methods*, **14**(6), pp. 587-589.

Kane, D.A. and Kimmel, C.B. (1993) 'The zebrafish midblastula transition'. *Development*, **119**(2), pp. 447-456.

Kane, D.A. (1998) 'Cell cycles and development in the embryonic zebrafish'. *Methods in Cell Biology*, **59**, pp. 11-26.

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C.C., Suzuki, M., Kawai, J. and Suzuki, H. (2005) 'Antisense transcription in the mammalian transcriptome'. *Science*, **309**(5740), pp. 1564-1566.

Katoh, K., Rozewicki, J. and Yamada, K.D. (2019) 'MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization'. *Briefings in Bioinformatics*, **20**(4), pp. 1160-1166.

Katzenback, B.A. (2015) 'Antimicrobial peptides as mediators of innate immunity in teleosts'. *Biology*, **4**(4), pp. 607-639.

Kelleher, M., Singh, R., O'Driscoll, C.M. and Melgar, S. (2019) 'Carcinoembryonic antigen (CEACAM) family members and inflammatory bowel disease'. *Cytokine & Growth Factor Reviews*, **47**, pp. 21-31.

Kettle, A.J., Gedye, C.A. and Winterbourn, C.C. (1993) 'Superoxide is an antagonist of anti-inflammatory drugs that inhibit hypochlorous acid production by myeloperoxidase'. *Biochemical Pharmacology*, **45**(10), pp. 2003-2010.

Khan, A.A., Alsahli, M.A. and Rahmani, A.H. (2018) 'Myeloperoxidase as an active disease biomarker: recent biochemical and pathological perspectives'. *Medical Sciences*, **6**(2), p. 33.

Khorkova, O., Myers, A.J., Hsiao, J. and Wahlestedt, C. (2014) 'Natural antisense transcripts'. *Human Molecular Genetics*, **23**(R1), pp. R54-R63.

Kibenge, F.S., Godoy, M.G., Fast, M., Workenhe, S. and Kibenge, M.J. (2012) 'Countermeasures against viral diseases of farmed fish'. *Antiviral Research*, **95**(3), pp. 257-281.

Kileng, Ø., Brundtland, M.I. and Robertsen, B. (2007) 'Infectious salmon anemia virus is a powerful inducer of key genes of the type I interferon system of Atlantic salmon, but is not inhibited by interferon'. *Fish & Shellfish Immunology*, **23**(2), pp. 378-389.

Kim, E.Y., Che, Y., Dean, H.J., Lorenzo-Redondo, R., Stewart, M., Keller, C.K., Whorf, D., Mills, D., Dulin, N.N., Kim, T. and Votoupal, M. (2022) 'Transcriptome-wide changes in gene expression, splicing, and lncRNAs in response to a live attenuated dengue virus vaccine'. *Cell Reports*, **38**(6), 110341.

Kim, H.M., Jeon, S., Chung, O., Jun, J.H., Kim, H.S., Blazyte, A., Lee, H.Y., Yu, Y., Cho, Y.S., Bolser, D.M. and Bhak, J. (2021a) 'Comparative analysis of 7 short-read sequencing platforms using the Korean Reference Genome: MGI and Illumina sequencing benchmark for whole-genome sequencing'. *GigaScience*, **10**(3), giab014.

Kim, H.S., Grimes, S.M., Hooker, A.C., Lau, B.T. and Ji, H.P. (2021b) 'Single-cell characterization of CRISPR-modified transcript isoforms with nanopore sequencing'. *Genome Biology*, **22**(331).

Kim, J.K., Kim, Y.S., Lee, H.M., Jin, H.S., Neupane, C., Kim, S., Lee, S.H., Min, J.J., Sasai, M., Jeong, J.H. and Choe, S.K. (2018) 'GABAergic signaling linked to autophagy enhances host protection against intracellular bacterial infections'. *Nature Communications*, **9**(1), 4184.

Kimmel, C.B., Warga, R.M. and Schilling, T.F. (1990) 'Origin and organization of the zebrafish fate map'. *Development*, **108**(4), pp. 581-594.

Kimmel, C.B., Ballard, W.W., Kimmel, S.R., Ullmann, B. and Schilling, T.F. (1995) 'Stages of embryonic development of the zebrafish'. *Developmental Dynamics*, **203**(3), pp. 253-310.

Kiron, V., Park, Y., Siriyappagouder, P., Dahle, D., Vasanth, G.K., Dias, J., Fernandes, J.M., Sørensen, M. and Trichet, V.V. (2020) 'Intestinal transcriptome analysis reveals soy derivative-linked changes in Atlantic salmon'. *Frontiers in Immunology*, **11**, 596514.

Kleppe, L., Wargelius, A., Johnsen, H., Andersson, E. and Edvardsen, R.B. (2015) 'Gonad specific genes in Atlantic salmon (*Salmon salar* L.): characterization of *tdrd7-2*, *dazl-2*, *piwil1* and *tdrd1* genes'. *Gene*, **560**(2), pp. 217-225.

Ko, R., Seo, J., Park, H., Lee, N. and Lee, S.Y. (2022) 'Pim1 promotes IFN-β production by interacting with IRF3'. *Experimental & Molecular Medicine*, **54**(11), pp. 2092-2103.

Kohonen, T. (1982) 'Self-organized formation of topologically correct feature maps'. *Biological Cybernetics*, **43**(1), pp. 59-69.

Kolde, R. (2019) 'Pheatmap: pretty heatmaps'. R Package Version 1.0.12. https://cran.r-project.org/package=pheatmap

Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C. and Teichmann, S.A. (2015) 'The technology and biology of single-cell RNA sequencing'. *Molecular Cell*, **58**(4), pp. 610-620.

Komal, A., Noreen, M. and El-Kott, A.F. (2021) 'TLR3 agonists: RGC100, ARNAX, and poly-IC: A comparative review'. *Immunologic Research*, **69**(4), pp. 312-322.

Kono, N. and Arakawa, K. (2019) 'Nanopore sequencing: Review of potential applications in functional genomics'. *Development, Growth & Differentiation*, **61**(5), pp. 316-326.

Kortner, T.M., Afanasyev, S., Koppang, E.O., Bjørgen, H., Krogdahl, Å. and Krasnov, A. (2024) 'A comprehensive transcriptional body map of Atlantic salmon unveils the vital role of the intestine in the immune system and highlights functional specialization within its compartments'. *Fish & Shellfish Immunology*, **146**, 109422.

Krasnov, A., Johansen, L.H., Karlsen, C., Sveen, L., Ytteborg, E., Timmerhaus, G., Lazado, C.C. and Afanasyev, S. (2021) 'Transcriptome responses of Atlantic Salmon (*Salmo salar* L.) to viral and bacterial pathogens, inflammation, and stress'*. Frontiers in Immunology*, **12**, 705601.

Kubiak, M.R. and Makałowska, I. (2017) 'Protein-coding genes' retrocopies and their functions'. *Viruses*, **9**(4), 80.

Kukurba, K.R. and Montgomery, S.B. (2015) 'RNA sequencing and analysis'. *Cold Spring Harbor Protocols*, **2015**(11), pdb-top084970.

Kuo, R.I., Tseng, E., Eory, L., Paton, I.R., Archibald, A.L. and Burt, D.W. (2017) 'Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human'. *BMC Genomics*, **18**(323).

Kuo, R.I., Cheng, Y., Zhang, R., Brown, J.W., Smith, J., Archibald, A.L. and Burt, D.W. (2020) 'Illuminating the dark side of the human transcriptome with long read transcript sequencing'. *BMC Genomics*, **21**(751).

Lages, M.A., Balado, M. and Lemos, M.L. (2019) 'The expression of virulence factors in *Vibrio anguillarum* is dually regulated by iron levels and temperature'. *Frontiers in Microbiology*, **10**, 2335.

Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R. and Weirauch, M.T. (2018) 'The human transcription factors'. *Cell*, **172**(4), pp. 650-665.

Lan, X., Witt, H., Katsumura, K., Ye, Z., Wang, Q., Bresnick, E.H., Farnham, P.J. and Jin, V.X. (2012) 'Integration of Hi-C and ChIP-seq data reveals distinct types of chromatin linkages'. *Nucleic Acids Research*, **40**(16), pp. 7690-7704.

Laurent, G.S., Wahlestedt, C. and Kapranov, P. (2015) 'The Landscape of long noncoding RNA classification'. *Trends in Genetics*, **31**(5), pp. 239-251.

Lawrence, M., Gentleman, R. and Carey, V. (2009) 'rtracklayer: an R package for interfacing with genome browsers'. *Bioinformatics*, **25**(14), 1841.

Lebrigand, K., Magnone, V., Barbry, P. and Waldmann, R. (2020) 'High throughput error corrected Nanopore single cell transcriptome sequencing'. *Nature Communications*, **11**(1), 4025.

Lee, H., Zhang, Z. and Krause, H.M. (2019) 'Long noncoding RNAs and repetitive elements: junk or intimate evolutionary partners?'. *TRENDS in Genetics*, **35**(12), pp. 892-902.

Lee, M.T., Bonneau, A.R. and Giraldez, A.J. (2014) 'Zygotic genome activation during the maternal-to-zygotic transition'. *Annual Review of Cell and Developmental Biology*, **30**, pp. 581-613.

Leger, A. and Leonardi, T. (2019). 'pycoQC, interactive quality control for Oxford Nanopore Sequencing'. *Journal of Open Source Software*, **4**(34), 1236.

Leiva, F., Rojas-Herrera, M., Reyes, D., Bravo, S., Garcia, K.K., Moya, J. and Vidal, R. (2020) 'Identification and characterization of miRNAs and lncRNAs of coho salmon (*Oncorhynchus kisutch*) in normal immune organs'. *Genomics*, **112**(1), pp. 45-54.

Levraud, J.P., Jouneau, L., Briolat, V., Laghi, V. and Boudinot, P. (2019) 'IFN-stimulated genes in zebrafish and humans define an ancient arsenal of antiviral immunity'. *The Journal of Immunology*, **203**(12), pp. 3361-3373.

Lewin, H.A., Robinson, G.E., Kress, W.J., Baker, W.J., Coddington, J., Crandall, K.A., Durbin, R., Edwards, S.V., Forest, F., Gilbert, M.T.P. and Goldstein, M.M. (2018) 'Earth BioGenome Project: Sequencing life for the future of life'. *Proceedings of the National Academy of Sciences*, **115**(17), pp. 4325-4333.

Lewin, H.A., Richards, S., Lieberman Aiden, E., Allende, M.L., Archibald, J.M., Bálint, M., Barker, K.B., Baumgartner, B., Belov, K., Bertorelle, G. and Blaxter, M.L. (2022) 'The earth BioGenome project 2020: Starting the clock.' *Proceedings of the National Academy of Sciences*, **119**(4), e2115635118.

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. and Pfister, H. (2014) 'UpSet: visualization of intersecting sets. IEEE transactions on visualization and computer graphics', **20**(12), pp. 1983-1992.

Li, B. and Dewey, C.N. (2011) 'RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome'. *BMC Bioinformatics*, **12**(323).

Li, B.J., Zhu, Z.X., Qin, H., Meng, Z.N., Lin, H.R. and Xia, J.H. (2020a) 'Genome-wide characterization of alternative splicing events and their responses to cold stress in tilapia'. *Frontiers in Genetics*, **11**, 244.

Li, H. (2021) ‘New strategies to improve minimap2 alignment accuracy’. *Bioinformatics*, **37**(23), 4572-4574.

Li, W. and Godzik, A. (2006) ‘Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences’. *Bioinformatics*, **22**(13), pp. 1658-1659.

Li, X., Nair, A., Wang, S., Wang, L. (2015) ‘Quality Control of RNA-Seq Experiments’, in Picardi, E. (eds) *RNA Bioinformatics. Methods in Molecular Biology*. New York: Humana Press, vol 1269.

Li, Y., Liu, Y., Yang, H., Zhang, T., Naruse, K. and Tu, Q. (2020b) ‘Dynamic transcriptional and chromatin accessibility landscape of medaka embryogenesis’. *Genome Research*, **30**(6), pp 924-937.

Liao, K.C. and Garcia-Blanco, M.A. (2021) ‘Role of alternative splicing in regulating host response to viral infection’. *Cells*, **10**(7), 1720.

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. and Grammes, F. (2016) ‘The Atlantic salmon genome provides insights into rediploidization’. *Nature*, **533**(7602), pp. 200-205.

Lin, H., Liang, Z.Y., Tang, H. and Chen, W. (2017) ‘Identifying sigma70 promoters with novel pseudo nucleotide composition’. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **16**(4), pp. 1316-1321.

Lin, J., Guan, L., Ge, L., Liu, G., Bai, Y. and Liu, X. (2021) ‘Nanopore-based full-length transcriptome sequencing of Muscovy duck (*Cairina moschata*) ovary’. *Poultry Science*, **100**(8), 101246.

Liu, C., Zhao, W., Su, J., Chen, X., Zhao, F., Fan, J., Li, X., Liu, X., Zou, L., Zhang, M. and Zhang, Z. (2022a) ‘HSP90AA1 interacts with CSFV NS5A protein and regulates CSFV replication via the JAK/STAT and NF-κB signaling pathway’. *Frontiers in Immunology*, **13**, 1031868.

Liu, F., Wang, T., Petit, J., Forlenza, M., Chen, X., Chen, L., Zou, J. and Secombes, C.J. (2020) ‘Evolution of IFN subgroups in bony fish-2. analysis of subgroup appearance and expansion in teleost fish with a focus on salmonids’. *Fish & Shellfish Immunology*, **98**, pp. 564-573.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L. and Law, M. (2012) ‘Comparison of next-generation sequencing systems’. *BioMed Research International*, **2012**, 251364.

Liu, P., Wang, L., Kwang, J., Yue, G.H. and Wong, S.M. (2016) ‘Transcriptome analysis of genes responding to NNV infection in Asian seabass epithelial cells’. *Fish & Shellfish Immunology*, **54**, pp. 342-352.

Liu, Q., Fang, L. and Wu, C. (2022b) ‘Alternative splicing and isoforms: from mechanisms to diseases’. *Genes*, **13**(3), 401.

Liu, W., Chen, B., Yao, J., Liu, J., Kuang, M., Wang, F., Wang, Y., Elkady, G., Lu, Y., Zhang, Y. and Liu, X. (2019) ‘Identification of fish CMPK2 as an interferon stimulated gene against SVCV infection’. *Fish & Shellfish Immunology*, **92**, pp. 125-132.

Liu, Z., Wang, W., Li, X., Zhao, X., Zhao, H., Yang, W., Zuo, Y., Cai, L. and Xing, Y. (2022c) ‘Temporal dynamic analysis of alternative splicing during embryonic development in zebrafish’. *Frontiers in Cell and Developmental Biology*, **10**, 879795.

Liu-Wei, W., van der Toorn, W., Bohn, P., Hölzer, M., Smyth, R.P. and von Kleist, M (2024) ‘Sequencing accuracy and systematic errors of nanopore direct RNA sequencing’. *BMC Genomics*, **25**(528).

Løken, O.M., Bjørgen, H., Hordvik, I. and Koppang, E.O. (2020) ‘A teleost structural analogue to the avian bursa of Fabricius’. *Journal of Anatomy*, **236**(5), pp. 798-808.

Love, M.I., Huber, W. and Anders, S. (2014) ‘Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2’. *Genome Biology*, **15**(550).

Love, M.I., Soneson, C. and Patro, R. (2018) ‘Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification’. *F1000Research*, **7**.

Lu, T.X. and Rothenberg, M.E. (2018) ‘MicroRNA’. *Journal of Allergy and Clinical Immunology*, **141**(4), pp. 1202-1207.

Luecken, M.D. and Theis, F.J. (2019) ‘Current best practices in single-cell RNA-seq analysis: a tutorial’. *Molecular Systems Biology*, **15**(6), e8746.

Lulijwa, R., Rupia, E.J. and Alfaro, A.C. (2020) ‘Antibiotic use in aquaculture, policies and regulation, health and environmental risks: a review of the top 15 major producers’. *Reviews in Aquaculture*, **12**(2), pp. 640-663.

Macqueen, D.J., Robb, D.H., Olsen, T., Melstveit, L., Paxton, C.G. and Johnston, I.A. (2008) ‘Temperature until the ‘eyed stage’of embryogenesis programmes the growth trajectory and muscle phenotype of adult Atlantic salmon’. *Biology Letters*, **4**(3), pp. 294-298.

Macqueen, D.J., Garcia de la serrana, D. and Johnston, I.A. (2013) ‘Evolution of ancient functions in the vertebrate insulin-like growth factor system uncovered by study of duplicated salmonid fish genomes’. *Molecular Biology and Evolution*, **30**(5), pp. 1060-1076.

Macqueen, D.J. and 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.

Macqueen, D.J., Primmer, C.R., Houston, R.D., Nowak, B.F., Bernatchez, L., Bergseth, S., Davidson, W.S., Gallardo-Escárate, C., Goldammer, T., Guiguen, Y. and Iturra, P. (2017) ‘Functional Annotation of All Salmonid Genomes (FAASG): an international initiative supporting future salmonid research, conservation and aquaculture’. *BMC Genomics*, **18**(484).

Madireddy, I. (2024) ‘First transcriptome sequencing, assembly, and annotation dataset for the freshwater angelfish, *Pterophyllum scalare’*. *Data in Brief*, **54**, 110400.

Magor, B.G. (2015) ‘Antibody affinity maturation in fishes—our current understanding’. *Biology*, **4**(3), pp. 512-524.

Magnadóttir, B. (2006) ‘Innate immunity of fish (overview)’. *Fish & Shellfish Immunology*, **20**(2), pp. 137-151.

Magnadóttir, B. (2010) ‘Immunological control of fish diseases’. *Marine Biotechnology*, **12**, pp. 361-379.

Maisey, K., Montero, R. and Christodoulides, M. (2017) ‘Vaccines for piscirickettsiosis (salmonid rickettsial septicaemia, SRS): the Chile perspective’. *Expert Review of Vaccines*, **16**(3), pp. 215-228.

Manicone, A.M. and McGuire, J.K. (2008) ‘Matrix metalloproteinases as modulators of inflammation’, in *Seminars in Cell & Developmental Biology*. Academic Press: vol. 19, no. 1, pp. 34-41.

Manuel, J.M., Guilloy, N., Khatir, I., Roucou, X. and Laurent, B. (2023) ‘Re-evaluating the impact of alternative RNA splicing on proteomic diversity’. *Frontiers in Genetics*, **14**, 1089053.

Mao, D., Yan, F., Zhang, X. and Gao, G. (2022) ‘TMEM106A inhibits enveloped virus release from cell surface’. *Iscience*, **25**(2), 103843.

Marana, M.H., Karami, A.M., Ødegård, J., Zuo, S., Jaafar, R.M., Mathiessen, H., von Gersdorff Jørgensen, L., Kania, P.W., Dalsgaard, I., Nielsen, T. and Buchmann, K. (2021) ‘Whole-genome association study searching for QTL for *Aeromonas salmonicida* resistance in rainbow trout’. *Scientific Reports*, **11**(1), 17857.

Marasco, L.E. and Kornblihtt, A.R. (2023) ‘The physiology of alternative splicing’. *Nature Reviews Molecular Cell Biology*, **24**(4), pp. 242-254.

Mardis, E.R. (2011) ‘A decade’s perspective on DNA sequencing technology’. *Nature*, **470**(7333), pp. 198-203.

Marlétaz, F., Timoshevskaya, N., Timoshevskiy, V.A., Parey, E., Simakov, O., Gavriouchkina, D., Suzuki, M., Kubokawa, K., Brenner, S., Smith, J.J. and Rokhsar, D.S. (2024) 'The hagfish genome and the evolution of vertebrates'. *Nature*, **627**(8005), pp. 811-820.

Marmorstein, R. and Roth, S.Y. (2001) 'Histone acetyltransferases: function, structure, and catalysis'. *Current Opinion in Genetics & Development*, **11**(2), pp. 155-161.

Marques-Coelho, D., Iohan, L.D.C.C., Melo de Farias, A.R., Flaig, A., Lambert, J.C. and Costa, M.R. (2021) 'Differential transcript usage unravels gene expression alterations in Alzheimer's disease human brains'. *npj Aging and Mechanisms of Disease*, **7**(1), 2.

Martin, S.A., Douglas, A., Houlihan, D.F. and Secombes, C.J. (2010) 'Starvation alters the liver transcriptome of the innate immune response in Atlantic salmon (*Salmo salar*)'. *BMC Genomics*, **11**(418).

Martin, M. (2011) 'Cutadapt removes adapter sequences from high-throughput sequencing reads'. *EMBnet Journal*, **17**(1), pp. 10-12.

Martín-Vicente, M., Medrano, L.M., Resino, S., García-Sastre, A. and Martínez, I. (2017) 'TRIM25 in the regulation of the antiviral innate immunity'. *Frontiers in Immunology*, **8**, 1187.

Martorell Ribera, J., Nipkow, M., Viergutz, T., Brunner, R.M., Bochert, R., Koll, R., Goldammer, T., Gimsa, U. and Rebl, A. (2020) 'Early response of salmonid head-kidney cells to stress hormones and toll-like receptor ligands'. *Fish & Shellfish Immunology*, **98**, pp. 950-961.

Matsumoto, K., Murakami, K. and Okada, N. (1986) 'Gene for lysine tRNA1 may be a progenitor of the highly repetitive and transcribable sequences present in the salmon genome'. *Proceedings of the National Academy of Sciences*, **83**(10), pp. 3156-3160.

Matsumoto, M. and Seya, T. (2008) 'TLR3: interferon induction by double-stranded RNA including poly (I: C)'. *Advanced Drug Delivery Reviews*, **60**(7), pp. 805-812.

Mattick, J.S. (2001) 'Non-coding RNAs: the architects of eukaryotic complexity'. *EMBO Reports*, **2**(11), pp. 986-991.

Matveev, V. and Okada, N. (2009) 'Retroposons of salmonoid fishes (Actinopterygii: Salmonoidei) and their evolution'. *Gene*, **434**(1-2), pp. 16-28.

McCarthy, A. (2010) 'Third Generation DNA Sequencing: Pacific Biosciences' Single Molecule Real Time Technology.' *Chemistry & Biology*, **17**(7), pp. 675-676.

McCormack, M., Talbot, A., Dillon, E., O'Connor, I. and MacCarthy, E. (2021) 'Host Response of Atlantic Salmon (*Salmo salar*) Re-Inoculated with *Paramoeba perurans'*. *Microorganisms*, **9**(5), 993.

McInnes, L., Healy, J. and Melville, J. (2020) 'Umap: Uniform manifold approximation and projection for dimension reduction'. *arXiv Preprint,* arXiv:1802.03426v3.

McManus, C.J. and Graveley, B.R. (2011) 'RNA structure and the mechanisms of alternative splicing'. *Current Opinion in Genetics & Development*, **21**(4), pp. 373-379.

McNab, F., Mayer-Barber, K., Sher, A., Wack, A. and O'garra, A. (2015) 'Type I interferons in infectious disease'. *Nature Reviews Immunology*, **15**(2), pp. 87-103.

Melville, J. (2024) 'uwot: The Uniform Manifold Approximation and Projection (UMAP) Method for Dimensionality Reduction'. R package version 0.2.2, https://CRAN.R-project.org/package=uwot.

Mertowska, P., Smolak, K., Mertowski, S. and Grywalska, E. (2023) 'Immunomodulatory role of interferons in viral and bacterial infections'. *International Journal of Molecular Sciences*, **24**(12), 10115.

Mesev, E.V., LeDesma, R.A. and Ploss, A. (2019) 'Decoding type I and III interferon signalling during viral infection'. *Nature Microbiology*, **4**(6), pp. 914-924.

Metzker, M.L. (2010) 'Sequencing technologies—the next generation'. *Nature Reviews Genetics*, **11**(1), pp. 31-46.

Micheel, J., Safrastyan, A. and Wollny, D. (2021) 'Advances in non-coding RNA sequencing'. *Non-Coding RNA*, **7**(4), 70.

Minh, B.Q., Nguyen, M.A.T. and Von Haeseler, A. (2013) 'Ultrafast approximation for phylogenetic bootstrap'. *Molecular Biology and Evolution*, **30**(5), pp. 1188-1195.

Mog, M., Ngasotter, S., Tesia, S., Waikhom, D., Panda, P., Sharma, S. and Varshney, S. (2020) 'Problems of antibiotic resistance associated with oxytetracycline use in aquaculture: A review'. *Journal of Entomology and Zoology Studies*, **8**(3), pp. 1075-1082.

Mola, S., Beauchamp, C., Boucher, G., Lesage, S., Karaky, M., Goyette, P., Foisy, S. and Rioux, J.D. (2023) 'Identifying transcript-level differential expression in primary human immune cells'. *Molecular Immunology*, **153**, pp. 181-193.

Moore, L.D., Le, T. and Fan, G. (2013) 'DNA methylation and its basic function'. *Neuropsychopharmacology*, **38**(1), pp. 23-38.

Mordecai, G.J., Miller, K.M., Bass, A.L., Bateman, A.W., Teffer, A.K., Caleta, J.M., Di Cicco, E., Schulze, A.D., Kaukinen, K.H., Li, S. and Tabata, A. (2021) 'Aquaculture mediates global transmission of a viral pathogen to wild salmon'. *Science Advances*, **7**(22), eabe2592.

Morris, K.V. and Mattick, J.S. (2014) 'The rise of regulatory RNA'. *Nature Reviews Genetics*, **15**(6), pp. 423-437.

Morrison, R.N., Young, N.D. and Nowak, B.F. (2012) 'Description of an Atlantic salmon (*Salmo salar* L.) type II interleukin-1 receptor cDNA and analysis of interleukin-1 receptor expression in amoebic gill disease-affected fish'. *Fish & Shellfish Immunology*, **32**(6), pp. 1185-1190.

Müller, T., Boileau, E., Talyan, S., Kehr, D., Varadi, K., Busch, M., Most, P., Krijgsveld, J. and Dieterich, C. (2021) 'Updated and enhanced pig cardiac transcriptome based on long-read RNA sequencing and proteomics'. *Journal of Molecular and Cellular Cardiology,* **150**, pp. 23-31.

Nachtergaele, S. and He, C. (2017) 'The emerging biology of RNA post-transcriptional modifications'. *RNA Biology*, **14**(2), pp. 156-163.

Nakato, R. and Sakata, T. (2021) 'Methods for ChIP-seq analysis: A practical workflow and advanced applications'. *Methods*, **187**, pp. 44-53.

Namjou, B., Kothari, P.H., Kelly, J.A., Glenn, S.B., Ojwang, J.O., Adler, A., Alarcón-Riquelme, M.E., Gallant, C.J., Boackle, S.A., Criswell, L.A. and Kimberly, R.P. (2011) 'Evaluation of the TREX1 gene in a large multi-ancestral lupus cohort'. *Genes & Immunity*, **12**(4), pp. 270-279.

Naseer, S. (2023) *Promoting healthy fish in aquaculture by genome-wide functional annotation of immune responses*. University of Aberdeen.

Natoli, G. and Andrau, J.C. (2012) 'Noncoding transcription at enhancers: general principles and functional models'. *Annual Review of Genetics*, **46**(1).

Nemeth, K., Bayraktar, R., Ferracin, M. and Calin, G.A. (2024) 'Non-coding RNAs in disease: from mechanisms to therapeutics'. *Nature Reviews Genetics*, **25**(3), pp. 211-232.

Nepal, C., Hadzhiev, Y., Previti, C., Haberle, V., Li, N., Takahashi, H., Suzuki, A.M.M., Sheng, Y., Abdelhamid, R.F., Anand, S. and Gehrig, J. (2013) 'Dynamic regulation of the transcription initiation landscape at single nucleotide resolution during vertebrate embryogenesis'. *Genome Research*, **23**(11), pp. 1938-1950.

Neph, S., Kuehn, M.S., Reynolds, A.P., Haugen, E., Thurman, R.E., Johnson, A.K., Rynes, E., Maurano, M.T., Vierstra, J., Thomas, S. and

Sandstrom, R. (2012) 'BEDOPS: high-performance genomic feature operations'. *Bioinformatics*, **28**(14), pp. 1919-1920.

Nguyen, L.T., Schmidt, H.A., Von Haeseler, A. and Minh, B.Q. (2015) 'IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies'. *Molecular Biology and Evolution*, **32**(1), pp. 268-274.

Nguyen, N.H. (2024) 'Genetics and genomics of infectious diseases in key aquaculture species'. *Biology*, **13**(1), p. 29.

Ni, T., Yang, W., Han, M., Zhang, Y., Shen, T., Nie, H., Zhou, Z., Dai, Y., Yang, Y., Liu, P. and Cui, K. (2016) 'Global intron retention mediated gene regulation during CD4+ T cell activation'. *Nucleic Acids Research*, **44**(14), pp. 6817-6829.

Nicolas, P. (2009) 'Multifunctional host defense peptides: intracellular-targeting antimicrobial peptides'. *The FEBS Journal*, **276**(22), pp. 6483-6496.

Nikom, D. and Zheng, S. (2023) 'Alternative splicing in neurodegenerative disease and the promise of RNA therapies'. *Nature Reviews Neuroscience*, **24**(8), pp. 457-473.

Nimalan, N., Sørensen, S.L., Fečkaninová, A., Koščová, J., Mudroňová, D., Gancarčíková, S., Vatsos, I.N., Bisa, S., Kiron, V. and Sørensen, M. (2022) 'Mucosal barrier status in Atlantic salmon fed marine or plant-based diets supplemented with probiotics'. *Aquaculture*, **547**, 737516.

Norris, A. (2017) 'Application of genomics in salmon aquaculture breeding programs by Ashie Norris. Who knows where the genomic revolution will lead us?'. *Marine Genomics*, **36**, pp. 13-15.

Norris, K., Hopes, T. and Aspden, J.L. (2021) 'Ribosome heterogeneity and specialization in development'. *Wiley Interdisciplinary Reviews: RNA*, **12**(4), e1644.

Nowicka, M. and Robinson, M.D. (2016) 'DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics'. *F1000Research*, **5**, 1356.

Ntini, E. and Marsico, A. (2019) 'Functional impacts of non-coding RNA processing on enhancer activity and target gene expression'. *Journal of Molecular Cell Biology*, **11**(10), pp. 868-879.

Nudelman, G., Frasca, A., Kent, B., Sadler, K.C., Sealfon, S.C., Walsh, M.J. and Zaslavsky, E. (2018) 'High resolution annotation of zebrafish transcriptome using long-read sequencing'. *Genome Research*, **28**(9), pp. 1415-1425.

OECD/FAO. (2020) *OECD-FAO Agricultural Outlook 2020-2029*. Rome: OECD Publishing, Paris/FAO.

Ohkura, N., Inoue, S., Ikeda, K. and Hayashi, K. (1994) ‘’Isolation and characterization of a phospholipase A2 inhibitor from the blood plasma of the Thailand cobra *Naja naja kaouthia*. *Biochemical and Biophysical Research Communications*, **200**(2), pp. 784-788.

Ohno, S. (1970) *Evolution by gene duplication*. New York: Springer.

Ohshima, K. (2013) ‘RNA-Mediated Gene Duplication and Retroposons: Retrogenes, LINEs, SINEs, and Sequence Specificity’. *International Journal of Evolutionary Biology*, **2013**(1), 424726.

Okoniewski, M.J. and Miller, C.J. (2006) ‘Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations’. *BMC Bioinformatics*, **7**(276).

Otis, J.P., Zeituni, E.M., Thierer, J.H., Anderson, J.L., Brown, A.C., Boehm, E.D., Cerchione, D.M., Ceasrine, A.M., Avraham-Davidi, I., Tempelhof, H. and Yaniv, K. (2015) ‘Zebrafish as a model for apolipoprotein biology: comprehensive expression analysis and a role for ApoA-IV in regulating food intake’. *Disease Models & Mechanisms*, **8**(3), pp. 295-309.

Overton, K., Dempster, T., Oppedal, F., Kristiansen, T.S., Gismervik, K. and Stien, L.H. (2019) ‘Salmon lice treatments and salmon mortality in Norwegian aquaculture: a review’. *Reviews in Aquaculture*, **11**(4), pp. 1398-1417.

Paik, D.T., Tian, L., Williams, I.M., Rhee, S., Zhang, H., Liu, C., Mishra, R., Wu, S.M., Red-Horse, K. and Wu, J.C. (2020) ‘Single-cell RNA sequencing unveils unique transcriptomic signatures of organ-specific endothelial cells’. *Circulation*, **142**(19), pp. 1848-1862.

Palmeirim, I., Henrique, D., Ish-Horowicz, D. and Pourquié, O. (1997) ‘Avian hairy gene expression identifies a molecular clock linked to vertebrate segmentation and somitogenesis’. *Cell*, **91**(5), pp. 639-648.

Pang, B., van Weerd, J.H., Hamoen, F.L. and Snyder, M.P. (2023) ‘Identification of non-coding silencer elements and their regulation of gene expression’. *Nature Reviews Molecular Cell Biology*, **24**(6), pp. 383-395.

Panigrahi, A. and O’Malley, B.W. (2021) ‘Mechanisms of enhancer action: the known and the unknown’. *Genome Biology*, **22**(108).

Panousopoulou, E., Hobbs, C., Mason, I., Green, J.B. and Formstone, C.J. (2016) ‘Epiboly generates the epidermal basal monolayer and spreads the nascent mammalian skin to enclose the embryonic body’. *Journal of Cell Science*, **129**(9), pp. 1915-1927.

Papalexi, E. and Satija, R. (2018) ‘Single-cell RNA sequencing to explore immune cell heterogeneity’. *Nature Reviews Immunology*, **18**(1), pp. 35-45.

Pardo-Palacios, F.J., Arzalluz-Luque, A., Kondratova, L., Salguero, P., Mestre-Tomás, J., Amorín, R., Estevan-Morió, E., Liu, T., Nanni, A., McIntyre, L. and Tseng, E. (2024a) ‘SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms’. *Nature Methods*, **21**(5), pp. 793-797.

Pardo-Palacios, F.J., Wang, D., Reese, F., Diekhans, M., Carbonell-Sala, S., Williams, B., Loveland, J.E., De María, M., Adams, M.S., Balderrama-Gutierrez, G. and Behera, A.K. (2024b) ‘Systematic assessment of long-read RNA-seq methods for transcript identification and quantification’. *Nature Methods*, **21**, pp.1349-1363.

Pareek, C.S., Smoczynski, R. and Tretyn, A. (2011) ‘Sequencing technologies and genome sequencing’. *Journal of Applied Genetics*, **52**, pp.413-435.

Park, P.J. (2009) ‘ChIP–seq: advantages and challenges of a maturing technology’. *Nature Reviews Genetics*, **10**(10), pp. 669-680.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A. and Kingsford, C. (2017) ‘Salmon provides fast and bias-aware quantification of transcript expression’. *Nature Methods*, **14**(4), pp. 417-419.

Pauli, A., Valen, E., Lin, M.F., Garber, M., Vastenhouw, N.L., Levin, J.Z., Fan, L., Sandelin, A., Rinn, J.L., Regev, A. and Schier, A.F. (2012) ‘Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis’. *Genome Research*, **22**(3), pp. 577-591.

Pauly, D. and Zeller, D. (2016) ‘Catch reconstructions reveal that global marine fisheries catches are higher than reported and declining’. *Nature Communications*, **7**(1), 10244.

Pelechano, V. and Steinmetz, L.M. (2013) ‘Gene regulation by antisense transcription’. *Nature Reviews Genetics*, **14**(12), pp. 880-893.

Peñaranda, M.M.D., Jensen, I., Tollersrud, L.G., Bruun, J.A. and Jørgensen, J.B. (2019) ‘Profiling the Atlantic salmon IgM+ B cell surface proteome: novel information on teleost fish B cell protein repertoire and identification of potential B cell markers’. *Frontiers in Immunology*, **10**, 37.

Peoples, J.N., Ghazal, N., Duong, D.M., Hardin, K.R., Manning, J.R., Seyfried, N.T., Faundez, V. and Kwong, J.Q. (2021) ‘Loss of the mitochondrial phosphate carrier SLC25A3 induces remodeling of the cardiac mitochondrial protein acylome’. *American Journal of Physiology-Cell Physiology*, **321**(3), pp. C519-C534.

Perojil-Morata, D. (2024) *Gene regulatory evolution after the salmonid-specific whole genome duplication*. University of Edinburgh.

Perry, A.K., Chen, G., Zheng, D., Tang, H. and Cheng, G. (2005) 'The host type I interferon response to viral and bacterial infections'. *Cell Research*, **15**(6), pp. 407-422.

Pertea, M., Shumate, A., Pertea, G., Varabyou, A., Breitwieser, F.P., Chang, Y.C., Madugundu, A.K., Pandey, A. and Salzberg, S.L. (2018) 'CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise'. *Genome Biology*, **19**(208).

Pettersen, J.M., Osmundsen, T., Aunsmo, A., Mardones, F.O. and Rich, K.M. (2015a). 'Controlling emerging infectious diseases in salmon aquaculture'. *Rev Sci Tech*, **34**(3), pp. 923-938.

Pettersen, J.M., Rich, K.M., Jensen, B.B. and Aunsmo, A. 2015b. 'The economic benefits of disease triggered early harvest: a case study of pancreas disease in farmed Atlantic salmon from Norway'. *Preventive Veterinary Medicine*, **121**(3-4), pp. 314-324.

Pinheiro, D. and Heisenberg, C.P. (2020) 'Zebrafish gastrulation: Putting fate in motion'. *Current Topics in Developmental Biology*, **136**, pp. 343-375.

Pollard, T.D. (2016) 'Actin and actin-binding proteins'. *Cold Spring Harbor Perspectives in Biology*, **8**(8), a018226.

Pooley, N.J., Tacchi, L., Secombes, C.J. and Martin, S.A. (2013) 'Inflammatory responses in primary muscle cell cultures in Atlantic salmon (*Salmo salar*)'. *BMC Genomics*, **14**(747).

Prawer, Y.D., Gleeson, J., De Paoli-Iseppi, R. and Clark, M.B. (2023) 'Pervasive effects of RNA degradation on Nanopore direct RNA sequencing'. *NAR Genomics and Bioinformatics*, **5**(2), lqad060.

Press, C.M. and Evensen, Ø. (1999) 'The morphology of the immune system in teleost fishes'. *Fish & Shellfish Immunology*, **9**(4), pp. 309-318.

Qi, M., Clark, J., Moody, E.R., Pisani, D. and Donoghue, P.C. (2024) 'Molecular dating of the teleost whole genome duplication (3R) is compatible with the expectations of delayed rediploidization'. *Genome Biology and Evolution*, **16**(7), evae128.

Qi, T., Wu, Y., Fang, H., Zhang, F., Liu, S., Zeng, J. and Yang, J. (2022) 'Genetic control of RNA splicing and its distinct role in complex trait variation'. *Nature Genetics*, **54**(9), pp. 1355-1363.

Qiao, Y., Ren, C., Huang, S., Yuan, J., Liu, X., Fan, J., Lin, J., Wu, S., Chen, Q., Bo, X. and Li, X. (2020) 'High-resolution annotation of the mouse

preimplantation embryo transcriptome using long-read sequencing'. *Nature Communications*, **11**(1), 2653.

Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P. and Gu, Y. (2012) 'A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers'. *BMC Genomics*, **13**(341).

Quinlan, A.R. and Hall, I.M. (2010) 'BEDTools: a flexible suite of utilities for comparing genomic features'. *Bioinformatics*, **26**(6), pp. 841-842.

R Core Team. (2024) 'R: A Language and Environment for Statistical Computing'. *R Foundation for Statistical Computing*. Vienna, Austria. https://www.R-project.org/.

Rajme-Manzur, D., Gollas-Galván, T., Vargas-Albores, F., Martínez-Porchas, M., Hernández-Oñate, M.Á. and Hernández-López, J. (2021) 'Granulomatous bacterial diseases in fish: An overview of the host's immune response'. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*, **261**, 111058.

Raju, S.V., Sarkar, P., Kumar, P. and Arockiaraj, J. (2021) 'Piscidin, fish antimicrobial peptide: structure, classification, properties, mechanism, gene regulation and therapeutical importance'. *International Journal of Peptide Research and Therapeutics*, **27**, pp. 91-107.

Ramberg, S., Høyheim, B., Østbye, T.K.K. and 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**, 656334.

Ran, F.A.F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A. and Zhang, F. (2013) 'Genome engineering using the CRISPR-Cas9 system'. *Nature Protocols*, **8**(11), pp. 2281-2308.

Rauta, P.R., Nayak, B. and Das, S. (2012) 'Immune system and immune responses in fish and their role in comparative immunity study: a model for higher organisms'. *Immunology Letters*, **148**(1), pp. 23-33.

Rawlings, J.S., Rosler, K.M. and Harrison, D.A. (2004) 'The JAK/STAT signaling pathway'. *Journal of Cell Science*, **117**(8), pp. 1281-1283.

Reading, B.J., Andersen, L.K., Ryu, Y.W., Mushirobira, Y., Todo, T. and Hiramatsu, N. (2018) 'Oogenesis and egg quality in finfish: yolk formation and other factors influencing female fertility'. *Fishes*, **3**(4), 45.

Reese, F., Williams, B., Balderrama-Gutierrez, G., Wyman, D., Çelik, M.H., Rebboah, E., Rezaie, N., Trout, D., Razavi-Mohseni, M., Jiang, Y. and

Borsari, B. (2023) ‘The ENCODE4 long-read RNA-seq collection reveals distinct classes of transcript structure diversity’. *bioRxiv*, [preprint], doi: 10.1101/2023.05.15.540865.

Reggiardo, R.E., Maroli, S.V., Peddu, V., Davidson, A.E., Hill, A., LaMontagne, E., Aaraj, Y.A., Jain, M., Chan, S.Y. and Kim, D.H. (2023) ‘Profiling of repetitive RNA sequences in the blood plasma of patients with cancer’. *Nature Biomedical Engineering*, **7**(12), pp. 1627-1635.

Ren, C., Liu, F., Ouyang, Z., An, G., Zhao, C., Shuai, J., Cai, S., Bo, X. and Shu, W. (2017) ‘Functional annotation of structural ncRNAs within enhancer RNAs in the human genome: implications for human disease’. *Scientific Reports*, **7**(1), 15518.

Ren, Y., Tseng, E., Smith, T.P., Hiendleder, S., Williams, J.L. and Low, W.Y. (2023) ‘Long read isoform sequencing reveals hidden transcriptional complexity between cattle subspecies’. *BMC Genomics*, **24**(1).

Reuter, J.A., Spacek, D.V. and Snyder, M.P. (2015) ‘High-throughput sequencing technologies’. *Molecular Cell*, **58**(4), pp. 586-597.

Revil, T., Gaffney, D., Dias, C., Majewski, J. and Jerome-Majewska, L.A. (2010) ‘Alternative splicing is frequent during early embryonic development in mouse’. *BMC Genomics*, **11**(399).

Reyes, A. and Huber, W. (2018) ‘Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues’. *Nucleic Acids Research*, **46**(2), pp. 582-592.

Riemondy, K., Henriksen, J.C. and Rissland, O.S. (2023) ‘Intron dynamics reveal principles of gene regulation during the maternal-to-zygotic transition’. *RNA*, **29**(5), pp. 596-608.

Ritchie, M.D., Holzinger, E.R., Li, R., Pendergrass, S.A. and Kim, D. (2015) ‘Methods of integrating data to uncover genotype–phenotype interactions’. *Nature Reviews Genetics*, **16**(2), pp. 85-97.

Robertsen, B. (2006) ‘The interferon system of teleost fish’. *Fish & Shellfish Immunology*, **20**(2), pp. 172-191.

Robertsen, B. (2018) ‘The role of type I interferons in innate and adaptive immunity against viruses in Atlantic salmon’. *Developmental & Comparative Immunology*, **80**, pp. 41-52.

Robertson, F.M., Gundappa, M.K., Grammes, F., Hvidsten, T.R., Redmond, A.K., Lien, S., Martin, S.A., Holland, P.W., Sandve, S.R. and Macqueen, D.J. (2017) ‘Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification’. *Genome Biology*, **18**(111).

Robledo, D., Gutiérrez, A.P., Barría, A., Yáñez, J.M. and Houston, R.D. (2018) 'Gene expression response to sea lice in Atlantic salmon skin: RNA sequencing comparison between resistant and susceptible animals'. *Frontiers in Genetics*, **9**, 287.

Rohde, L.A. and Heisenberg, C.P. (2007) 'Zebrafish gastrulation: cell movements, signals, and mechanisms'. International Review of Cytology, **261**, pp. 159-192.

Rokka, A., Antonenkov, V.D., Soininen, R., Immonen, H.L., Pirilä, P.L., Bergmann, U., Sormunen, R.T., Weckström, M., Benz, R. and Hiltunen, J.K. (2009) 'Pxmp2 is a channel-forming protein in mammalian peroxisomal membrane'. *PLoS ONE*, **4**(4), e5090.

Rossi, R., Arjmand, S., Bærentzen, S.L., Gjedde, A. and Landau, A.M. (2022) 'Synaptic vesicle glycoprotein 2A: features and functions'. *Frontiers in Neuroscience*, **16**, 864514.

Sahlin, K. and Medvedev, P. (2021) 'Error correction enables use of Oxford Nanopore technology for reference-free transcriptome analysis'. *Nature Communications*, **12**(1).

Salmena, L. (2021) 'Pseudogenes: four decades of discovery', in *Pseudogenes: Functions and Protocols.* New York: Springer, pp. 3-18.

Salzman, J. (2016) 'Circular RNA expression: its potential regulation and function'. *Trends in Genetics*, **32**(5), pp. 309-316.

Sanger, F., Nicklen, S. and Coulson, A.R. (1977) 'DNA sequencing with chain-terminating inhibitors'. *Proceedings of the National Academy of Sciences*, **74**(12), pp. 5463-5467.

Santini, F., Harmon, L.J., Carnevale, G. and Alfaro, M.E. (2009) 'Did genome duplication drive the origin of teleosts? A comparative study of diversification in ray-finned fishes'. *BMC Evolutionary Biology*, **9**(194).

Santoro, M.M., Pesce, G. and Stainier, D.Y. (2009) 'Characterization of vascular mural cells during zebrafish development'. *Mechanisms of Development*, **126**(8-9), pp. 638-649.

Sartorelli, V. and Lauberth, S.M. (2020) 'Enhancer RNAs are an important regulatory layer of the epigenome'. *Nature Structural & Molecular Biology*, **27**(6), pp. 521-528.

Saurabh, S. and Sahoo, P.K. (2008) 'Lysozyme: an important defence molecule of fish innate immune system'. *Aquaculture Research*, **39**(3), pp. 223-239.

Schar, D., Zhao, C., Wang, Y., Larsson, D.J., Gilbert, M. and Van Boeckel, T.P. (2021) 'Twenty-year trends in antimicrobial resistance from aquaculture and fisheries in Asia'. *Nature Communications*, **12**(1), 5384.

Schaub, A. and Glasmacher, E. (2017) 'Splicing in immune cells—mechanistic insights and emerging topics'. *International Immunology*, **29**(4), pp. 173-181.

Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995) 'Quantitative monitoring of gene expression patterns with a complementary DNA microarray'. *Science*, **270**(5235), pp. 467-470.

Schloss, J.A., Gibbs, R.A., Makhijani, V.B. and Marziali, A. (2020) 'Cultivating DNA sequencing technology after the human genome project'. *Annual Review of Genomics and Human Genetics*, **21**(1), pp. 117-138.

Schoggins, J.W., Wilson, S.J., Panis, M., Murphy, M.Y., Jones, C.T., Bieniasz, P. and Rice, C.M. (2011). 'A diverse range of gene products are effectors of the type I interferon antiviral response'. *Nature*, **472**(7344), pp. 481-485.

Schroder, K., Hertzog, P.J., Ravasi, T. and Hume, D.A. (2004) 'Interferon-γ: an overview of signals, mechanisms and functions'. *Journal of Leucocyte Biology*, **75**(2), pp. 163-189.

Seki, M., Katsumata, E., Suzuki, A., Sereewattanawoot, S., Sakamoto, Y., Mizushima-Sugano, J., Sugano, S., Kohno, T., Frith, M.C., Tsuchihara, K. and Suzuki, Y. (2019) 'Evaluation and application of RNA-Seq by MinION'. *DNA Research*, **26**(1), pp. 55-65.

Seki, M., Oka, M., Xu, L., Suzuki, A. and Suzuki, Y. (2021) 'Transcript identification through long-read sequencing', in Picardi, E. (ed) *RNA Bioinformatics*. *Methods in Molecular Biology.* New York: Humana, vol. 2284, pp. 531-541.

Sessegolo, C., Cruaud, C., Da Silva, C., Cologne, A., Dubarry, M., Derrien, T., Lacroix, V. and Aury, J.M. (2019) 'Transcriptome profiling of mouse samples using nanopore sequencing of cDNA and RNA molecules'. *Scientific Reports*, **9**(1), 14908.

Sharon, D., Tilgner, H., Grubert, F. and Snyder, M. (2013) 'A single-molecule long-read survey of the human transcriptome'. *Nature Biotechnology*, **31**(11), pp. 1009-1014.

Sherwood, E.R. and Toliver-Kinsky, T. (2004) 'Mechanisms of the inflammatory response'. *Best Practice & Research Clinical Anaesthesiology*, **18**(3), pp. 385-405.

Shi, K.P., Dong, S.L., Zhou, Y.G., Li, Y., Gao, Q.F. and Sun, D.J. (2019) 'RNA-seq reveals temporal differences in the transcriptome response to acute heat stress in the Atlantic salmon (*Salmo salar*)'. *Comparative Biochemistry and Physiology Part D: Genomics and Proteomics*, **30**, pp. 169-178.

Shiau, C.K., Lu, L., Kieser, R., Fukumura, K., Pan, T., Lin, H.Y., Yang, J., Tong, E.L., Lee, G., Yan, Y. and Huse, J.T. (2023) 'High throughput single cell long-read sequencing analyses of same-cell genotypes and phenotypes in human tumors'. *Nature Communications*, **14**(1), 4124.

Shlyueva, D., Stampfel, G. and Stark, A. (2014) 'Transcriptional enhancers: from properties to genome-wide predictions'. *Nature Reviews Genetics*, **15**(4), pp. 272-286.

Simakov, O., Marlétaz, F., Yue, J.X., O'Connell, B., Jenkins, J., Brandt, A., Calef, R., Tung, C.H., Huang, T.K., Schmutz, J. and Satoh, N. (2020) 'Deeply conserved synteny resolves early events in vertebrate evolution'. *Nature Ecology & Evolution*, **4**(6), pp. 820-830.

Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G. and Kasprzyk, A. (2009) 'BioMart–biological queries made easy'. *BMC Genomics*, **10**(22).

Solnica-Krezel, L. (2006) 'Gastrulation in zebrafish—all just about adhesion?' *Current Opinion in Genetics & Development*, **16**(4), pp. 433-441.

Solnica-Krezel, L. (2020) 'Maternal contributions to gastrulation in zebrafish'. *Current Topics in Developmental Biology*, **140**, pp. 391-427.

Sommerset, I., Krossøy, B., Biering, E. and Frost, P. (2005) 'Vaccines for fish in aquaculture'. *Expert Review of Vaccines*, **4**(1), pp. 89-101.

Sommerset, I., Wiik-Nielsen, J., Moldal, T., Oliveira, V.H.S., Svendsen, J.C., Haukaas, A. and Brun, E. (2024) *Norwegian Fish Health Report 2023, Norwegian Veterinary Institute Report, series #8b/2024*. Norwegian Veterinary Institute.

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

Song, L. and Crawford, G.E. (2010) 'DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells'. *Cold Spring Harbor Protocols*, **2010**(2), pdb-prot5384.

Song, Q., Xiao, Y., Xiao, Z., Liu, T., Li, J., Li, P. and Han, F. (2021) 'Lysozymes in fish'. *Journal of Agricultural and Food Chemistry*, **69**(50), pp. 15039-15051.

Spitz, F. and Furlong, E.E. (2012) ‘Transcription factors: from enhancer binding to developmental control’. *Nature Reviews Genetics*, **13**(9), pp. 613-626.

Stark, R., Grzelak, M. and Hadfield, J., (2019) ‘RNA sequencing: the teenage years’. *Nature Reviews Genetics*, **20**(11), pp. 631-656.

Statello, L., Guo, C.J., Chen, L.L. and Huarte, M. (2021) ‘Gene regulation by long non-coding RNAs and its biological functions’. *Nature Reviews Molecular Cell Biology*, **22**(2), pp. 96-118.

Steijger, T., Abril, J.F., Engström, P.G., Kokocinski, F., Hubbard, T.J., Guigó, R., Harrow, J. and Bertone, P. (2013) ‘Assessment of transcript reconstruction methods for RNA-seq’. *Nature Methods*, **10**(12), pp. 1177-1184.

Stien, L.H., Dempster, T., Bui, S., Glaropoulos, A., Fosseidengen, J.E., Wright, D.W. and Oppedal, F. (2016) ‘‘Snorkel’sea lice barrier technology reduces sea lice loads on harvest-sized Atlantic salmon with minimal welfare impacts’. *Aquaculture*, **458**, pp. 29-37.

Stentiford, G.D., Sritunyalucksana, K., Flegel, T.W., Williams, B.A., Withyachumnarnkul, B., Itsathitphaisarn, O. and Bass, D. (2017) ‘New paradigms to help solve the global aquaculture disease crisis’. *PLoS Pathogens*, **13**(2), e1006160.

Su, Y., Yu, Z., Jin, S., Ai, Z., Yuan, R., Chen, X., Xue, Z., Guo, Y., Chen, D., Liang, H. and Liu, Z. (2024) ‘Comprehensive assessment of mRNA isoform detection methods for long-read sequencing data’. *Nature Communications*, **15**(1), 3972.

Su, Z., Wang, J., Yu, J., Huang, X. and Gu, X. (2006) ‘Evolution of alternative splicing after gene duplication’. *Genome Research*, **16**(2), pp. 182-189.

Su, Z. and Huang, D. (2021) ‘Alternative splicing of pre-mRNA in the control of immune activity’. *Genes*, **12**(4), 574.

Sullivan, C.V., Chapman, R.W., Reading, B.J. and Anderson, P.E. (2015) ‘Transcriptomics of mRNA and egg quality in farmed fish: some recent developments and future directions’. *General and Comparative Endocrinology*, **221**, pp. 23-30.

Sun, B., van Dissel, D., Mo, I., Boysen, P., Haslene-Hox, H. and Lund, H. (2022) ‘Identification of novel biomarkers of inflammation in Atlantic salmon (*Salmo salar* L.) by a plasma proteomic approach’. *Developmental & Comparative Immunology*, **127**, 104268.

Sun, J., Daniels, R.R., Balic, A., Andresen, A.M., Bjørgen, H., Dobie, R., Henderson, N.C., Koppang, E.O., Martin, S.A., Fosse, J.H. and Taylor,

R.S. (2024) ‘Cell atlas of the Atlantic salmon spleen reveals immune cell heterogeneity and cell-specific responses to bacterial infection’. *Fish & Shellfish Immunology*, **145**, 109358.

Tacchi, L., Bickerdike, R., Douglas, A., Secombes, C.J. and Martin, S.A. (2011) ‘Transcriptomic responses to functional feeds in Atlantic salmon (*Salmo salar*)’. *Fish & Shellfish Immunology*, **31**(5), pp. 704-715.

Talbot, A., Gargan, L., Moran, G., Prudent, L., O’Connor, I., Mirimin, L., Carlsson, J. and MacCarthy, E. (2021) ‘Investigation of the transcriptomic response in Atlantic salmon (*Salmo salar*) gill exposed to *Paramoeba perurans* during early onset of disease’. *Scientific Reports*, **11**(1), 20682.

Tammas, I., Bitchava, K. and Gelasakis, A.I. (2024) ‘Transforming aquaculture through vaccination: A review on recent developments and milestones’. *Vaccines*, **12**(7), 732.

Tan, H., Onichtchouk, D. and Winata, C. (2016) ‘DANIO-CODE: toward an encyclopedia of DNA elements in zebrafish’. *Zebrafish*, **13**(1), pp. 54-60.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A. and Lao, K. (2009) ‘mRNA-Seq whole-transcriptome analysis of a single cell’. *Nature Methods*, **6**(5), pp. 377-382.

Tang, S., Lomsadze, A. and Borodovsky, M. (2015) ‘Identification of protein coding regions in RNA transcripts’. *Nucleic Acids Research*, **43**(12).

Tao, Y., Zhang, Q., Wang, H., Yang, X. and Mu, H. (2024) ‘Alternative splicing and related RNA binding proteins in human health and disease’. *Signal Transduction and Targeted Therapy*, **9**(1), p. 26.

Tardaguila, M., De La Fuente, L., Marti, C., Pereira, C., Pardo-Palacios, F.J., Del Risco, H., Ferrell, M., Mellado, M., Macchietto, M., Verheggen, K. and Edelmann, M. (2018) ‘SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification’. *Genome Research*, **28**(3), pp. 396-411.

Tarifeño-Saldivia, E., Valenzuela-Miranda, D. and Gallardo-Escárate, C. (2017) ‘In the shadow: The emerging role of long non-coding RNAs in the immune response of Atlantic salmon’. *Developmental & Comparative Immunology*, **73**, pp. 193-205.

Tawfik, M.M., Betancor, M.B., McMillan, S., Norambuena, F., Tocher, D.R., Douglas, A. and Martin, S.A. (2024) ‘Modulation of metabolic and immunoregulatory pathways in the gut transcriptome of Atlantic salmon (*Salmo salar* L.) after early nutritional programming during first feeding with plant-based diet’. *Frontiers in Immunology*, **15**, 1412821.

Taylor, R.S., Ruiz Daniels, R., Dobie, R., Naseer, S., Clark, T.C., Henderson, N.C., Boudinot, P., Martin, S.A. and Macqueen, D.J. (2022) 'Single cell transcriptomics of Atlantic salmon (*Salmo salar* L.) liver reveals cellular heterogeneity and immunological responses to challenge by *Aeromonas salmonicida*'. *Frontiers in Immunology*, **13**, 984799.

Thompson, M.R., Kaminski, J.J., Kurt-Jones, E.A. and Fitzgerald, K.A. (2011) 'Pattern recognition receptors and the innate immune response to viral infection'. *Viruses*, **3**(6), pp. 920-940.

Tian, L., Jabbari, J.S., Thijssen, R., Gouil, Q., Amarasinghe, S.L., Voogd, O., Kariyawasam, H., Du, M.R., Schuster, J., Wang, C. and Su, S. (2021) 'Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing'. *Genome Biology*, **22**(310).

Tomalty, K.M., Meek, M.H., Stephens, M.R., Rincón, G., Fangue, N.A., May, B.P. and Baerwald, M.R. (2015) 'Transcriptional response to acute thermal exposure in juvenile Chinook salmon determined by RNAseq'. *G3: Genes, Genomes, Genetics*, **5**(7), pp. 1335-1349.

Torre, D., Francoeur, N.J., Kalma, Y., Gross Carmel, I., Melo, B.S., Deikus, G., Allette, K., Flohr, R., Fridrikh, M., Vlachos, K. and Madrid, K. (2023) 'Isoform-resolved transcriptome of the human preimplantation embryo'. *Nature Communications*, **14**(1), 6902.

Torrissen, O., Jones, S., Asche, F., Guttormsen, A., Skilbrei, O.T., Nilsen, F., Horsberg, T.E. and Jackson, D. (2013) 'Salmon lice–impact on wild salmonids and salmon aquaculture'. *Journal of Fish Diseases*, **36**(3), pp. 171-194.

Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L. and Pachter, L. (2013) 'Differential analysis of gene regulation at transcript resolution with RNA-seq'. *Nature Biotechnology*, **31**(1), pp. 46-53.

Travers, K.J., Chin, C.S., Rank, D.R., Eid, J.S. and Turner, S.W. (2010) 'A flexible and efficient template format for circular consensus sequencing and SNP detection'. *Nucleic Acids Research*, **38**(15), e159.

Trifinopoulos, J., Nguyen, L.T., von Haeseler, A. and Minh, B.Q. (2016) 'W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis'. *Nucleic Acids Research*, **44**(W1), pp. W232-W235.

Troskie, R.L., Jafrani, Y., Mercer, T.R., Ewing, A.D., Faulkner, G.J. and Cheetham, S.W. (2021) 'Long-read cDNA sequencing identifies functional pseudogenes in the human transcriptome'. *Genome Biology*, **22**(146).

Tsai, M.C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J.K., Lan, F., Shi, Y., Segal, E. and Chang, H.Y. (2010) 'Long noncoding RNA as

modular scaffold of histone modification complexes'. *Science*, **329**(5992), pp. 689-693.

United Nations Department of Economic and Social Affairs, Population Division. (2022). *World Population Prospects 2022: Summary of Results*. UN DESA/POP/2022/TR/NO. 3.

Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigó, R. and Johnson, R. (2018) 'Towards a complete map of the human long non-coding RNA transcriptome'. *Nature Reviews Genetics*, **19**(9), pp. 535-548.

Valero, Y., Saraiva-Fraga, M., Costas, B. and Guardiola, F.A. (2020) 'Antimicrobial peptides from fish: beyond the fight against pathogens'. *Reviews in Aquaculture*, **12**(1), pp. 224-253.

Van den Berge, K., Soneson, C., Robinson, M.D. and Clement, L. (2017) 'stageR: a general stage-wise method for controlling the gene-level false discovery rate in differential expression and differential transcript usage'. *Genome Biology*, **18**(151).

Van der Vaart, M., Spaink, H.P. and Meijer, A.H. (2012) 'Pathogen recognition and activation of the innate immune response in zebrafish'. *Advances in Hematology,* **2012**, 159807.

van Der Werf, I., Mondala, P.K., Steel, S.K., Balaian, L., Ladel, L., Mason, C.N., Diep, R.H., Pham, J., Cloos, J., Kaspers, G.J. and Chan, W.C. (2023) 'Detection and targeting of splicing deregulation in pediatric acute myeloid leukemia stem cells'. *Cell Reports Medicine*, **4**(3), 100962.

Van Dijk, E.L., Auger, H., Jaszczyszyn, Y. and Thermes, C. (2014) 'Ten years of next-generation sequencing technology'. *Trends in Genetics*, **30**(9), pp. 418-426.

Van Dijk, E.L., Jaszczyszyn, Y., Naquin, D. and Thermes, C. (2018) 'The third revolution in sequencing technology'. *Trends in Genetics*, **34**(9), pp. 666-681.

Varadharajan, S., Sandve, S.R., Gillard, G.B., Tørresen, O.K., Mulugeta, T.D., Hvidsten, T.R., Lien, S., Asbjørn Vøllestad, L., Jentoft, S., Nederbragt, A.J. and Jakobsen, K.S. (2018) 'The grayling genome reveals selection on gene expression regulation after whole-genome duplication'. *Genome Biology and Evolution*, **10**(10), pp. 2785-2800.

Vargas-Chacoff, L., Regish, A.M., Weinstock, A. and McCormick, S.D. (2018) 'Effects of elevated temperature on osmoregulation and stress responses in Atlantic salmon *Salmo salar* smolts in fresh water and seawater'. *Journal of Fish Biology*, **93**(3), pp. 550-559.

Vastenhouw, N.L., Cao, W.X. and Lipshitz, H.D. (2019) 'The maternal-to-zygotic transition revisited'. *Development*, **146**(11), dev161471.

Vesterlund, L., Jiao, H., Unneberg, P., Hovatta, O. and Kere, J. (2011) 'The zebrafish transcriptome during early development'. *BMC Developmental Biology*, **11**(30).

Vidgren, J., Svensson, L.A. and Liljas, A. (1994) 'Crystal structure of catechol O-methyltransferase'. Nature, 368(6469), pp. 354-358.

Vitting-Seerup, K. and Sandelin, A. (2019) 'IsoformSwitchAnalyzeR: analysis of changes in genome-wide patterns of alternative splicing and its functional consequences'. *Bioinformatics*, **35**(21), pp. 4469-4471.

Wahlestedt, C. (2013) 'Targeting long non-coding RNA to therapeutically upregulate gene expression'. *Nature Reviews Drug Discovery*, **12**(6), pp. 433-446.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K. and Earl, A.M. (2014) 'Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement'. *PLoS One*, **9**(11), e112963.

Walser, C.B. and Lipshitz, H.D. (2011) 'Transcript clearance during the maternal-to-zygotic transition'. *Current Opinion in Genetics & Development*, **21**(4), pp. 431-443.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P. and Burge, C.B. (2008) 'Alternative isoform regulation in human tissue transcriptomes'. *Nature*, **456**(7221), pp. 470-476.

Wang, K.C. and Chang, H.Y. (2011) 'Molecular mechanisms of long noncoding RNAs'. *Molecular Cell*, **43**(6), pp. 904-914.

Wang, Y., Zhao, Y., Bollas, A., Wang, Y. and Au, K.F. (2021) 'Nanopore sequencing technology, bioinformatics and applications'. *Nature Biotechnology*, **39**(11), pp.1348-1365.

Wang, Z., Gerstein, M. and Snyder, M. (2009) 'RNA-Seq: a revolutionary tool for transcriptomics'. *Nature Reviews Genetics*, **10**(1), pp. 57-63.

Wang, Z., You, X., Zhang, Y., Liu, Q. and Yang, D. (2024) 'Poly (I: C) induces anti-inflammatory response against secondary LPS challenge in zebrafish larvae'. *Fish & Shellfish Immunology*, **144**, 109285.

Watson, J.D. and Crick, F.H. (1953) 'Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid'. *Nature*, **171**(4356), pp. 737-738.

Weber, G.M., Birkett, J., Martin, K., Dixon, D., Gao, G., Leeds, T.D., Vallejo, R.L. and Ma, H. (2021) 'Comparisons among rainbow trout, *Oncorhynchus mykiss*, populations of maternal transcript profile associated with egg viability. *BMC Genomics*, **22**(1), 448.

Wehrens, R. and Kruisselbrink, J. (2018) 'Flexible self-organizing maps in kohonen 3.0'. *Journal of Statistical Software*, **87**, pp. 1-18.

Weirather, J.L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.J., Buck, D. and Au, K.F. (2017) 'Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis'. *F1000Research*, **6**(100).

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. and Töpfer, A. (2019) 'Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome'. *Nature Biotechnology*, **37**(10), pp. 1155-1162.

Werner, A., Cockell, S., Falconer, J., Carlile, M., Alnumeir, S. and Robinson, J. (2014) 'Contribution of natural antisense transcription to an endogenous siRNA signature in human cells'. *BMC Genomics*, **15**(19).

White, R.J., Collins, J.E., Sealy, I.M., Wali, N., Dooley, C.M., Digby, Z., Stemple, D.L., Murphy, D.N., Billis, K., Hourlier, T. and Füllgrabe, A. (2017a) 'A high-resolution mRNA expression time course of embryonic development in zebrafish'. *elife*, **6**, e30860.

White, R., Pellefigues, C., Ronchese, F., Lamiable, O. and Eccles, D. (2017b) 'Investigation of chimeric reads using the MinION'. *F1000Research*, **6**, 100.

Whyte, S.K. (2007) 'The innate immune response of finfish–a review of current knowledge'. *Fish & Shellfish Immunology*, **23**(6), pp. 1127-1151.

Wick, R.R., Judd, L.M. and Holt, K.E. (2019) 'Performance of neural network basecalling tools for Oxford Nanopore sequencing'. *Genome Biology*, **20**(129).

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

Wight, M. and Werner, A. (2013) 'The functions of natural antisense transcripts'. *Essays in Biochemistry*, **54**, pp. 91-101.

Woolley, S.A., Salavati, M. and Clark, E.L. (2023) 'Recent advances in the genomic resources for sheep'. *Mammalian Genome*, **34**(4), pp. 545-558.

Workman, R.E., Tang, A.D., Tang, P.S., Jain, M., Tyson, J.R., Razaghi, R., Zuzarte, P.C., Gilpatrick, T., Payne, A., Quick, J. and Sadowski, N. (2019)

‘Nanopore native RNA sequencing of a human poly (A) transcriptome’. *Nature Methods*, **16**(12), pp. 1297-1305.

Wragg, J. and Müller, F. (2016) ‘Transcriptional regulation during zygotic genome activation in zebrafish and other anamniote embryos’. *Advances in Genetics*, **95**, pp. 161-194.

Wright, C.J., Smith, C.W. and Jiggins, C.D. (2022) ‘Alternative splicing as a source of phenotypic diversity’. *Nature Reviews Genetics*, **23**(11), pp. 697-710.

Wringe, B.F., Devlin, R.H., Ferguson, M.M., Moghadam, H.K., Sakhrani, D. and Danzmann, R.G. (2010) ‘Growth-related quantitative trait loci in domestic and wild rainbow trout (*Oncorhynchus mykiss*)’. *BMC Genetics*, **11**(63).

Wu, R., Chen, F., Wang, N., Tang, D. and Kang, R. (2020) ‘ACOD1 in immunometabolism and disease’. *Cellular & Molecular Immunology*, **17**(8), pp. 822-833.

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L.I. and Fu, X. (2021) ‘clusterProfiler 4.0: A universal enrichment tool for interpreting omics data’. *The Innovation*, **2**(3), 100141.

Xia, Y.Q., Cheng, J.X., Liu, Y.F., Li, C.H., Liu, Y. and Liu, P.F. (2022) ‘Genome-wide integrated analysis reveals functions of lncRNA-miRNA-mRNA interactions in Atlantic salmon challenged by *Aeromonas salmonicida’*. *Genomics*, **114**(1), pp. 328-339.

Xiang, R., Fang, L., Liu, S., Macleod, I.M., Liu, Z., Breen, E.J., Gao, Y., Liu, G.E., Tenesa, A., Mason, B.A. and Chamberlain, A.J. (2023) ‘Gene expression and RNA splicing explain large proportions of the heritability for complex traits in cattle’. *Cell Genomics*, **3**(10).

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C.Y., Feng, Y., Liu, Z., Zeng, Q., Cheng, L., Sun, Y.E. and Liu, J.Y. (2013) ‘Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing’. *Nature*, **500**(7464), pp. 593-597.

Yamaguchi, K., Ishigaki, K., Suzuki, A., Tsuchida, Y., Tsuchiya, H., Sumitomo, S., Nagafuchi, Y., Miya, F., Tsunoda, T., Shoda, H. and Fujio, K. (2022) ‘Splicing QTL analysis focusing on coding sequences reveals mechanisms for disease susceptibility loci’. *Nature Communications*, **13**(1), 4659.

Ye, J., Kaattari, I.M. and Kaattari, S.L. (2011) ‘The differential dynamics of antibody subpopulation expression during affinity maturation in a teleost’. *Fish & Shellfish Immunology*, **30**(1), pp. 372-377.

Ye, J., Xu, M., Tian, X., Cai, S. and Zeng, S. (2019) ‘Research advances in the detection of miRNA’. *Journal of Pharmaceutical Analysis*, **9**(4), pp. 217-226.

Yi, L., Pimentel, H., Bray, N.L. and Pachter, L. (2018) ‘Gene-level differential analysis at transcript-level resolution’. *Genome Biology*, **19**(53).

Yu, D., Zhou, M., Chen, W., Ding, Z., Wang, C., Qian, Y., Liu, Y., He, S. and Yang, L. (2024) ‘Characterization of transcriptome changes in saline stress adaptation on *Leuciscus merzbacheri* using PacBio Iso-Seq and RNA-Seq’. *DNA Research*, **31**(3), dsae019.

Yu, G., Wang, L.G., Han, Y. and He, Q.Y. (2012) ‘clusterProfiler: an R package for comparing biological themes among gene clusters’. *Omics: a Journal of Integrative Biology*, **16**(5), pp. 284-287.

Yusuf, F. and Brand-Saberi, B. (2006) ‘The eventful somite: patterning, fate determination and cell division in the somite’. *Anatomy and Embryology*, **211**(Suppl 1), pp. 21-30.

Zang, H., Guo, S., Dong, S., Song, Y., Li, K., Fan, X., Qiu, J., Zheng, Y., Jiang, H., Wu, Y. and Lü, Y. (2024) ‘Construction of a Full-Length Transcriptome of Western Honeybee Midgut Tissue and Improved Genome Annotation’. *Genes*, **15**(6), 728.

Zapata, A. and Amemiya, C.T. (2000) ‘Phylogeny of lower vertebrates and their immunological structures’, in Du Pasquier, L. and Litman, G.W. (eds) *Origin and evolution of the vertebrate immune system*. Berlin: Springer, Heidelberg, pp.67-107.

Zhang, F., Tao, Y., Zhang, Z., Guo, X., An, P., Shen, Y., Wu, Q., Yu, Y. and Wang, F. (2012) ‘Metalloreductase Steap3 coordinates the regulation of iron homeostasis and inflammatory responses’. *Haematologica*, **97**(12), 1826.

Zhang, J., Zhang, Y.Z., Jiang, J. and Duan, C.G. (2020) ‘The crosstalk between epigenetic mechanisms and alternative RNA processing regulation’. *Frontiers in Genetics*, **11**, 998.

Zhang, P.G., Huang, S.Z., Pin, A.L. and Adams, K.L. (2010) ‘Extensive divergence in alternative splicing patterns after gene and genome duplication during the evolutionary history of *Arabidopsis’*. *Molecular Biology and Evolution*, **27**(7), pp. 1686-1697.

Zhang, R., Kuo, R., Coulter, M., Calixto, C.P., Entizne, J.C., Guo, W., Marquez, Y., Milne, L., Riegler, S., Matsui, A. and Tanaka, M. (2022) ‘A high-resolution single-molecule sequencing-based Arabidopsis transcriptome using novel methods of Iso-seq analysis’. *Genome Biology*, **23**(1), 149.

Zhang, S., Bhattacharya, H. and Li, H. (2019) ‘Embryogenesis and development’, in *Reproductive Biology and Phylogeny of Fishes, Vol 8B: Part B: Sperm Competion Hormones*. CRC Press, pp. 485-517.

Zhang, Y., Qian, J., Gu, C. and Yang, Y. (2021) ‘Alternative splicing and cancer: a systematic review’. *Signal Transduction and Targeted Therapy*, 6(78).

Zhao, L., Huang, J., Li, Y., Wu, S. and Kang, Y. (2023) ‘Skin immune response of rainbow trout (*Oncorhynchus mykiss*) infected with infectious hematopoietic necrosis virus’. *Aquaculture International*, **31**(6), pp. 3275-3295.

Zhao, S. (2019) ‘Alternative splicing, RNA-seq and drug discovery’. *Drug Discovery Today*, **24**(6), pp. 1258-1267.

Zhong, L., Carvalho, L.A., Gao, S., Whyte, S.K., Purcell, S.L., Fast, M.D. and Cai, W. (2023) ‘Transcriptome analysis revealed immune responses in the kidney of Atlantic salmon (*Salmo salar*) co-infected with sea lice (*Lepeophtheirus salmonis*) and infectious salmon anemia virus’. *Fish & Shellfish Immunology*, **143**, 109210.

Zhou, R., Liu, L. and Wang, Y. (2021) ‘Viral proteins recognized by different TLRs’. *Journal of Medical Virology*, **93**(11), pp. 6116-6123.

Zinngrebe, J., Montinaro, A., Peltzer, N. and Walczak, H. (2014) ‘Ubiquitin in the immune system’. *EMBO Reports*, **15**(1), pp. 28-45.

Zou, J. and Secombes, C.J. (2011) ‘Teleost fish interferons and their role in immunity’. *Developmental & Comparative Immunology*, **35**(12), pp. 1376-1387.