Norwegian University of Life Sciences
Faculty of Chemistry, Biotechnology and Food Science

Philosophiae Doctor (PhD)
Thesis 2019:53

Evolution of gene expression following the whole genome duplication in salmonid fish

Evolusjon av genuttrykk etter helgenomduplikasjon i laksefisk

Gareth Benjamin Gillard

Evolution of gene expression following the whole genome duplication in salmonid fish

Evolusjon av genuttrykk etter helgenomduplikasjon i laksefisk

Philosophiae Doctor (PhD) Thesis

Gareth Benjamin Gillard

Norwegian University of Life Sciences
Faculty of Chemistry, Biotechnology and Food Science

Ås 2019

Thesis number 2019:53
ISSN 1894-6402
ISBN 978-82-575-1612-3

Summary

Whole genome duplication (WGD) is a rare mutational event that provides additional duplicates of all genes in the entire genome, resulting in functional redundancy. This redundancy leads to relaxation of selective constraints and can in turn spark evolution of novel phenotypes. Although there seem to be an association between WGD and the propensity to survive and adapt to novel environments, this potential link between WGD events and a surge of adaptive evolution is rather anecdotal and not well supported by empirical evidence at this point. In this thesis, we apply various comparative transcriptomics approaches to investigate the impact of a salmonid-specific WGD (4R) on gene expression evolution. In paper I, we investigate the consequences of the WGD on gene regulation in Atlantic salmon lipid metabolism-related pathways. We found pathway specific differences in duplicate retention which was independent of how conserved regulation was between duplicates. We identified gene dosage effects in only certain pathways related to the biosynthesis of unsaturated fatty acids. In paper II, we investigated and compared the consequences of the WGD on the regulation of genes in European grayling and Atlantic salmon. We classify 4R duplicate pairs into different evolutionary scenarios and found that, only a very small fraction (~5%) displayed hallmarks of adaptive evolution of novel tissue regulation. In paper III, we use a phylogenetic statistical framework (based on the Ornstein-Uhlenbeck process) to detect evolutionary shifts in liver gene expression levels in the salmonid lineage compared to outgroup species without the 4R WGD. We observe higher gene expression evolution rates following WGD, with some examples of likely adaptive increases in liver gene expression. However, the majority of expression level shifts conserved across salmonid species represented a decrease in expression compared to the pre-4R ancestral expression levels. This suggests that strong selection for dosage compensation is acting on early evolution of gene expression following WGD. Taken together, this thesis describes how gene expression diverged after the WGD in salmonids and represents a first step towards a genome wide understanding of the consequences of WGD on evolution of gene expression.

Sammendrag

En helgenomduplikasjon (HGD) er en sjelden mutasjonshendelse som gir ekstra duplikater av alle genene i ett genom, og som derfor resulterer i funksjonell redundans. Denne redundansen muliggjør akkumulasjon av nye mutasjoner i gener med en ekstra 'backup' kopi, som igjen kan lede til evolusjon av nye fenotyper. Selv om mye tyder på at det finnes en assosiasjon mellom HGD og sannsynligheten for å overleve og tilpasse seg nye miljøer, så er den empiriske støtten for at HGD leder til økt adaptiv evolusjon relativt anekdotisk. I denne avhandlingen bruker vi ulike metoder for komparativ transkriptomikk til å undersøke hvilken innvirkning en laksefisk-spesifikk HGD (4R) har hatt på evolusjon av genuttrykk. I artikkel 1 undersøker vi hvilken innvirkning HGD har hatt på genreguleringen av metabolske stier relatert til lipidmetabolisme i atlantisk laks. Vi fant forskjeller i duplikat-bevaring som var spesifikk for utvalgte metabolske stier og som var uavhengig av i hvor stor grad duplikatene var regulert likt. Vi identifiserte bare gendose-effekter i metabolske stier relatert til biosyntese av umettede fettsyrer. I artikkel 2 undersøkte og sammenlignet vi innvirkning HGD har hatt på evolusjon av genuttrykk i harr og atlantisk laks. Vi klassifiserte genduplikater i ulike evolusjonære senarioer og fant at bare en liten andel (~5%) viste tydelige tegn på adaptiv evolusjon av ny vevsregulering. I artikkel 3 brukte vi et fylogenetisk statistisk rammeverk (basert på Ornstein-Uhlenbeck prosessen) til å detektere skift i genuttrykksnivå i leveren til laksefisk sammenlignet med utgruppearter uten 4R HGD. Vi observerte høyere evolusjonsrater på genuttrykk etter HGD, og identifiserte noen eksempler på det som sannsynligvis er adaptiv økning av genuttrykk i lever. Likevel representerte de fleste uttrykksskiftene som var konservert i alle laksefiskene en nedregulering av uttrykksnivå sammenlignet med det som fantes før 4R. Dette tyder på at det finnes en sterk seleksjon på dosekompensasjon i den tidlige fasen av genuttrykksevolusjon etter HGD. Denne avhandlingen beskriver hvordan genuttrykk divergerte etter HGD i laksefiskene og representerer et første steg mot en forståelse av hvordan HGD påvirker evolusjon av genuttrykk på genomnivå.

“Change… is good”

Kha’Zix, the voidreaver

Acknowledgements

I am forever grateful to the many people who have made my PhD and Norway experience what it was. To my supervisors Torgeir and Simen, you have been invaluable for my progression in academia and life. I felt continuous support and investment from you both. Thank you for the many opportunities you provided me to broaden my research connections, attend conferences in some amazing places, and facilitate my unforgettable research stay in San Francisco.

To Rori, thank you for all your hospitality during my stay at San Francisco State University. I appreciate all the invested interest I felt from you during our work together. I also extend my thanks to everyone else I meet during my stay there.

To Chris, my Masters supervisor, I'm grateful for your recommendation of me for this PhD project. What I learnt during my time under you has shaped me into the bioinformatician I am today.

Thank you to many people at CIGENE, NUMBU, and elsewhere who have been instrumental as colleges and coauthors on my work.

To everyone at the NMBU biostatistics group, past and present, thank you for all fun times, whether it was chatting over coffee or lunch, or even ice fishing. You provided me with assistance when I needed it, and helped make my parents a little less worried about me being on the other side of the world. You made Norway feel like my second home.

Many of my coworkers quickly become close friends. To Tom, Line, Yang, and others, my time spent with you will never be forgotten. I'm thankful I got to share many amazing and important experiences with you, whether it was flying around the world to conferences or playing board games together. I look forward to whatever else we do next.

To my love Erica, during my time in Norway you've shared with me so many highlights of my life, as well as sharing my burden of completing a PhD. You've helped make me who I am today. I can't write enough words in this little space to say how much you've meant to me, I'd need to write it down in its own thesis. So instead, I'll endeavor to let you know each and every day in person. I extend my love to all of Erica's family as well for sharing their love and support with me during our times together.

To my Mother, Farther, and all my family, thank you for all the non-stop love and support you have given. You've shown so much care and understanding for me during my time as a PhD student. You hide the burden of being so far apart. Our distance has limited the time we could spend together, yet you're always flexible with times to video call. I hope you all share with me this accomplishment.

List of papers

The thesis is based on the following three papers, referred to by their Roman numerals.

I. **Gillard, G.*, Harvey, T. N.*, Gjuvsland, A., Jin, Y., Thomassen, M., Lien, S., Leaver, M., Torgersen, J. S., Hvidsten, T. R., Vik, J. O. and Sandve, S. R.** (2018). Life-stage-associated remodelling of lipid metabolism regulation in Atlantic salmon. *Molecular Ecology* **27**(5), 1200–1213.

II. **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), 2785–2800.

III. **Gillard, G. B., Rohlfs, R. V., Koop, B. F., Rondeau, E. B., Sandve, S. R.**, and **Hvidsten, T. R.** (2019). Gene regulatory evolution following salmonid whole genome duplication. *Manuscript*

* Equal contribution

Paper contributions

My contributions to the papers included in the thesis.

I. Performed all analysis of transcriptome data, including the differential expression analysis and the duplicate analysis.

II. Performed the identification of ortholog groups and the expression level comparison with liver transcriptome data.

III. Performed all data analysis, except for the co-expression network analysis.

Table of Contents

1

Introduction

Genomic variation is the fundamental basis for the evolution of all the diverse life that exists on this planet. Gene duplication is one mechanism that give rise to novel genomic variation and contribute to the evolution of species and adaptation of novel traits (Zhang 2003; Stephens 1951a; Ohno 1970). An extreme example of duplication is whole genome duplication (WGD) in which all chromosomes of an individual become duplicated, resulting in a huge influx of new genetic material all at once. Understanding the consequences that whole genome duplications have had on gene and genome evolution is an important step towards understanding the evolution of all life.

1.1 The role of gene duplication in evolution

1.1.1 Gene duplication

As early as 1936, a report by Bridges (Bridges 1936) described how a duplication of the Bar gene in fruit fly (*Drosophila melanogaster)* was responsible for an extreme reduction in eye-size. The impact that gene duplication has on a species´ phenotype and evolution has continued to be investigated to this day. Following that study, scenarios began to be proposed on how the duplication of genes could contribute to evolution (Stephens 1951b; Ohno 2013; Nei 1969) including the famous book by Ohno: *Evolution by Gene Duplication* (Ohno 1970). However, it was not until advances in whole genome sequencing delivered a massive increase in the number of sequenced genomes that we realised how prolific gene duplication was. Gene duplication was found prevalent throughout all three domains of life with a large portion of known genes originating from a duplication.

The expected rate of gene duplication in eukaryotes is about one duplication per gene per 100 million years (Lynch and Conery 2000), comparable to the nucleotide substitution rate in vertebrates (0.1 to 0.5 per site per 100 million years) (Li 1997). The proportion of genes with a duplicated copy (also referred to as paralogs) varies in organisms from different domains of life. For example, the number of duplicated genes is 38% in humans (Li et al. 2001), 65% in the plant *Arabidopsis thaliana* (Arabidopsis Genome Initiative 2000), 41% in the fruit fly *Drosophila melanogaster* (Rubin et al. 2000), 44% in the bacteria *Mycoplasma pneumoniae* (Himmelreich et al. 1996), and 17% in the bacteria *Haemophilus influenzae* (Rubin et al. 2000). Repeated duplication of a gene can result in large gene families containing genes with similar functions, and the size of such families can vary between genes and species (Lespinet et al. 2002). For example, the biggest gene family in fruit fly is the trypsin gene family (Gu et al. 2002) with 111 members, while the biggest family in mammals is the olfactory receptor family with around 1000 members (Mombaerts 2001).

1.1.2 How genes become duplicated

Several scenarios may occur that result in the duplication of genes (Figure 1). These may be generalised as either a small-scale duplication when it involves the duplication of a single gene or a section of the genome containing several genes, or a large-scale duplication when it involves the duplication of entire chromosomes or even the entire genome at once. The mode of the duplication may be consequential to the evolutionary fate of the gene or genes that are duplicated (Zhang 2003).

A small-scale duplication may occur from the unequal crossing over of chromosomes during cell division (Figure 1A). The duplicated sequence may be a part of a gene, an entire gene, or several genes in tandem. A property of this mode is that duplicated genes may be copied complete with their flanking regulatory regions, and multiple genes remain linked in chromosomal space. Another mode of small-scale duplication that contrasts with the previous is retrotransposition (Figure 1B). Retrotransposition occurs when gene's transcript becomes retrotranscribed from RNA to a cDNA sequence by a retrotransposase protein and is then inserted back into the genome. This mode only duplicates a single gene to a random location in the genome, and the gene's intron and regulatory sequences are not copied as those regions are not transcribed. Without the gene's regulatory sequence copied the transposed gene becomes a nonfunctional pseudogene by default, and must rely on the recruitment of regulatory elements to be expressed (Long 2001)

A large-scale duplication may occur from a lack of disjunction between daughter chromosomes after DNA replication. This may result in the duplication of entire chromosomes or even a whole genome duplication (WGD). This mode of duplication results in a huge number of gene duplications, all in the same chromosomal space and with their regulatory regions intact. These large-scale duplications are important events given the large sudden influx of functional genes they provide.

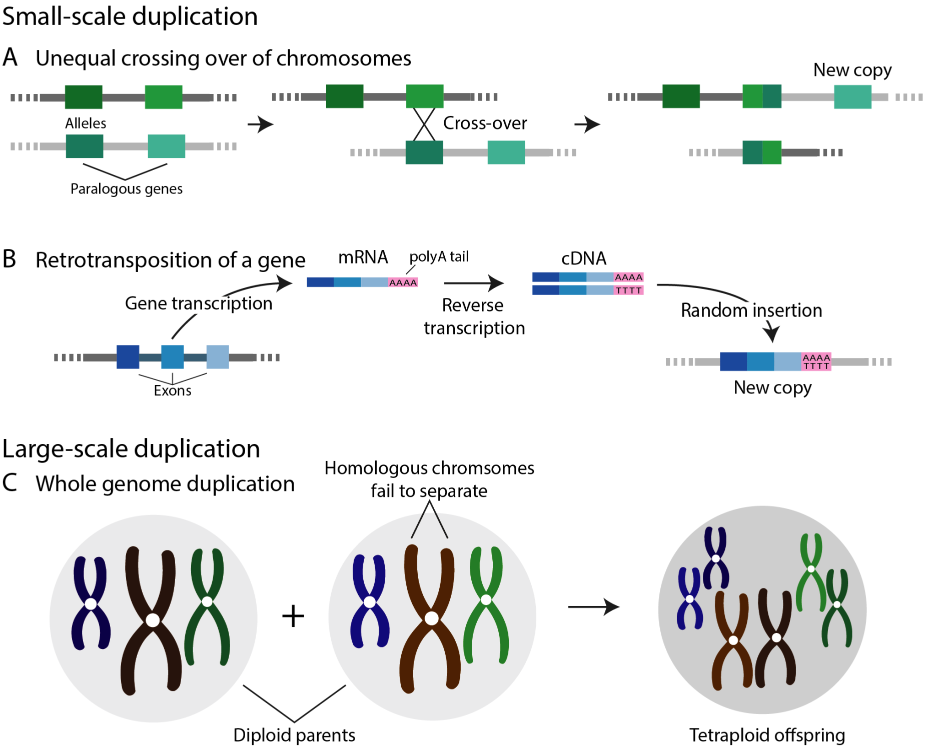


Figure 1: Common modes of gene duplication. A small-scale duplication may result from (A) an unequal crossover of chromosomes or (B) the retrotransposition of a gene's transcript into a new part of the genome. A large-scale duplication may occur when chromosomes fail to separate after RNA replication, resulting in offspring with a chromosome or (C) whole genome duplication.

1.1.3 Evolutionary fate of duplicated genes

Gene duplications first occur in single individuals and then may either be lost or fixed in the population, similarly to point mutations. However, if retained, the long-term fate of a duplication is dependent on subsequent genetic changes that occur to the new copy that determines its role in the organism. The gain and loss of duplicates throughout the genome is a constant theme (Hughes and Nei 1989; Nei et al. 2000). There have been different scenarios theorised to explain the evolutionary fate (loss or retention) of a duplicated gene. These include gene pseudogenisation, subfunctionalisation, and neofunctionalisation (Figure 2).

Pseudogenisation is the degeneration of a gene into a nonfunctional gene (pseudogene) (Figure 2A). Duplication creates functional redundancy, as the resulting gene copies are initially identical. The redundancy of having two functional copies removes the selection pressures against mutations to one copy. The build-up of mutations can eventually turn one copy into a pseudogene, which may then become deleted from, or evolve into unidentifiable gene-fossils in the genome. This process has been demonstrated through population genetic modelling (Walsh 1995; Lynch et al. 2001) and genomic analysis (Lynch and Conery 2000; Harrison et al. 2002). We may still identify these pseudogenes from duplications by sequence similarity to its copy, if the duplication was recent enough.

Subfunctionalisation involves the partitioning of original gene functions between the two copies after duplication (Figure 2B, (Jensen 1976; Orgel 1977; Hughes 1994). In this scenario, duplicates lose their redundancy by diverging in function, leading to both copies being stably maintained in the genome (Nowak et al. 1997). Subfunctionalisation may involve the division of gene expression activity between the duplicates. For example, the zebrafish engrailed-1 and engrailed-1b transcription factors are duplicates that have diverged to be expressed in different tissues: the pectoral appendage bud and the neurons of the hindbrain/spinal cord, respectively (Force et al. 1999). The nonduplicated engrailed-1 gene in mouse is expressed across all tissues. Another scenario may be the partitioning of protein function. One copy may become specialised in one of the original functions. For example, specialised digestive enzymes in the leaf-eating monkey douc langur originated from the duplication of a bifunctional gene (Zhang et al. 2002).

Neofunctionalisation involves the novel gain of function in a duplicate copy, the most impactful scenario for the evolution of novel traits in a species (Figure 2C). The concept of adaptive evolution of novel function following a gene's duplication was hypothesised by Ohno (Ohno 1970). One duplicate copy, being functionally redundant, evolves under no or relaxed purifying selection pressure. Subsequent sequence mutations may lead to the gain of novel function. The random gain of a novel biological function may seem improbable, but examples of this happening exist. The two human RNase A genes, eosinophil-derived neurotoxin (EDN) and eosinophil cationic protein (ECP), originated from a gene duplication (Zhang et al. 1998). After duplication, the ECP gene through many arginine additions to the protein developed novel antibacterial activity absent in the original EDN gene (Rosenberg 1995). Neofunctionalisation often results in the evolution of related function rather than something completely novel. For example, the duplication of a human opsin gene gave rise to both red and green sensitive opsin genes, giving humans and related primates their sensitivity to a wider range of colours (Yokoyama and Yokoyama 1989). The amount of mutation needed to cause a functional change will vary from gene to gene. Many substitutions were probably needed for the ECP gene to evolve (Zhang et al. 1998), while there were mainly two substitutions responsible for the evolution of the opsin gene (Asenjo et al. 1994).

An additional scenario that should be mentioned is the selection for both duplications to be retained with the same function (Figure 2D). This may happen when it is beneficial to have an extra dosage of RNA or protein product from two copies instead of one. For example, genes with high demand products like rRNAs and histones. Purifying selection acts against modifying mutations to either copy, preventing divergence (Nei et al. 2000; Piontkivska et al. 2002). Retention of both copies may also happen when the two duplicates exist in a stoichiometric balance that is dosage sensitive (Veitia 2004).

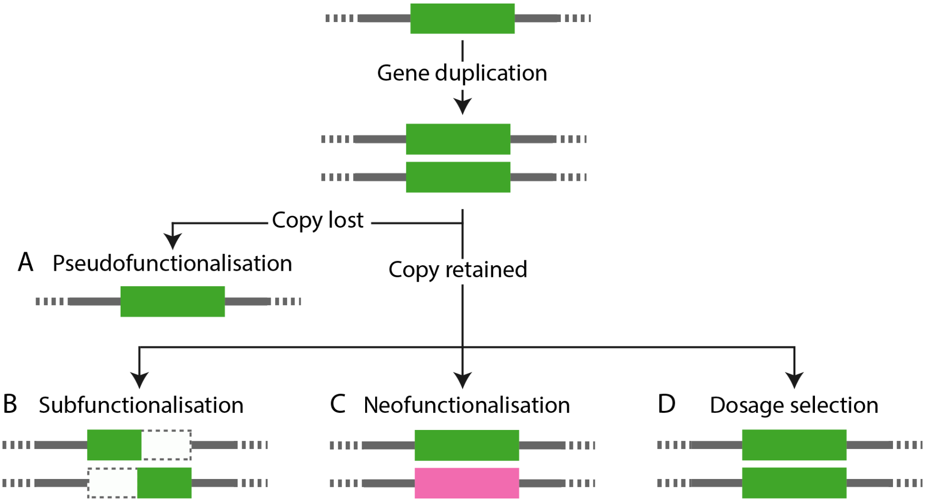


Figure 2: Fate of gene duplication. Following duplication, the most common outcome is the loss of a duplicated copy through (A) pseudofunctionalisation. If both copies have been selected to be retained it may be through (B) subfunctionalisation: the division of the original function between the copies, (C) neo-functionalisation: the gain of a new function in one copy, or (D) dosage selection: an advantage having multiples of the gene.

The evolutionary forces that act upon the duplicates control the divergence of function. Two models describe divergence with or without positive selection. First, the Dykhuizen-Hartl effect does not require positive selection for functional divergence (Kimura 1979; Zhang et al. 1998; Dykhuizen and Hartl 1980; Li 1983). In this model, random mutations are fixed in one duplicate copy under relaxed purifying selection, and this mutation may later confer a functional change in response to an environmental shift. Second, a model involving positive selection has two scenarios: either neutral mutations lead to a new function in one copy which is later refined and fixed by positive selection (Zhang et al. 1998), or each copy specialises in one ancestral function and positive selection refines this specialisation (Hughes 1994). When functional divergence is complete, each of the duplicates are likely maintained under different functional constraints (Gu 1999; Knudsen and Miyamoto 2001). The previous scenarios of duplicate loss (pseudogenisation) or functional gain (sub- or neofunctionalisation) also do not act independently, but may interact to determine the fate of a duplicated gene.

1.1.4 Changes to regulation of gene expression

Evolution of a gene's function may occur from changes to the protein coding region, or to regulatory control of the gene, changing when, where, or how much the gene is expressed. Evolution of gene regulation is known to play an important role in species evolution (King and Wilson 1975; Wang et al. 1996; Pierce and Crawford 1997; Ferea et al. 1999; Fraser et al. 2010; Berthelot et al. 2018), and has been considered as the major contributor to species differences, rather than protein sequence evolution (King and Wilson 1975; Wray 2007).

The expression of a gene is generally regulated by the regions of regulatory sequence surrounding it. These regions contain promoter and enhancer sequences that include regulatory elements that are recognized by numerous transcription factors in combination (Spitz and Furlong 2012; Moorthy et al. 2017; Shin et al. 2016). These transcription factors may recruit the transcriptional machinery that activate (transcribe) the gene to be expressed or tune the baseline expression level. Studies in mammals have shown that these regulatory regions can evolve to change the expression patterns of genes (Cotney et al. 2013; Xiao et al. 2012; Vierstra et al. 2014; Villar et al. 2015; Reilly et al. 2015; Young et al. 2015; Kunarso et al. 2010; Schmidt et al. 2010). Another mechanism that can change gene expression is related to chromatin structure. Chromosome regions may be tightly packed around histone proteins which reduces the accessibility of the DNA to transcription factors, and leads to the suppression of associated genes (Klemm et al. 2019).

1.2 Consequences of whole genome duplication in vertebrates

1.2.1 Genome duplication and the speciation of vertebrates

While whole genome duplication is a common occurrence for plants, it is a rare event to occur in animals (Van de Peer et al. 2009). There have been two WGDs at the base of all vertebrates referred as the 1R and 2R duplications (Dehal and Boore 2005). These duplication event are hypothesised to have shaped vertebrate lineages by driving speciation and the evolution of novel traits.

After a WGD the previous diploid individual has now become a tetraploid. This individual now has problems reproducing with the rest of the population. A tetraploid and diploid will produce triploid offspring that will likely be sterile because of problems segregating uneven chromosome numbers. The tetraploid genome is unstable, and will over time revert to a diploid state by the process of rediploidisation (Wolfe 2001). Paralogous chromosomes with high sequence similarity can easily cross over. Thus, rediploidisation critically relies on genomic changes that prevent cross over. Genomic rearrangements and gene losses may modify the ancestral structure and decrease the similarity between paralogous chromosomes over time. Reproductive isolation after WGD may drive speciation (Jaillon et al. 2009). A major factor for speciation is reciprocal gene loss, where one copy of an essential gene is lost in one population while another population losses the reciprocal copy. Offspring resulting from mating between these two populations then have a 1/16 chance of being a lethal double null homozygote. This chance increases in proportion to the number of essential gene copies that have been lost (Lynch and Conery 2000; Werth and Windham 1991). Reciprocal gene loss has been shown to have occurred between zebrafish and medaka (Naruse et al. 2004; Sémon and Wolfe 2007). In plants, there is a strong link between WGD events and increased speciation rates (Bowers et al. 2003; De Bodt et al. 2005; Magallón and Castillo 2009; Soltis et al. 2009; Soltis et al. 2014) Magallón and Castillo 2009), but in vertebrates such a link remains a hypothesis, requiring more empirical evidence.

A limitation of studying gene evolution from the vertebrate 1R and 2R WGDs is their extreme age (>600 mya (Vandepoele et al. 2004)), meaning that few gene duplications from these events can be identified reliably. There have however been several more subsequent WGDs in vertebrate fish lineages that are recent enough to facilitate study of the evolution of the gene duplicates arising from these vertebrate WGD.

1.2.2 Teleost and salmonid fish genome duplications

The ray-finned fish have diversified into more than 30,000 species, about half of all vertebrates, and inhabit a wide range of aquatic environments (Nelson 2006). The vast majority of these species belong to the infraclass teleost, an old diverse lineage spanning more than 400 million years (Near et al. 2012; Betancur-R et al. 2013; Broughton et al. 2013). Around 320-350 million years ago, after the teleosts diverged from the holostean lineage (containing gars and bowfins), a third WGD (3R) occurred (Smith et al. 2013; Jaillon et al. 2004; Kasahara et al. 2007; Nakatani et al. 2007). Roughly 12-24% of gene duplications are retained from the 3R WGD (Braasch and Postlethwait 2012). Links have been made between the 3R WGD and gene evolution, such as the expansion of gene family size and lineage specific expression evolution (Ahn et al. 2012; Braasch et al. 2009; Opazo et al. 2013; Voldoire et al. 2017). While the evolution of 3R duplicates may be studied, the limited number of remaining duplicates greatly restricts the power of such studies. With respect to the gene duplicate number we can study, the relatively recent fourth WGD in the salmonid lineage represents a better study system (Figure 3). Another benefit with the salmonid 4R study system is that the Esociformes provides a close sister lineage from which we can infer pre-4R ancestral gene function or regulation.

After their divergence from the Esociformes (containing their closest species: northern pike) about 125 million years ago, the salmonid lineage experienced a fourth WGD (4R) about >80 million years ago (Near et al. 2012; Macqueen and Johnston 2014). The relatively recent nature of the 4R duplication is evident by the fact that the Atlantic salmon genome is still in the process of rediploidisation from a tetraploid back to a diploid-behaving genome (Lien et al. 2016). About half of the 4R genes in salmonid genomes are still retained as duplicates, and parts of the duplicated salmonid genomes are still indistinguishable in sequence content (Lien et al. 2016; Robertson et al. 2017).

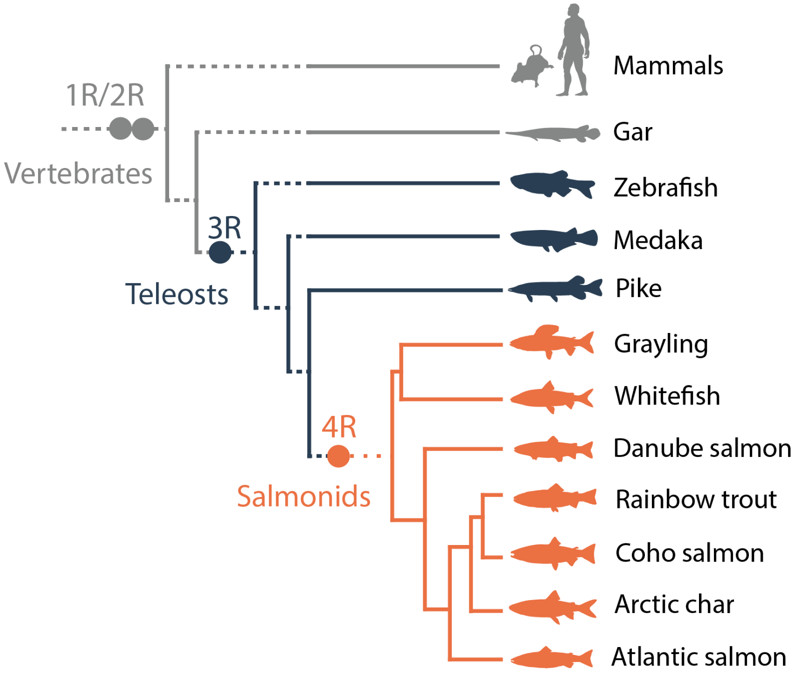


Figure 3: Whole genome duplication events leading to salmonids. Four whole genome duplications (WGD) occurred before the evolution of salmonid fish. The first two duplications (1R and 2R) were at the base of all vertebrates. The third (3R) was at the base of the teleost lineage after divergence from holostean lineage (gars). The fourth (4R) was at the base of the salmonid lineage after divergence from Esociformes (pike).

There have been some studies that have investigated the consequences of the 4R WGD on gene expression evolution in Atlantic salmon (Carmona-Antoñanzas et al. 2016; Lien et al. 2016). For 4R duplicates, the most common fate was no divergence in tissue expression profiles between duplicates, but cases of evolution of novel tissue regulation (i.e. regulatory neofunctionalisation) were also common, while subfunctionalisation was very rare (Lien et al. 2016; Sandve et al. 2018). It has been hypothesised that the increased activity of transposable elements seen after the 4R duplication in Atlantic salmon may have been an important mechanism both for rediploidisation and for expression evolution (Lien et al. 2016). Transposable elements aid to rearrange gene regulatory elements, thus changing the regulation of genes. One possible example of this in salmon 4R duplicates is the promotor regions of the Atlantic salmon fatty acid elongase 5 (*elovl5*) gene duplicates that have acquired different transposable element sequences and divergent gene regulatory mechanisms (Carmona-Antoñanzas et al. 2016; Carmona-Antoñanzas et al. 2014).

1.3 Approaches to study expression evolution of gene duplicates

Progressive research into gene expression evolution is now possible given the gains in sequencing technology allowing large scale expression studies across many species, tissues, and replicates. In addition, sequenced genomes are rapidly becoming available for more and more species. There have also been advances in approaches to studying expression evolution, within the context of single genome or comparative analysis across multiple genomes.

1.3.1 Finding gene duplicates

The first step to study duplicate evolution is to identify duplicated genes within a species´ genome (paralogs), and to identify genes with a common origin (orthologs) across multiple species. A standard approach to finding paralogs and orthologs is by protein sequence similarity. We can detect duplicate genes within a single genome, as well as orthologous genes between species, by finding the best reciprocal matches between protein sequences. Orthologs from the salmonid WGD can be identified more easily than older WGDs by their higher sequence similarity, given the shorter time since the duplication. We may detect if a given duplication occurred from a WGD by looking at the chromosome positions of the two copies, which should be on separate paralogous sections of chromosomes. Ortholog detection across species involves aligning all proteins, both within and between species, and clustering the best matches into groups of gene orthologs (orthogroups). From these orthogroups, we can refine the groups further by constructing gene trees from the sequence alignment of orthogroups and, using the position of proteins from outgroup species (i.e. rooting), find subsets of orthogroups (clades) with correct species phylogeny. Further, more detailed gene trees can be generated using the coding nucleotide sequence (CDS) for the proteins. Software such as OrthoFinder (Emms and Kelly 2015) generates orthogroups from sequence similarities, and can handle many species including those with a WGD event. With orthogroups we can find cases of gene duplication present across species by using the gene copy number within a given orthogroup. When gene duplication is retained the species with an extra duplication (e.g. salmonids) should have two gene copies in the orthogroup compared to one for species that did not undergo the duplication (e.g. 4R WGD). We can also find shared duplicate lost when all genes are in single copy, as well as mixtures of different lineage-specific duplicate retention and loss.

1.3.2 Measuring gene expression and sample normalisation

High-throughput RNA sequencing (RNA-Seq) is now the standard way of measuring gene expression. RNA extracted from a biological sample is fragmented into shorter sequences (fragments) that are amplified before their nucleotide bases are sequenced. The sequenced 'reads' are mapped to the species' transcriptome or genome sequence, and the number of reads mapped to a given gene is the gene's 'read count', a raw measure of transcript abundance. The read count value is often normalised to remove bias, accounting for the total number of reads sequenced for the sample, and the length of a gene's transcript. This allows the comparison of counts between samples or genes, respectively. Normalised counts could be calculated in Fragments Per Kilobases exon per Million reads (FPKM) or in Transcripts Per Million reads (TPM).

Expression values from RNA-Seq data are relative measurements: the raw read count for a given gene is proportional to the total number of reads that have been sequenced for a given sample. When comparing gene expression between different types of samples, from different tissues, conditions, or species, the landscape of the types of genes expressed (transcriptome) will be undoubtedly different. The composition of the RNA population influences the read counts, for example, if many genes are expressed uniquely in one experimental condition (tissue type, species), the sequencing depth for the remaining genes will be lower. This bias is commonly accounted for in methods for differential gene expression analysis. For example, the Trimmed Mean of M-values (TMM) normalisation method (Robinson and Oshlack 2010) is used in the software edgeR (Robinson et al. 2010) for normalisation. This method assumes that most genes between different samples are not differentially expressed. The average differences in gene expression between samples is measured by a weighted trimmed mean of the log expression ratios (trimmed mean of M values (TMM)). Then from the difference in TMM values, sample specific scaling factors are calculated to normalise average gene differences between samples. When conducting comparative transcriptomics analysis between species with and without a WGD, using a normalisation protocol like the TMM method is essential to account for expected gene expression differences based on the RNA landscape.

1.3.3 Analysis within the genome of single species

At the smallest scope, analysing duplicate expression evolution involves the comparison of the expression of one duplicate gene to its copy within the context of a single species' genome. For example, the comparison of the zebrafish engrailed-1 and engrailed-1b genes to find that they had different tissue expression profiles was done only using data from zebrafish. An important example for salmonids is the study of the fatty acid elongase gene *elovl5* in Atlantic salmon. A few papers have focused solely on this gene that has two functional copies, *elovl5a* and *elovl5b*, a result of the salmonid 4R WGD (Carmona-Antoñanzas et al. 2013; Carmona-Antoñanzas et al. 2016). These copies have evolved differences in their regulatory regions, a loss and gain of a transcription factor binding site in one copy, leading to a difference in tissue expression (a scenario of neofunctionalisation).

The next scope is the comparison of multiple genes within the same genome, perhaps a gene family with duplications or duplications within a specific pathway. This scope may be extended to the analysis of all gene duplicates in a species' genome, to investigate patterns in expression evolution after a WGD, and is often seen in genome papers (e.g. Atlantic salmon genome (Lien et al. 2016) and rainbow trout genome (Berthelot et al. 2014)). When the scope of the analysis is within a single species' genome, comparison of gene expression between duplicates is straightforward. Gene expression does not need to be normalised for species differences, and duplicate genes need only be identified in one genome. Although analysis of a single species is straightforward and can lead to interesting insights for that species, comparative analysis across multiple species adds greater insights into evolutionary processes following WGD.

1.3.4 Analysis across the genomes of multiple species

At a larger scope, duplicate expression evolution may be analysed across the genomes of multiple species. This provides some unique information about how gene duplicates have evolved, such as where in a lineage did the evolution occur (judging by the presence of the duplicate across related species), the degree of conservation in multiple species (retention across species suggests functional importance), and distinguishing the ancestral function from the evolved function. For example, the evolutionary fates of the *sox* gene family after the teleost 3R WGD was investigated by comparing duplicate copy number and expression level retention across multiple species, finding instances of species-specific differences in duplicate tissue expression patterns (Voldoire et al. 2017) The PhyloFish database was made available for such cross-species transcriptome comparisons across the WGDs in fish lineages. In a case study, they highlight species-specific differences in the tissue expression patterns of the *sta8* gene (Pasquier et al. 2016). Understanding the ancestral state of a gene duplicate is essential for describing how it has evolved. In the example previously given for the zebrafish engrailed-1 and engrailed-1b genes, these genes were described as subfunctionalised by comparing their tissue expression profiles to the state of the mouse engrailed-1 ortholog. In studies on *elovl5* duplication in Atlantic salmon, the duplicate copies are compared to the closest species without the salmonid 4R WGD, pike. Comparison to pike highlights how the salmon duplicates had evolved specialised expression in liver (*elovl5b*) or intestine (*elovl5a*), which is possibly an adaptation to an invertebrate rich diet that young salmon go through that is poor in essential omega-3 lipids (*elovl5* is involved in LC-PUFA biosynthesis: Carmona-Antoñanzas et al. 2013; Carmona-Antoñanzas et al. 2016). In the Atlantic salmon genome paper (Lien et al. 2016) the tissue expression profiles of duplicates from the salmonid 4R WGD were compared to pike to find cases of neo- or subfunctionalisation in tissue regulation.

Cross-species comparisons can be very informative, but there are limitations with this approach. These studies can be classified as a pairwise approach because species pairs are being compared independently without utilising information on their evolutionary relationship (Dunn et al. 2018). A more sophisticated way to detect gene expression evolution across multiple species is to model gene expression as a trait using evolutionary models.

1.3.5 Modelling expression changes

Earlier comparative approaches with many species typically relied on traditional ANOVA tests to detect genes with significant expression divergence (Nuzhdin et al. 2004; Gilad et al. 2006; Khaitovich et al. 2006; Whitehead and Crawford 2006). This approach may account for variation within species, but ignores the evolutionary relationships between species, treating them as independent. Evolutionary models have been specifically developed to account for evolutionary relationships and time.

We may consider gene expression levels as a quantitative trait that can evolve over time across a phylogeny, and thus make use of evolutionary models for trait evolution for modelling gene expression evolution. The Ornstein-Uhlenbeck (OU) process, proposed by Hansen (Hansen 1997), is such an evolutionary model. It is a stochastic process that models the accumulation of random changes in gene expression levels over time (random walk), but unlike a similar process involving random walk, Brownian motion, OU assumes that for a given gene there is a biologically optimum level for the gene to be expressed at, and bounds exist surrounding this optimum, creating an acceptable range that expression variation is constrained to. The stabilising selection pressure of having these bounds means that expression variation increases less and less over time (i.e. non-linear relationship). We can assess to what degree the OU assumptions fit the expression data for a given set of species by comparing the trend of expression distance between species to their evolutionary distance (e.g. sequence substitutions). In fruit fly, unlike changes to sequence, divergence of gene expression was not continuously linear over time, but reached a saturation point because of a stabilizing selection pressure, supporting the use of a model like the OU process over a standard neutral drift model (Bedford and Hartl 2009; Kalinka et al. 2010).

The OU process has been used so far to infer fitness and selection of evolving expression levels (Bedford and Hartl 2009; Kalinka et al. 2010; Nourmohammad et al. 2017) and applied to detect selection on gene expression across mammalian phylogenies (Brawand et al. 2011; Rohlfs and Nielsen 2015). It has also been used to predict if expression is evolving under adaptive or neutral selection (Chen et al. 2019). An extension of the OU process involves incorporating the biological variance within- and between-species, similar to an ANOVA test. This species variance is incorporated together with the OU process in the Expression Variance and Evolution model (EVE) (Rohlfs and Nielsen 2015). This model enables the comparison of the likelihoods of different evolutionary hypotheses: e.g. if the optimum expression level of a duplicate has diverged from the ancestral optimum or not. The lineage specific hypothesis testing that is supported by EVE is ideal for testing if genes on the salmonid branch have experienced increased levels of expression evolution or not after the 4R WGD compared to species that did not undergo the WGD.

1.4 Aim of this thesis

The aim of this thesis was to investigate the effect of WGD on vertebrate gene expression evolution, and to test to what degree gene duplication promotes adaptive evolution. We used the salmonid 4R WGD as a system to study the consequences of a relatively recent WGD on evolution of duplicate gene expression. We used existing genomic data and supplement this by generating novel genomic and transcriptomic data. We made use of various bioinformatic approaches to transcriptome analysis, starting from a comparative analysis between duplicates within a single genome, moving to a comparative analysis between duplicates across pairs of genomes, to lastly modelling gene expression as an evolutionary trait across many species. This research presented novel findings about expression evolution in salmonids that aid understanding of vertebrate gene and species evolution.

2

Paper summaries

In Paper I we investigate the consequences of the salmonid 4R WGD on the regulation of genes in Atlantic salmon lipid metabolism pathways. In Paper II we investigate the consequences of the 4R WGD on the regulation of genes in European grayling and Atlantic salmon lineages. In Paper III, we investigate gene expression evolution in the salmonid lineage by testing for shifts in expression between multiple species with and without the 4R WGD.

2.1 Paper I – Life-stage-associated remodelling of lipid metabolism regulation in Atlantic salmon

Atlantic salmon plays a central role in the understanding of expression evolution following the salmonid 4R WGD. Given the economic importance of Atlantic salmon for fishing and aquaculture industries, a lot of research is conducted on Atlantic salmon, especially towards the improvement of omega-3 content in farmed salmon. This interest has resulted in a high-quality genome and a plethora of transcriptomic data (Lien et al. 2016).

In this paper, we produce comprehensive gene annotations for lipid metabolism pathways for Atlantic salmon and gene expression data from a feeding trial with contrasting diets to facilitate research on omega-3 biosynthesis. The feeding experiment showed the effects that high or low omega-3 precursors in the diet had on the regulation of Atlantic salmon lipid pathways in the liver and gut during both fresh- and saltwater life-stages. We found life-stage associated remodelling of lipid metabolism from liver centric in freshwater to gut centric in saltwater. Genes relating to lipogenesis and lipid transport in liver decrease in expression and become less responsive to diet, while genes for lipid uptake in gut becomes more highly expressed.

Evolution acting upon 4R duplicates has been suggested to have adaptively increased the potential for omega-3 biosynthesis (*elovl5*: Carmona-Antoñanzas et al. 2013; Carmona-Antoñanzas et al. 2016). We thus investigated the consequences of the 4R duplications on Atlantic salmon lipid pathways. We found that more genes in lipid pathways were retained in duplicates compared to all genes, and that duplicate retention varies between lipid pathways. Moreover, we showed that pathways differ in how many duplicates had correlated expression profiles during the feeding trial. Regulatory conservation was not associated with duplicate retention, e.g. 'biosynthesis of unsaturated fatty acids' was a pathway with fewer duplicates retained, but more duplicates with highly similar expression. We investigated relationships between gene duplication and increased gene dosage using the expression of northern pike orthologs as the assumed ancestral dosage level. For three lipid pathways (including 'biosynthesis of unsaturated fatty acids') we found several genes (*hadhab*, *elovl6*, and *elovl5)* showing a link between duplicate co-expression and higher total gene dosage. The signatures we found of pathway-specific selection pressure on gene duplicates, including increased gene dosage in three genes involved in fatty acid metabolism, illustrates possible adaptive consequences of the salmonid 4R WGD on evolution of lipid metabolism.

2.2 Paper II – The grayling genome reveals selection on gene expression regulation after whole-genome duplication

Genome studies have shown that rediploidisation of paralogous chromosomes after the salmonid 4R WGD temporally overlaps with species radiation, resulting in most gene duplicates (75%) having diverged in sequence before salmonid speciation (ancestral ohnolog resolution, AORe), while some duplicates (25%) have diverged after speciation in a species-specific manner (lineage-specific ohnolog resolution (LORe)) (Robertson et al. 2017). This process provides potential for differences between the salmonid genomes to evolve. Salmonids split into the Salmoninae and Thymallinae clades, that evolved different genome structures, and ecological adaptations. The former includes species such as Atlantic salmon that evolved the capacity to migrate between fresh- and saltwater. The later includes European grayling which does not migrate to saltwater. The unique combination of shared and lineage-specific duplicate divergence and different life-style adaptations between these salmonid clades provides an ideal system to study evolutionary consequences of WGD.

To study effects of the 4R duplication on grayling and Atlantic salmon comparatively, we generated an annotated genome assembly for grayling. We identified duplicate gene orthologs across grayling and Atlantic salmon, and compared tissue expression profiles between duplicates to assign orthologs into different evolutionary scenarios including conserved expression, ancestral divergence, and species-specific divergence.

About a third of the duplicates reflected nonneutral tissue expression evolution, with strong purifying selection maintained over the ~50 million years since grayling and

Atlantic salmon lineages diverged. Of these, the majority reflected conserved tissue regulation, including genes enriched in brain and neural functions along with higher-order protein-protein interactions. A small subset of duplicates showed evidence of ancestral duplicate divergence in tissue expression that has been maintained since the speciation, which suggests adaptive divergence following WGD. The candidate duplicates for adaptive tissue expression divergence had elevated rates of protein coding and promoter sequence evolution, and are enriched for immune and lipid metabolism functions. Lineage-specific duplicate divergence points towards underlying differences in adaptive pressures in the two species and highlights cases of regulatory divergence of salmonid 4R duplicates, possibly related to a niche shift in early salmonid evolution.

2.3 Paper III – Gene regulatory evolution following salmonid whole genome duplication

The salmonid 4R WGD presents an ideal system for testing Ohno's hypothesis: that genes undergo adaptive evolution more readily when selection pressures are relaxed due to the redundancy provided by duplication. The abundance of genome and transcriptome data for salmonid and other teleost species now makes it possible to apply comparative transcriptomics across many species both with and without the 4R WGD. Novel approaches to detect expression evolution is now available, including methods using proper evolutionary models to test evolutionary hypotheses. We present a first attempt to apply novel methods to analyse gene evolution after salmonid WGD.

We use liver expression data from seven species to detect significant shifts in a gene's expression in the salmonid lineage compared to ancestral expression levels as observed in species that did not experience the 4R WGD. We identify gene orthologs across the selected species that have retained or lost their duplication, and used the Expression Variance and Evolution (EVE) method to test for expression shifts.

We revealed that proportionately more salmonid duplicates shifted in expression (26%) compared to salmonid singletons (16%), and compared to individual teleost outgroup species (6-10%), indicating that the redundancy produced by the 4R WGD has acted as a catalyst for expression divergence. Most of the shifts for duplicates was a shift down in expression level (62%), possibly explained by one duplicate evolving under relaxed selection pressure towards pseudogenisation. However, further analysis using a tissue atlas co-expression network go against that pseudogenisation is the major driver of evolutionary down tuning of gene duplicate expression levels. Instead it seems likely that strong selection on some form of gene product dosage balance has been important post 4R WGD for genes in liver. Functional enrichment in diverged genes highlighted lipid metabolism-related functions in duplicates that had one copy shifted up in expression, including three elongase genes (*elovl1*, *elovl5*, and *elovl6*). These genes and others present potential cases of adaptive regulatory neofunctionalisation of salmonid duplicates.

3

Discussion

There are various approaches to characterising gene expression differences (Hermansen et al. 2016): (a) the binary comparison of expression, i.e. 'on-off', (b) a differential expression (DE) tests, (c) and comparing the correlation of expression patterns, the results of which may be (d) clustered into groups representing co-expression. Lastly (e) in studies with many species we may use evolutionary modelling for comparisons in a phylogenetic context. Throughout the papers in this thesis, there was a progression of different bioinformatic methodologies used for comparing differences in gene expression. Our analysis started off within the context of a single salmonid genome in paper I, moved on to a comparison between a pair of salmonids in paper II, and finally ended up with a phylogenetic analysis of multiple salmonid and outgroup species in paper III. The different methodologies had their advantages and limitations, along with challenges in implementation which are discussed below.

Our first analysis of salmonid 4R duplicate expression, in paper I, was within the scope of a single salmonid genome, Atlantic salmon. We used the correlation of expression patterns over a feeding trial in liver to compare how duplicated genes were coregulated. While single genome analysis has revealed many insights about expression divergence of duplicates stemming from WGD, the scope and support of such inference is limited to the context of only that species. A comparative analysis across multiple species in a given lineage (e.g. salmonids), like in papers II and III, gives more certainty about lineage evolutions (e.g. 4R effects on salmonids). Gene expression divergence collaborated by several species in a lineage adds support to predictions of adaptive evolution, as similar expression changes conserved over multiple species are less likely to be neural.

In paper I we also leveraged expression levels of pike orthologs to investigate dosage effects from duplication. Pike has been used as an outgroup in papers I and II, as well as previous studies (Braasch et al. 2016; Lien et al. 2016; Varadharajan et al. 2018), representing a proxy for pre-4R ancestral expression levels. Outgroup species are important for insight on how duplicate expression has evolved, telling us about the direction of the change in expression and if evolution occurred in one or both duplicates. We extended the number of outgroup species to three in paper III, which gives more certainty about the pre-4R ancestral state and more statistical power.

We widen our scope in paper II when we compared duplicate expression between two salmonid species, grayling and Atlantic salmon. We used correlation-based clustering of orthologs duplicated in both species to divide duplicates into groups based on multi-tissue expression data, and assigned duplicates as conserved or ancestral/lineage diverged depending on the groups expression profile (i.e. evolutionary scenarios). Other studies have also previously compared tissue expression profiles, directly (Pasquier et al. 2016) or based on correlation clustering (Lien et al. 2016). A limitation often seen in these studies is the lack of biological replication (Lien et al. 2016; Pasquier et al. 2016), which was also a problem for paper II. We will likely see such multi-tissue datasets with replication in the future as sequencing costs decrease, but right now being able to sequence multiple tissues from multiple individuals at an appropriate depth has often been at the cost of sample replication. We relied instead on expression differences across tissue types being greater than individual variation within a single tissue, meaning that we assumed that gene assignment to clusters would be robust to biological variation. This, however, is not necessarily the case for all genes, meaning that the results from this method can be unreliable for genes with high variance. A problem in paper II was that most duplicates remained unclassified as their cross-species expression profiles displayed no interpretable evolutionary scenario.

In paper II, we went further that previous studies (Lien et al. 2016) in validating the classifications of diverged duplicates from the correlation analysis. Here we tested for differential expression between duplicates using replicated liver data, and showed that the shifts in liver expression characteristic to the tissue group with dominant expression in liver, generally were supported by statistically significant changes in liver expression between diverged duplicates. This analysis in paper II shows how a combination of comparative transcriptomic methods may add confidence to the results, especially when there are limitations with one part, such as the lack of sample replication for multiple tissues.

What was missing from the previous analysis in papers I and II, was a formal statistical framework to test for adaptive over neutral expression evolution of duplicates after the 4R WGD (Sandve et al. 2018). In paper III the comparative analysis involved modelling expression evolution across orthologs from more salmonid and outgroup species using the Expression Variance and Evolution (EVE) model, an extension of the Ornstein-Uhlenbeck (OU) process that allows for integrating within-species variation by leveraging sample replication (Rohlfs and Nielsen 2015; Rohlfs et al. 2014). This allowed a statistical comparison of alternative evolutionary hypotheses for duplicate expression evolution, accounting for biological and evolutionary variance. Although expression modelling has been previously used to detect expression evolution (Bedford and Hartl 2009; Kalinka et al. 2010; Perry et al. 2012; Chen et al. 2019), the application here to a phylogeny with a WGD is novel, and we found that this presented several methodological challenges detailed below.

We had to devise a novel approach for using expression modelling given a ortholog tree containing duplications. We first considered duplicates in the same orthogroup as separate species, and testing each duplicate branch for divergence in expression from the other duplicate branch and from the outgroup species. The problems with this approach was first, the expression of duplicates are not necessary independent as there may be dosage balancing effects in play, and second, we were concerned with a difference in statistical power between testing duplicate orthogroups and testing singleton orthogroups due to the presence of more orthologs in the duplicate groups. Therefore, we favoured to solely use outgroup species as the ancestral expression level (same approach in papers I & II, and (Lien et al. 2016), where pike is the outgroup) and opted to independently test each salmonid duplicate branch for divergence in expression, retaining the same outgroup data for both duplicates.

While the relatively short time since the 4R WGD (~80 mya) makes identification of gene duplicates easier than other vertebrate WGD events, species-specific gene loss still introduces problems in this analysis. The current implementation of EVE requires every orthogroup to be complete, that is, singleton orthogroups must contain exactly one ortholog for every species, and duplicate orthogroups exactly one ortholog for every outgroup species and exactly two orthologs from every salmonid. Such complete orthogroups becomes increasing unlikely the more species that are analysed. We observed a decrease in the number of complete orthogroups when using more species, especially with certain salmonid species such as grayling and Danube salmon, with especially fragmented genome assemblies, causing unreasonable limitations in the number of orthogroups that could be analysed. Thus, we opted to not include such possible species in the EVE analysis in favour of analysing more orthogroups. This is a current major limitation of comparative analysis using EVE. In contrast, the paper I comparison of duplicates within a single salmonid genome, or the paper II comparison of duplicate orthologs across two salmonid genomes, was not limited to this degree and many more duplicates were analysed in these papers. Increasing the number of species in the analysis is desirable for statistical power and biological interpretation, but it is at odds with the number of complete orthogroups that can be analysed. The ability to handle missing ortholog data will be critical for future comparative studies.

In paper III, where we analysed species with and without the 4R WGD, we took much consideration into normalising expression data between the species. We were concerned that differences in genome sizes because of the 4R WGD would create a bias when we compared expression levels. We did not have the same concern in paper II, as we compared patterns of expression and not levels. We opted to settle on normalising expression data between species by comparing the expression of singleton orthologs across all species, and calculated normalisation factors (TMM normalisation method: Robinson and Oshlack 2010). The assumption was that singletons are maintained under selection pressure to have more similar expression levels than duplicates. Interestingly, other differences in the transcriptional landscape between species seems to have more influence on the expression distribution than the number of genes in the genomes. This is not unreasonable, as the number of annotated genes in the genome does not determine the number genes expressed in a given tissue/condition, or their level of expression. Although a complex and unsolved problem, we have made reasonable efforts to normalise between species. Understanding how the transcriptional landscape influences gene expression is increasingly needed as we analyse more phylogenetically diverse species, especially across WGD events.

3.1 Future perspectives

While some of the problems identified above have straightforward solutions, like the lack of replication of tissue expression data may be soon solved from more transcriptomic studies, the other challenges, mostly associated with comparative analysis in paper III, will require more thought and effort to solve. We argue that the most critical problem is to handle missing species in orthogroups, something not currently supported in comparative analysis using EVE. The future of comparative transcriptome analysis requires a dynamic test that can handle null expression values when orthologs are missing in a species, or in a perfect scenario, allows testing of various orthogroup structures. Developing appropriate methods for normalising expression data between species with large genome differences will be important for all comparative analysis between diverse species. Lastly, comparative analysis using expression modelling could be extended to include expression profile data (e.g. multi-tissue data) as suggested below.

The current implementation of EVE is designed to test for expression shifts across species, given replicates in a single condition/tissue. Our analysis in paper III for example is limited to the scope of genes expressed in liver. Ideally we would like to analyse samples from multiple tissues at once to detect expression changes within and across tissues. This is currently outside the scope of EVE, but perhaps a multivariate OU model (Beaulieu et al. 2012) may be implemented to analyse multi-tissue or other gradient expression data, ideally retaining how EVE accounts for within-species variation using sample replication. The multiple dimensions of data from this kind of analysis may also present a challenge in interpreting results.

Advancements in sequencing and bioinformatic methods of comparative transcriptomics have enabled novel research into systems such as expression evolution of duplicates in the salmonid lineage. Over time as more genomic and transcriptomic data becomes available and challenges with comparative transcriptomics are met with solutions, a clearer picture will emerge about how the salmonid 4R WGD has shaped gene expression evolution, and how duplication contributes in general to the evolution of vertebrates.

References

Ahn, D., You, K.-H. and Kim, C.-H. 2012. Evolution of the tbx6/16 subfamily genes in vertebrates: insights from zebrafish. *Molecular Biology and Evolution* 29(12), pp. 3959–3983.

Arabidopsis Genome Initiative 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408(6814), pp. 796–815.

Asenjo, A.B., Rim, J. and Oprian, D.D. 1994. Molecular determinants of human red/green color discrimination. *Neuron* 12(5), pp. 1131–1138.

Beaulieu, J.M., Jhwueng, D.-C., Boettiger, C. and O'Meara, B.C. 2012. Modeling stabilizing selection: expanding the Ornstein-Uhlenbeck model of adaptive evolution. *Evolution* 66(8), pp. 2369–2383.

Bedford, T. and Hartl, D.L. 2009. Optimization of gene expression by natural selection. *Proceedings of the National Academy of Sciences of the United States of America* 106(4), pp. 1133–1138.

Berthelot, C., Brunet, F., Chalopin, D., et al. 2014. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nature Communications* 5, p. 3657.

Berthelot, C., Villar, D., Horvath, J.E., Odom, D.T. and Flicek, P. 2018. Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nature ecology & evolution* 2(1), pp. 152–163.

Betancur-R, R., Broughton, R.E., Wiley, E.O., et al. 2013. The tree of life and a new classification of bony fishes. *PLoS Currents. Influenza* 5.

Bowers, J.E., Chapman, B.A., Rong, J. and Paterson, A.H. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422(6930), pp. 433–438.

Braasch, I., Brunet, F., Volff, J.-N. and Schartl, M. 2009. Pigmentation pathway evolution after whole-genome duplication in fish. *Genome Biology and Evolution* 1, pp. 479–493.

Braasch, I., Gehrke, A.R., Smith, J.J., et al. 2016. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nature Genetics* 48(4), pp. 427–437.

Braasch, I. and Postlethwait, J.H. 2012. Polyploidy in Fish and the Teleost Genome Duplication. In: Soltis, P. S. and Soltis, D. E. eds. *Polyploidy and Genome Evolution*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 341–383.

Brawand, D., Soumillon, M., Necsulea, A., et al. 2011. The evolution of gene expression levels in mammalian organs. *Nature* 478(7369), pp. 343–348.

Bridges, C.B. 1936. The bar "gene" a duplication. *Science* 83(2148), pp. 210–211.

Broughton, R.E., Betancur-R, R., Li, C., Arratia, G. and Ortí, G. 2013. Multi-locus phylogenetic analysis reveals the pattern and tempo of bony fish evolution. *PLoS Currents. Influenza* 5.

Carmona-Antoñanzas, G., Tocher, D.R., Martinez-Rubio, L. and Leaver, M.J. 2014. Conservation of lipid metabolic gene transcriptional regulatory networks in fish and mammals. *Gene* 534(1), pp. 1–9.

Carmona-Antoñanzas, G., Tocher, D.R., Taggart, J.B. and Leaver, M.J. 2013. An evolutionary perspective on Elovl5 fatty acid elongase: comparison of Northern pike and duplicated paralogs from Atlantic salmon. *BMC Evolutionary Biology* 13, p. 85.

Carmona-Antoñanzas, G., Zheng, X., Tocher, D.R. and Leaver, M.J. 2016. Regulatory divergence of homeologous Atlantic salmon elovl5 genes following the salmonid-specific whole-genome duplication. *Gene* 591(1), pp. 34–42.

Chen, J., Swofford, R., Johnson, J., et al. 2019. A quantitative framework for characterizing the evolutionary history of mammalian gene expression. *Genome Research* 29(1), pp. 53–63.

Cotney, J., Leng, J., Yin, J., et al. 2013. The evolution of lineage-specific regulatory activities in the human embryonic limb. *Cell* 154(1), pp. 185–196.

De Bodt, S., Maere, S. and Van de Peer, Y. 2005. Genome duplication and the origin of angiosperms. *Trends in Ecology & Evolution* 20(11), pp. 591–597.

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

Dunn, C.W., Zapata, F., Munro, C., Siebert, S. and Hejnol, A. 2018. Pairwise comparisons across species are problematic when analyzing functional genomic data. *Proceedings of the National Academy of Sciences of the United States of America* 115(3), pp. E409–E417.

Dykhuizen, D. and Hartl, D.L. 1980. Selective neutrality of 6PGD allozymes in E. coli and the effects of genetic background. *Genetics* 96(4), pp. 801–817.

Emms, D.M. and Kelly, S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* 16, p. 157.

Ferea, T.L., Botstein, D., Brown, P.O. and Rosenzweig, R.F. 1999. Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proceedings of the National Academy of Sciences of the United States of America* 96(17), pp. 9721–9726.

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.

Fraser, H.B., Moses, A.M. and Schadt, E.E. 2010. Evidence for widespread adaptive evolution of gene expression in budding yeast. *Proceedings of the National Academy of Sciences of the United States of America* 107(7), pp. 2977–2982.

Gilad, Y., Oshlack, A. and Rifkin, S.A. 2006. Natural selection on gene expression. *Trends in Genetics* 22(8), pp. 456–461.

Gu, X. 1999. Statistical methods for testing functional divergence after gene duplication. *Molecular Biology and Evolution* 16(12), pp. 1664–1674.

Gu, Z., Cavalcanti, A., Chen, F.-C., Bouman, P. and Li, W.-H. 2002. Extent of gene duplication in the genomes of Drosophila, nematode, and yeast. *Molecular Biology and Evolution* 19(3), pp. 256–262.

Hansen, T.F. 1997. Stabilizing selection and the comparative analysis of adaptation. *Evolution* 51(5), pp. 1341–1351.

Harrison, P.M., Hegyi, H., Balasubramanian, S., et al. 2002. Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Research* 12(2), pp. 272–280.

Hermansen, R.A., Hvidsten, T.R., Sandve, S.R. and Liberles, D.A. 2016. Extracting functional trends from whole genome duplication events using comparative genomics. *Biological procedures online* 18, p. 11.

Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B.C. and Herrmann, R. 1996. Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae. *Nucleic Acids Research* 24(22), pp. 4420–4449.

Hughes, A.L. 1994. The evolution of functionally novel proteins after gene duplication. *Proceedings of the Royal Society of London …*.

Hughes, A.L. and Nei, M. 1989. Evolution of the major histocompatibility complex: independent origin of nonclassical class I genes in different groups of mammals. *Molecular Biology and Evolution* 6(6), pp. 559–579.

Jaillon, O., Aury, J.-M., Brunet, F., et al. 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature* 431(7011), pp. 946–957.

Jaillon, O., Aury, J.-M. and Wincker, P. 2009. "Changing by doubling", the impact of Whole Genome Duplications in the evolution of eukaryotes. *Comptes Rendus Biologies* 332(2–3), pp. 241–253.

Jensen, R.A. 1976. Enzyme recruitment in evolution of new function. *Annual Review of Microbiology* 30, pp. 409–425.

Kalinka, A.T., Varga, K.M., Gerrard, D.T., et al. 2010. Gene expression divergence recapitulates the developmental hourglass model. *Nature* 468(7325), pp. 811–814.

Kasahara, M., Naruse, K., Sasaki, S., et al. 2007. The medaka draft genome and insights into vertebrate genome evolution. *Nature* 447(7145), pp. 714–719.

Khaitovich, P., Enard, W., Lachmann, M. and Pääbo, S. 2006. Evolution of primate gene expression. *Nature Reviews. Genetics* 7(9), pp. 693–702.

Kimura, M. 1979. The neutral theory of molecular evolution. *Scientific American* 241(5), p. 98–100, 102, 108 passim.

King, M.C. and Wilson, A.C. 1975. Evolution at two levels in humans and chimpanzees. *Science* 188(4184), pp. 107–116.

Klemm, S.L., Shipony, Z. and Greenleaf, W.J. 2019. Chromatin accessibility and the regulatory epigenome. *Nature Reviews. Genetics* 20(4), pp. 207–220.

Knudsen, B. and Miyamoto, M.M. 2001. A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. *Proceedings of the National Academy of Sciences of the United States of America* 98(25), pp. 14512–14517.

Kunarso, G., Chia, N.-Y., Jeyakani, J., et al. 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature Genetics* 42(7), pp. 631–634.

Lespinet, O., Wolf, Y.I., Koonin, E.V. and Aravind, L. 2002. The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Research* 12(7), pp. 1048–1059.

Li, W.H. 1983. Evolution of duplicate genes and pseudogenes. In: Nei, M. and Koehn, R. K. eds. *Evolution of Genes and Proteins*. Sunderland, MA: Sinauer Associates, pp. 14–37.

Li, W.H. 1997. Molecular evolution.

Li, W.H., Gu, Z., Wang, H. and Nekrutenko, A. 2001. Evolutionary analyses of the human genome. *Nature* 409(6822), pp. 847–849.

Lien, S., Koop, B.F., Sandve, S.R., et al. 2016. The Atlantic salmon genome provides insights into rediploidization. *Nature* 533(7602), pp. 200–205.

Long, M. 2001. Evolution of novel genes. *Current Opinion in Genetics & Development* 11(6), pp. 673–680.

Lynch, M. and Conery, J.S. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290(5494), pp. 1151–1155.

Lynch, M., O'Hely, M., Walsh, B. and Force, A. 2001. The probability of preservation of a newly arisen gene duplicate. *Genetics* 159(4), pp. 1789–1804.

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. Biological Sciences / the Royal Society* 281(1778), p. 20132881.

Magallón, S. and Castillo, A. 2009. Angiosperm diversification through time. *American Journal of Botany* 96(1), pp. 349–365.

Mombaerts, P. 2001. The human repertoire of odorant receptor genes and pseudogenes. *Annual Review of Genomics and Human Genetics* 2, pp. 493–510.

Moorthy, S.D., Davidson, S., Shchuka, V.M., et al. 2017. Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes. *Genome Research* 27(2), pp. 246–258.

Nakatani, Y., Takeda, H., Kohara, Y. and Morishita, S. 2007. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Research* 17(9), pp. 1254–1265.

Naruse, K., Tanaka, M., Mita, K., Shima, A., Postlethwait, J. and Mitani, H. 2004. A medaka gene map: the trace of ancestral vertebrate proto-chromosomes revealed by comparative gene mapping. *Genome Research* 14(5), pp. 820–828.

Near, T.J., Eytan, R.I., Dornburg, A., et al. 2012. Resolution of ray-finned fish phylogeny and timing of diversification. *Proceedings of the National Academy of Sciences of the United States of America* 109(34), pp. 13698–13703.

Nei, M. 1969. Gene Duplication and Nucleotide Substitution in Evolution. *Nature* 221(5175), pp. 40–42.

Nei, M., Rogozin, I.B. and Piontkivska, H. 2000. Purifying selection and birth-and-death evolution in the ubiquitin gene family. *Proceedings of the National Academy of Sciences of the United States of America* 97(20), pp. 10866–10871.

Nelson, J. 2006. *Fishes of the World*.

Nourmohammad, A., Rambeau, J., Held, T., Kovacova, V., Berg, J. and Lässig, M. 2017. Adaptive evolution of gene expression in drosophila. *Cell reports* 20(6), pp. 1385–1395.

Nowak, M.A., Boerlijst, M.C., Cooke, J. and Smith, J.M. 1997. Evolution of genetic redundancy. *Nature* 388(6638), pp. 167–171.

Nuzhdin, S.V., Wayne, M.L., Harmon, K.L. and McIntyre, L.M. 2004. Common pattern of evolution of gene expression level and protein sequence in Drosophila. *Molecular Biology and Evolution* 21(7), pp. 1308–1317.

Ohno, S. 1970. *Evolution by Gene Duplication*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Ohno, S. 2013. Sex chromosomes and sex-linked genes.

Opazo, J.C., Butts, G.T., Nery, M.F., Storz, J.F. and Hoffmann, F.G. 2013. Whole-genome duplication and the functional diversification of teleost fish hemoglobins. *Molecular Biology and Evolution* 30(1), pp. 140–153.

Orgel, L.E. 1977. Gene-duplication and the origin of proteins with novel functions. *Journal of Theoretical Biology* 67(4), p. 773.

Pasquier, J., Cabau, C., Nguyen, T., et al. 2016. Gene evolution and gene expression after whole genome duplication in fish: the PhyloFish database. *BMC Genomics* 17, p. 368.

Perry, G.H., Melsted, P., Marioni, J.C., et al. 2012. Comparative RNA sequencing reveals substantial genetic variation in endangered primates. *Genome Research* 22(4), pp. 602–

610.

Pierce, V.A. and Crawford, D.L. 1997. Phylogenetic analysis of glycolytic enzyme expression. *Science* 276(5310), pp. 256–259.

Piontkivska, H., Rooney, A.P. and Nei, M. 2002. Purifying selection and birth-and-death evolution in the histone H4 gene family. *Molecular Biology and Evolution* 19(5), pp. 689–697.

Reilly, S.K., Yin, J., Ayoub, A.E., et al. 2015. Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science* 347(6226), pp. 1155–1159.

Robertson, F.M., Gundappa, M.K., Grammes, F., et al. 2017. Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification. *Genome Biology* 18(1), p. 111.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26(1), pp. 139–140.

Robinson, M.D. and Oshlack, A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* 11(3), p. R25.

Rohlfs, R.V., Harrigan, P. and Nielsen, R. 2014. Modeling gene expression evolution with an extended Ornstein-Uhlenbeck process accounting for within-species variation. *Molecular Biology and Evolution* 31(1), pp. 201–211.

Rohlfs, R.V. and Nielsen, R. 2015. Phylogenetic ANOVA: the expression variance and evolution model for quantitative trait evolution. *Systematic Biology* 64(5), pp. 695–708.

Rosenberg, H.F. 1995. Recombinant human eosinophil cationic protein. Ribonuclease activity is not essential for cytotoxicity. *The Journal of Biological Chemistry* 270(14), pp. 7876–7881.

Rubin, G.M., Yandell, M.D., Wortman, J.R., et al. 2000. Comparative genomics of the eukaryotes. *Science* 287(5461), pp. 2204–2215.

Sandve, S.R., Rohlfs, R.V. and Hvidsten, T.R. 2018. Subfunctionalization versus neofunctionalization after whole-genome duplication. *Nature Genetics* 50(7), pp. 908–909.

Schmidt, D., Wilson, M.D., Ballester, B., et al. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* 328(5981), pp. 1036–1040.

Sémon, M. and Wolfe, K.H. 2007. Reciprocal gene loss between Tetraodon and zebrafish after whole genome duplication in their ancestor. *Trends in Genetics* 23(3), pp. 108–112.

Shin, H.Y., Willi, M., HyunYoo, K., et al. 2016. Hierarchy within the mammary STAT5-driven Wap super-enhancer. *Nature Genetics* 48(8), pp. 904–911.

Smith, J.J., Kuraku, S., Holt, C., et al. 2013. Sequencing of the sea lamprey (Petromyzon marinus) genome provides insights into vertebrate evolution. *Nature Genetics* 45(4), p. 415–21, 421e1.

Soltis, D.E., Buggs, R.J.A., Barbazuk, W.B., Schnable, P.S. and Soltis, P.S. 2009. On the origins of species: does evolution repeat itself in polyploid populations of independent origin? *Cold Spring Harbor Symposia on Quantitative Biology* 74, pp. 215–223.

Soltis, D.E., Visger, C.J. and Soltis, P.S. 2014. The polyploidy revolution then…and now: Stebbins revisited. *American Journal of Botany* 101(7), pp. 1057–1078.

Spitz, F. and Furlong, E.E.M. 2012. Transcription factors: from enhancer binding to developmental control. *Nature Reviews. Genetics* 13(9), pp. 613–626.

Stephens, S.G. 1951a. Possible significance of duplication in evolution. In: Advances in Genetics. Elsevier, pp. 247–265.

Stephens, S.G. 1951b. Possible significance of duplication in evolution. In: Advances in Genetics. Elsevier, pp. 247–265.

Van de Peer, Y., Maere, S. and Meyer, A. 2009. The evolutionary significance of ancient genome duplications. *Nature Reviews. Genetics* 10(10), pp. 725–732.

Vandepoele, K., De Vos, W., Taylor, J.S., Meyer, A. and Van de Peer, Y. 2004. Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proceedings of the National Academy of Sciences of the United States of America* 101(6), pp. 1638–1643.

Varadharajan, S., Sandve, S.R., Gillard, G.B., et al. 2018. The Grayling Genome Reveals Selection on Gene Expression Regulation after Whole-Genome Duplication. *Genome Biology and Evolution* 10(10), pp. 2785–2800.

Veitia, R.A. 2004. Gene dosage balance in cellular pathways: implications for dominance and gene duplicability. *Genetics* 168(1), pp. 569–574.

Vierstra, J., Rynes, E., Sandstrom, R., et al. 2014. Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. *Science* 346(6212), pp. 1007–1012.

Villar, D., Berthelot, C., Aldridge, S., et al. 2015. Enhancer evolution across 20 mammalian species. *Cell* 160(3), pp. 554–566.

Voldoire, E., Brunet, F., Naville, M., Volff, J.-N. and Galiana, D. 2017. Expansion by whole genome duplication and evolution of the sox gene family in teleost fish. *Plos One* 12(7), p. e0180936.

Walsh, J.B. 1995. How often do duplicated genes evolve new functions? *Genetics* 139(1), pp. 421–428.

Wang, D., Marsh, J.L. and Ayala, F.J. 1996. Evolutionary changes in the expression pattern of a developmentally essential gene in three Drosophila species. *Proceedings of the National Academy of Sciences of the United States of America* 93(14), pp. 7103–7107.

Werth, C.R. and Windham, M.D. 1991. A Model for Divergent, Allopatric Speciation of Polyploid Pteridophytes Resulting from Silencing of Duplicate-Gene Expression. *The American Naturalist* 137(4), pp. 515–526.

Whitehead, A. and Crawford, D.L. 2006. Neutral and adaptive variation in gene expression. *Proceedings of the National Academy of Sciences of the United States of America* 103(14), pp. 5425–5430.

Wolfe, K.H. 2001. Yesterday's polyploids and the mystery of diploidization. *Nature Reviews. Genetics* 2(5), pp. 333–341.

Wray, G.A. 2007. The evolutionary significance of cis-regulatory mutations. *Nature Reviews. Genetics* 8(3), pp. 206–216.

Xiao, S., Xie, D., Cao, X., et al. 2012. Comparative epigenomic annotation of regulatory DNA. *Cell* 149(6), pp. 1381–1392.

Yokoyama, S. and Yokoyama, R. 1989. Molecular evolution of human visual pigment genes. *Molecular Biology and Evolution* 6(2), pp. 186–197.

Young, R.S., Hayashizaki, Y., Andersson, R., et al. 2015. The frequent evolutionary birth and death of functional promoters in mouse and human. *Genome Research* 25(10), pp. 1546–1557.

Zhang, J. 2003. Evolution by gene duplication: an update. *Trends in Ecology & Evolution* 18(6), pp. 292–298.

Zhang, J., Rosenberg, H.F. and Nei, M. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proceedings of the National Academy of Sciences of the United States of America* 95(7), pp. 3708–3713.

Zhang, J., Zhang, Y. and Rosenberg, H.F. 2002. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nature Genetics* 30(4), pp. 411–415.

Paper I

Received: 7 July 2017 | Revised: 7 December 2017 | Accepted: 13 December 2017

DOI: 10.1111/mec.14533

WILEY MOLECULAR ECOLOGY

ORIGINAL ARTICLE

Life-stage-associated remodelling of lipid metabolism regulation in Atlantic salmon

Gareth Gillard[1]* | Thomas N. Harvey[2]* | Arne Gjuvsland[2] | Yang Jin[3] | Magny Thomassen[4] | Sigbjørn Lien[2] | Michael Leaver[5] | Jacob S. Torgersen[6] | Torgeir R. Hvidsten[1] | Jon Olav Vik[2] | Simen R. Sandve[2]

[1]Faculty of Chemistry, Biotechnology and Food Sciences, Norwegian University of Life Sciences, Ås, Norway

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

[3]Department of Biology, Norwegian University of Science and Technology, Trondheim, Norway

[4]Department of Animal and Aquacultural Sciences, Faculty of Biosciences, Norwegian University of Life Sciences, Ås, Norway

[5]Institute of Aquaculture, School of Natural Sciences, University of Stirling, Stirling, Scotland, UK

[6]AquaGen AS, Ås, Norway

Correspondence
Simen R. Sandve and Jon Olav Vik, Centre for Integrative Genetics (CIGENE), Department of Animal and Aquacultural Sciences, Faculty of Biosciences, Norwegian University of Life Sciences, Ås, Norway.
Emails: simen.sandve@nmbu.no; jonovik@gmail.com

Abstract

Atlantic salmon migrates from rivers to sea to feed, grow and develop gonads before returning to spawn in freshwater. The transition to marine habitats is associated with dramatic changes in the environment, including water salinity, exposure to pathogens and shift in dietary lipid availability. Many changes in physiology and metabolism occur across this life-stage transition, but little is known about the molecular nature of these changes. Here, we use a long-term feeding experiment to study transcriptional regulation of lipid metabolism in Atlantic salmon gut and liver in both fresh- and saltwater. We find that lipid metabolism becomes significantly less plastic to differences in dietary lipid composition when salmon transitions to saltwater and experiences increased dietary lipid availability. Expression of genes in liver relating to lipogenesis and lipid transport decreases overall and becomes less responsive to diet, while genes for lipid uptake in gut become more highly expressed. Finally, analyses of evolutionary consequences of the salmonid-specific whole-genome duplication on lipid metabolism reveal several pathways with significantly different ($p < .05$) duplicate retention or duplicate regulatory conservation. We also find a limited number of cases where the whole-genome duplication has resulted in an increased gene dosage. In conclusion, we find variable and pathway-specific effects of the salmonid genome duplication on lipid metabolism genes. A clear life-stage-associated shift in lipid metabolism regulation is evident, and we hypothesize this to be, at least partly, driven by nondietary factors such as the preparatory remodelling of gene regulation and physiology prior to sea migration.

KEYWORDS
adaptation, fish, life stage, metabolism, transcriptomics

1 | INTRODUCTION

Atlantic salmon lives a "double life." It starts its life in rivers, before transforming its physiology and behaviour and migrating to sea to grow and accumulate resources for reproduction. This shift in environment requires preparatory remodelling of physiology prior to sea migration (referred to as smoltification), which encompasses a suite of coordinately regulated processes involving hormonal changes and large-scale alteration of gene expression. The resulting adaptations to a marine environment include transformation of salt tolerance, coloration, behaviour, growth rate and metabolism (reviewed in Stefansson, Björnsson, Ebbesson, & McCormick, 2008).

*Shared first authors.

A key difference between freshwater and sea habitats is the dietary availability of essential long-chain polyunsaturated fatty acids. Salmon in rivers mostly eat invertebrates that are low in physiologically critical n-3 and n-6, 20 and 22 carbon long-chain polyunsaturated fatty acids (n-3LC-PUFA and n-6LC-PUFA), arachidonic acid (20:4n-6), eicosapentaenoic acid (20:5n-3) and docosahexaenoic (22:6n-3), while marine habitat food chains are high in available LC-PUFAs. Possibly, as an adaptation to this (Leaver, Bautista et al., 2008), salmon have evolved a high capacity for endogenous production of LC-PUFAs by elongation and desaturation of essential dietary 18 carbon precursor linoleic and linolenic acids (18:2n-6 and 18:3n-3; Figure 4) and the ability to increase or decrease this endogenous production as a response to the dietary availability (Kennedy et al., 2006; Leaver, Villeneuve et al., 2008; Morais et al., 2011; Ruyter, Røsjø, Måsøval, Einen, & Thomassen, 2000; Tocher, Bell, MacGlaughlin, McGhee, & Dick, 2001; Tocher et al., 2002; Zheng et al., 2005). During smoltification and after sea migration, Atlantic salmon have been shown to undergo transformation of lipid metabolism function, by decreasing lipid syntheses and increasing lipid breakdown (Sheridan, 1989). However, very little is known about the molecular nature of this life-stage-associated transformation physiological function.

The evolution of novel traits in salmonids, such as increased plasticity and the ability to migrate to sea, may have been facilitated by their ancestral whole-genome duplication (called Ss4R) some 80 Ma (Allendorf & Thorgaard, 1984; Lorgen et al., 2015; Macqueen & Johnston, 2014; Robertson et al., 2017). Gene duplication can give rise to new adaptive phenotypes in different ways: through evolution of novel functions or gene regulation, subdivision and/or specialization of function among duplicates, or via an adaptive increase in gene dosage. The Atlantic salmon genome contains ~10,000 pairs of Ss4R gene duplicates, of which ~50% have evolved some novel regulation (Lien et al., 2016; Robertson et al., 2017). Indeed, in the context of lipid metabolism, it has recently been shown that a Ss4R duplicate of elovl5, a key enzyme in LC-PUFA syntheses, has gained expression compared to its ancestral regulation with likely implications for the ability to synthesize LC-PUFAs (Carmona-Antoñanzas, Zheng, Tocher, & Leaver, 2016). This is believed to have facilitated evolution of novel traits, including flexible phenotypes necessary for an anadromous life history (Stefansson et al., 2008). However, no systematic genomewide study has yet been conducted to assess the importance of the Ss4R in evolution of salmon lipid metabolism.

In this study, we integrate comparative genomics with transcriptomic data from a feeding trial carried out across the freshwater to saltwater transition to build a functional annotation of lipid metabolism pathway genes in salmon. We use this annotation to elucidate (i) the nature of the transformation of lipid metabolism from freshwater to saltwater life stages and (ii) the impact of whole-genome duplication on evolution of the lipid gene repertoire and metabolic function. Our results indicate a striking shift in lipid metabolism after transition to sea water and show that lipid pathways differ with respect to selection pressure on gene duplicates from the salmonid whole-genome duplication.

2 | MATERIALS AND METHODS

2.1 | Orthogroup prediction

Protein sequences were obtained from seven teleost fish species: *Danio rerio* (zebrafish), *Gasterosteus aculeatus* (three-spined stickleback), *Oryzias latipes* (medaka), *Oncorhynchus mykiss* (rainbow trout), *Oncorhynchus kisutch* (coho salmon), *Salmo salar* (Atlantic salmon), *Thymallus thymallus* (grayling), *Esox lucius* (northern pike), and two mammalian outgroup species: *Homo sapiens* (human), *Mus musculus* (house mouse). Human, mouse, zebrafish, medaka and stickleback protein fasta data were obtained from ENSEMBL (release 83). Atlantic salmon (RefSeq assembly GCF_000233375.1, Annotation Release 100) and northern pike (RefSeq assembly GCF_000721915.2, Annotation Release 101) proteins were obtained from NCBI RefSeq. Rainbow trout proteins were obtained from an assembly and annotation of the genome (Berthelot et al., 2014). Grayling proteins were obtained from an assembly and annotation of the genome (Varadharajan et al., 2017). The coho salmon transcriptome (Kim, Leong, Koop, & Devlin, 2016) was obtained from NCBI (GDQG00000000.1). Where transcriptome data were used, protein sequences were translated using TRANSDECODER (v2.0.1, http://transdecoder.github.io/). Protein fasta files were filtered to retrieve only the longest protein isoform per gene. ORTHOFINDER (v0.2.8; Emms et al., 2015) assigned groups of orthologs based on protein sequence similarity. Proteins within an orthogroups were further aligned using MAFFT (v7.130; Katoh, Misawa, Kuma, & Miyata, 2002), and maximum-likelihood trees were estimated using FASTTREE (v2.1.8; Price et al., 2010).

2.2 | Annotation of salmon lipid metabolism genes

A list of zebrafish proteins obtained from 19 manually selected zebrafish KEGG pathways related to lipid metabolism (Appendix S1: Table S1) were used to search for Atlantic salmon orthologs. Orthogroups that contained a selected zebrafish protein were identified. Salmon proteins within those orthogroups were assigned as orthologs of the closest zebrafish protein based on the orthogroup tree distance. A lipid metabolism gene list was created including salmon orthologs to the selected zebrafish genes. Additional salmon genes related to lipid metabolism not included in KEGG pathways (e.g., regulators or transporters, SREBP, LXR, FABP) were manually searched for through NCBI and added to the list.

2.3 | Tissue expression

Atlantic salmon RNA-Seq samples from 15 different tissues (liver, gut, pyloric caeca, heart, kidney, muscle, gill, eye, skin, ovary, nose, testis, brain, head kidney and spleen) were obtained from NCBI SRA (PRJNA72713; Lien et al., 2016). Fastq files were adapter trimmed before alignment to the Atlantic salmon genome (RefSeq assembly GCF_000233375.1; Lien et al., 2016) using STAR (v2.5.2a; Dobin et al., 2013). HTSeq-count (v0.6.1p1; Anders, Pyl, & Huber, 2015) counted the sum of uniquely aligned reads in exon regions of each gene in the annotation (RefSeq Annotation Release 100). Gene FPKM values were calculated based on the gene count over the samples effective library size (see TMM method from EDGER (Robinson, McCarthy, & Smyth, 2010) user manual) and the mean gene transcript isoform length.

2.4 | Feed trial

Atlantic salmon fry were obtained from AquaGen Breeding Centre, Kyrksæterøra, Norway, and reared in the Norwegian Institute for Water Research (NIVA), Solbergstranda, Norway, in four partitioned 1,000-L tanks on vegetable oil (VO)- or fish oil (FO)-based diets continuously from first feeding (fry weight <0.2 g). Daily feed amount was calculated based on total biomass in each tank and decreased as the fish grew, from 3% at first feeding to 1.2% by the end of the trial. Fish were euthanized periodically throughout the experiment to maintain appropriate levels of dissolved oxygen. VO-based feeds contained a combination of linseed oil and palm oil at a ratio of 1.8:1, and FO-based feeds contained only North Atlantic fish oil. Percentage of protein in feed decreased with fish size from 56% at first feeding to 41% at the end of the trial. This corresponded to an increase in percentage of lipid from 16% at first feeding to 31% at the end of the trial. At the time of sampling, the proportion of lipid in the feed was 22% in freshwater and 31% in saltwater (Appendix S1: Table S2). Increasing lipid proportion in feed with fish size is standard practice in the aquaculture industry as this maintains optimal growing conditions by decreasing the digestible protein to digestible energy ratio (Storebakken, 2002). All feeds were formulated and produced by EWOS innovation (Supplementary File 3). Local groundwater was UV-sterilized for use in the freshwater life stage, and water from the Oslofjord taken from 60 metres below sea surface (~3%–3.5% salinity) was UV-sterilized for use in the saltwater life stage. Fish were raised under constant light and water temperature (~12°C) for 26 weeks. Then, 40 presmolt salmon (~50 g) from each control tank (~240 fish per control tank) were switched to the contrasting diet (VO to FO and vice versa) by physically moving them to the empty partition of the tank receiving the appropriate feed (Figure 8a). Five fish from each of the control tanks (2 VO tanks and 2 FO tanks) were sampled before switching feeds (D0), and then, fish from both control and feed switch conditions were similarly sampled 1, 2, 5, 9, 16 and 20 days after switching feeds (5 fish × 2 replicate tanks × 4 conditions = 40 fish per time point, Figure 8b). Two weeks after freshwater sampling (31 weeks after first feeding), smoltification was triggered by 5 weeks of winter-like conditions with decreased light (12 hr/day) and water temperature (~8°C), immediately followed by 5 weeks of spring-like conditions, returning to normal light (24 hours per day) and water temperature (~12°C). All salmon from the control groups (VO or FO) were then switched to saltwater and allowed to acclimate for 3 weeks. The feed switch was repeated in saltwater by transferring half (~40 fish) of the postsmolt salmon (~200 g) from each control tank to the contrasting feed condition. Again, preswitch control samples were taken (D0) followed by sampling 1, 2, 6, 9, 16 and 20 days postdiet switch (Figure 8b). For both freshwater and saltwater samplings, feeding was stopped in the mornings of each of the sampling days. All fish were euthanized by a blow to the head and samples of liver and midgut (gut section between pyloric caeca and hindgut) were flash-frozen in liquid nitrogen and stored under −80 °C. A subset of the samples taken were used for further RNA-Seq analysis (see Figure 8c for details).

2.5 | RNA sequencing

Total RNA was extracted from selected feed trial samples (see Figure 8c for details) using the RNeasy Plus Universal Kit (QIAGEN). Quality was determined on a 2100 Bioanalyzer using the RNA 6000 Nano Kit (Agilent). Concentration was determined using a Nanodrop 8000 spectrophotometer (Thermo Scientific). cDNA libraries were prepared using the TruSeq Stranded mRNA HT Sample Prep Kit (Illumina). Library mean length was determined by running on a 2100 Bioanalyzer using the DNA 1000 Kit (Agilent) and library concentration was determined with the Qbit BR Kit (Thermo Scientific). Single-end sequencing of sample libraries was completed on an Illumina HiSeq 2500 with 100-bp reads.

2.6 | Differential expression analysis between feed conditions and life stages

To analyse gene expression differences between feed conditions and life stages, samples from the feed trial were selected for RNA-Seq. Liver and gut tissue RNA were sequenced from fish fed each of the feeds (FO, VO) at day 0 of the diet switch, both before (freshwater) and after (saltwater) smoltification (see Figure 8c for the number of RNA-Seq replicates and sampling details). Fastq files were processed to produce gene count and FPKM data using the same protocol described under the Section 2.3. For the feed comparison, changes in gene expression were tested between FO and VO feed conditions for both freshwater and saltwater samples, and liver and gut tissues. For the life-stage comparison, changes in gene expression were tested between freshwater and saltwater stages for both FO and VO feed conditions, and liver and gut tissues. Using RNA-Seq gene count data, lowly expressed genes were filtered prior to testing, retaining genes with a minimum of one read count per million (CPM) in two or more samples. Differential expression analysis was carried out using a standard EDGER (Robinson et al., 2010) protocol. Effective library sizes were calculated using the EDGER TMM normalization procedure allowing effective comparison of expression data between different sample types (see EDGER user manual). An exact test between expression levels of a pair of conditions gave the log2-fold change, *p*-value and false discovery rate (FDR) for each gene. Genes with FDR <0.05 were considered differentially expressed genes (DEGs).

2.7 | Identification of Ss4R duplicates

To identify putative gene duplicates stemming from the Ss4R, we used the same approach as in Lien et al. (2016). All-vs-all protein blast was run with e-value cut-off of 1e−10 and pident (percentage of identical matches) ≥80 and blast hit coverage of ≥50% of protein length. Only the best protein hits between the 98 defined synteny blocks (see Lien et al., 2016) were considered as putative Ss4R duplicates. Blast result ranking was carried out using the product of pident times bitscore to avoid spurious "best blast matches" with low pident (<85), but high bitscore.

2.8 | Duplicate analysis

Genes from the lipid metabolism gene list were paired together with their putative Ss4R duplicates identified above. The retention of gene duplicates (i.e., whether both genes in a pair were retained, or just one) was compared between all identified duplicates in the salmon genome annotation and the lipid metabolism gene list. Pathway-level retention was explored by comparing the number of genes in each of the 19 selected KEGG pathways (Appendix S1: Table S1) in a duplicate pairing to that of the total list of lipid genes, to find pathways with significantly less or more duplicate retention (Fisher's exact test, *p*-value <.05). Regulatory conservation of lipid gene duplicates was explored by correlation of gene expression changes between duplicates over the course of the feed trial described above. RNA-Seq data were generated from liver samples of salmon from 38 sampling time points (19 in freshwater and 19 in saltwater). Fastq files were processed to produce gene count and FPKM data using the same protocol described under the Section 2.3. For each duplicate pair, mean FPKM values were retrieved for each time point and used to calculate a freshwater and saltwater correlation value. Duplicates with Pearson correlation ≥0.6 were considered correlated (*p*-value <.003 from 19 sample points). The number of duplicates with correlated expression profiles was counted for each pathway and compared to all lipid genes to find pathways with significantly less or more correlated duplicates (Fisher's exact test, *p*-value <.05). The effect of gene duplication on gene dosage was estimated by calculating a dosage ratio between the FPKM value of a salmon ortholog (sum of gene expression in duplicate pairs) over the FPKM value of the nonduplicated ortholog from northern pike. For salmon, the RNA-Seq data from the freshwater and saltwater FO feed trial was used (samples used in Section 2.6). For pike, RNA-Seq from livers of four individuals were aligned (see Section 2.3 for protocol) to their respective genomes (see genomes in Section 2.1). RSEM (v1.2.31; Li & Dewey, 2011) was used to generate FPKM values for genes so that nonuniquely mapped reads between salmon duplicate genes were not ignored but instead assigned proportionately to each gene to match the proportions of uniquely mapped reads between the genes. Gene dosage levels for duplicate pairs with correlated expression (see above), noncorrelated expression and single genes were compared for all lipid metabolism genes and for each pathway.

3 | RESULTS AND DISCUSSION

3.1 | Annotation of lipid metabolism genes

To identify genes involved in lipid metabolism in Atlantic salmon, we initially assembled groups of orthologous genes (orthogroups) using protein sequence similarity. We included proteins from four salmonid species sharing the Ss4R genome duplication, in addition to four nonsalmonid fish genomes and two model mammalian outgroup species (Figure 1a) to aid in distinguishing Ss4R copies from other gene duplicates. Next, we aligned orthogroup proteins and constructed

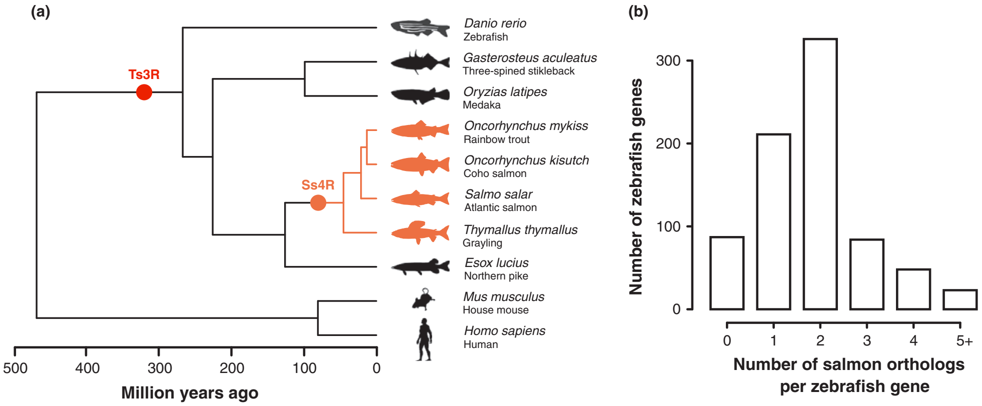


FIGURE 1 Ortholog annotation. (a) Species used to construct ortholog groups and their evolutionary distance. Points in the phylogenetic tree show the time of the teleost-specific (Ts3R) and salmonid-specific (Ss4R) whole-genome duplications. (b) The number of salmon orthologs found (1,421 genes in total) per zebrafish gene in 19 selected KEGG pathways involved in lipid metabolism [Colour figure can be viewed at wileyonlinelibrary.com]

maximum-likelihood gene trees. The majority (82%–98%) of proteins from each species were represented in 23,782 ortholog gene trees. The salmonid species had significantly higher number of proteins included in ortholog gene trees compared to nonsalmonid fish (Appendix S1: Figure S1), reflecting the salmonid-specific whole-genome duplication. We then used the evolutionary distances in gene trees to infer the most likely salmon sequence orthologs of zebrafish genes selected from 19 KEGG pathways involved in lipid metabolism (File S1). This resulted in the annotation of 1421 (File S2) salmon lipid metabolism genes, of which 326 (23%) showed a 2:1 ortholog ratio between salmon and zebrafish (Figure 1b). Only 87 (6%) of the zebrafish genes could not be assigned a salmon ortholog.

To validate our ortholog annotation pipeline used to identify lipid metabolism genes, we analysed the tissue specificity of these genes using gene expression data from 15 tissues (File S3) of Atlantic salmon (Lien et al., 2016). Genes in certain fatty acid metabolism-related pathways ("fatty acid metabolism," "PPAR signalling pathway," "fat digestion and absorption") had higher overall expression in tissues known to have high lipid metabolism activity (i.e., pyloric caeca, liver, heart and brain; Glatz, Luiken, & Bonen, 2010; Rimoldi, Benedito-Palos, Terova, & Pérez-Sánchez, 2016; Tocher, 2003; Figure 2). Examples include the following: (i) liver was the site of highest expression for all genes in the LC-PUFA biosynthesis pathway (the desaturases Δ6FAD and Δ5FAD, and the elongases elovl5, elovl2 and elovl4). (ii) Bile acids are essential for fat digestion in the gut, but are synthesized in liver. As expected, the rate-limiting step for bile syntheses, cytochrome P450 7A1 (CYP7A1), has the highest expression in the liver. (iii) Cholesterol, an essential component of cell membranes and precursor to bile acids, is known to be synthesized in all tissues, but primarily in liver, intestine and brain (Brown & Sharpe, 2016). This is reflected in our annotation by high expression of the key cholesterol biosynthesis genes 3-hydroxy-3-methylglutaryl-CoA reductase (HMGCR), isopentenyl-diphosphase Δisomerase (IDI1), squalene epoxidase (SM) and lanosterol synthase (LS) in these tissues. (iv) Several known regulators of lipid metabolism show high expression in liver, heart, brain and pyloric caeca, as expected, including liver X receptor (LXR), peroxisome proliferator-activated receptor-alpha (PPARα), sterol regulatory element-binding protein 1 (SREBP1), and sterol regulatory element-binding protein 2 (SREBP2). Taken together, the tissue distribution of lipid metabolism gene expression is in line with knowledge about vertebrate physiology in general, and supports the validity of our annotation of lipid metabolism genes in salmon. To make all data underlying our annotation easily available, and to facilitate further refinement through manual community curation, we have created an interactive Web server available online (goo.gl/8Ap89a).

3.2 | Life-stage-associated remodelling of lipid metabolism

We conducted a feeding trial to study how salmon adjusts its lipid metabolism to different levels of LC-PUFA in freshwater and saltwater (see Figure 8 for experimental details). Groups of salmon were fed contrasting diets from hatching until after transition to sea water. One feed was based on vegetable oil (VO) and hence low in LC-PUFA, similar to river ecosystem diets, whereas the other was based on fish oil (FO) and high in LC-PUFA as expected in a marine-type diet (see Appendix S1: Tables S2 and S3 for details on feed composition). VO-based diets are also low in cholesterol (Ciftci, Przybylski, & Rudzińska, 2012; Verleyen et al., 2002). The proportion of fat in feed also increased between FW and SW (Appendix S1: Table S2), as is standard practice in the aquaculture industry to maintain optimal growth conditions (Storebakken, 2002). Moreover, total lipid availability is also expected to increase between natural

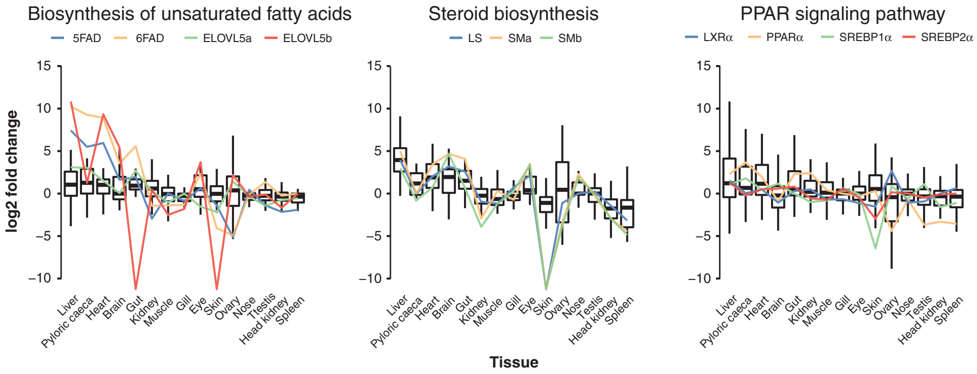


FIGURE 2 Tissue expression profiles of salmon genes in lipid metabolism pathways. Tissue expression profiles of our annotated lipid metabolism genes were consistent with expectations. Gene expression levels are shown as the log2-fold change difference between the FPKM value of each tissue and the median FPKM across all tissues. Expression profiles for selected genes in each pathway are shown (see Figures S2 and S3 for all pathways and gene details) [Colour figure can be viewed at wileyonlinelibrary.com]

riverine and marine ecosystem diets. The contrasting levels of EPA/DHA between FO and VO diets remained constant across life stages. In total, 32 and 23 fish were sampled for RNA-Seq of liver and gut, respectively, including up to eight biological replicates from each diet and life stage (freshwater and saltwater, see Figure 8c for details). Fish in the different dietary groups were given FO and VO feed from first feeding (<0.2 g body weight) until sampling.

In general, global gene expression levels were more affected by dietary composition in liver than in gut (which was largely unresponsive), and the effect was more pronounced in freshwater than in saltwater (Figure 3a). VO diets, compared to FO diets, increased lipid metabolism-related gene expression in liver. In freshwater, 66 genes were differentially expressed with 57 (86%) of these upregulated, while in saltwater, 31 genes were differentially expressed with 23 (74%) of these upregulated (Figure 3b). The increased activity of liver lipid metabolism under VO diets confirms the well-known ability of salmon to regulate endogenous synthesis of LC-PUFA and cholesterol in response to VO diets (Kortner, Björkhem, Krasnov, Timmerhaus, & Krogdahl, 2014; Leaver, Villeneuve et al., 2008; Zheng et al., 2005).

Fish sampled in freshwater and saltwater shared a relatively small number of differentially expressed genes (DEGs) for each pathway (Appendix S1: Table S4). We found that most pathways had more DEGs in freshwater ("fatty acid biosynthesis," "steroid biosynthesis" and its precursor "terpenoid backbone biosynthesis"), whereas few had more DEGs in saltwater ("fat digestion and absorption" and "steroid hormone biosynthesis"; Figure 3c). Of 87 lipid metabolism DEGs in the dietary contrast, 56 (64%) were freshwater-specific, 21 (24%) were saltwater-specific, and 10 (11%) shared dietary response. For example, only two genes in the FA and LC-PUFA biosynthesis pathways (Δ6FADa and Δ5FAD) shared response to diet in freshwater and saltwater (Figure 4). Similarly, in the pathways responsible for cholesterol biosynthesis, there were more DEGs between diets in FW (21 DEGs in FW, 4 shared and no SW-specific; Figure 5). The few genes that showed diet effects specific to saltwater included bile salt-activated lipase, responsible for the hydrolysis of free fatty acids from TAG obtained from the diet (Tocher, 2003). Two of these genes, carboxyl ester lipase, tandem duplicate 2a (CEL2a) and b (CEL2b), are highly upregulated in saltwater in response to VO diet. Taken together, our results show higher metabolic plasticity in parr-stage salmon, suggesting a life-stage-associated remodelling of lipid metabolism in liver. This corroborates the idea of a postsmoltification phenotype adapted to an environment with a surplus of n-3LC-PUFA.

To further investigate the life-stage-associated changes in lipid metabolism, we tested for differential expression between salmon in freshwater and saltwater fed diets with identical n-3LC-PUFA profiles (Figure 6). Liver and gut showed contrasting effects of saltwater on lipid gene expression with extensive downregulation in liver and upregulation in gut (Figure 6b). The number of DEGs in each tissue was similar for the environment comparison (Figure 6a), unlike for the diet comparison (Figure 3).

Further examination of key lipid metabolism genes revealed that after life-stage transition, the systemwide lipid metabolism

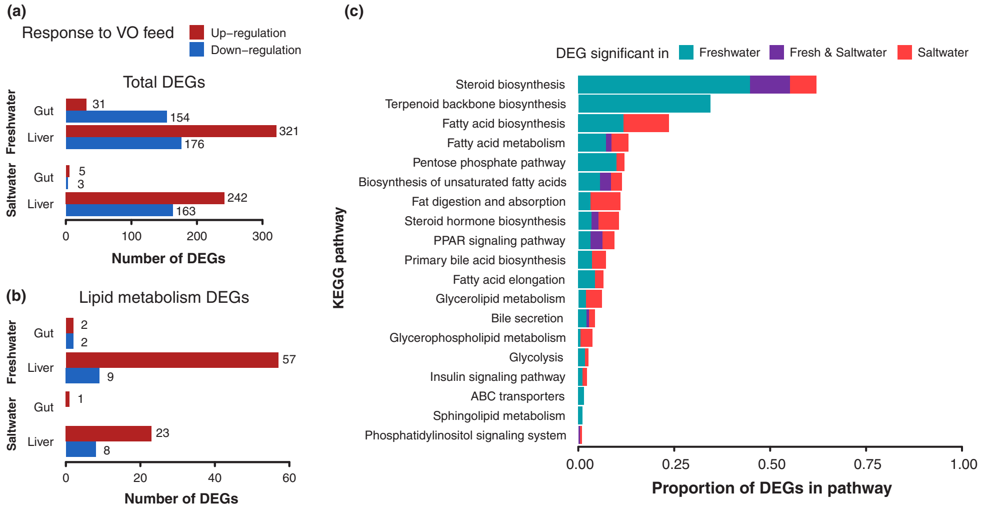


FIGURE 3 Gene regulation in response to feed type. (a) Total number of significant (FDR <0.05) differentially expressed genes (DEGs) between fish oil (FO)- and vegetable oil (VO)-fed salmon in the liver and gut tissues of freshwater and saltwater stage Atlantic salmon (see Files S4 (liver) and S5 (gut) for underlying data). (b) As above, but for lipid-associated genes only. (c) Proportions of genes in each KEGG pathway that had significantly different liver expression between the two feed types only in freshwater, only in saltwater or in both stages [Colour figure can be viewed at wileyonlinelibrary.com]

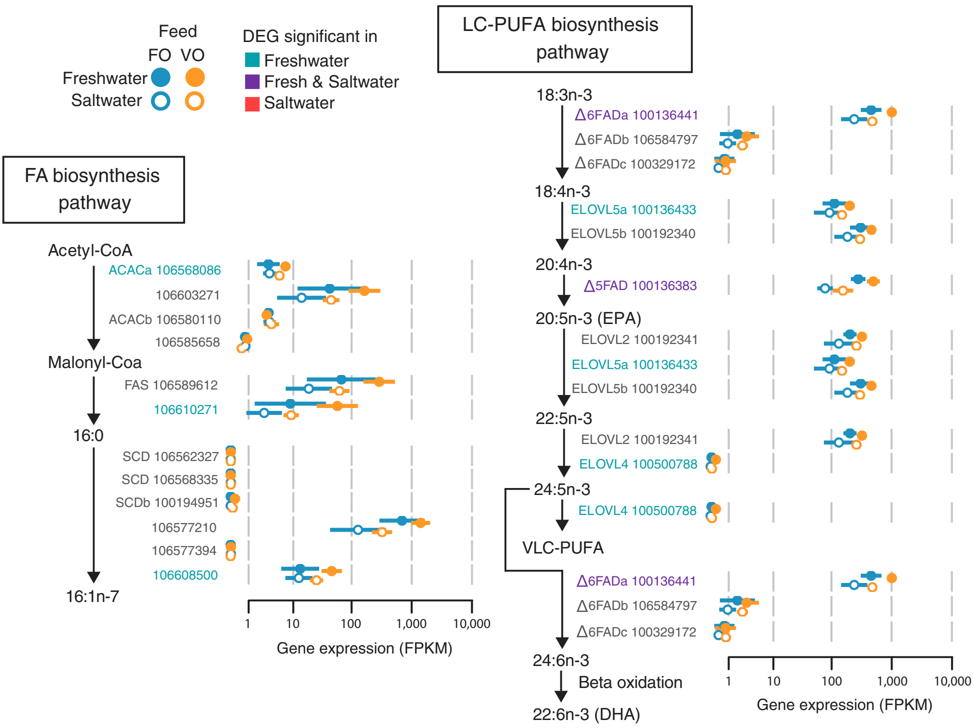


FIGURE 4 Diet and life-stage effects on FA and LC-PUFA biosynthesis in salmon liver. Core fatty acid (FA) biosynthesis and biosynthesis of unsaturated fatty acids pathways with Atlantic salmon genes annotated to each catalytic step (enzyme names followed by NCBI gene numbers). Gene expression levels are shown as mean (point) and standard deviation (line) of expression in eight samples (measured in log (FPKM + 1)) from each diet (FO, VO feeds) and life-stage (freshwater, saltwater) combination. Genes significantly (FDR < 0.05) differentially expressed (DEG) between diets in a life stage are highlighted [Colour figure can be viewed at wileyonlinelibrary.com]

remodelling represented a concerted shift in the metabolic role of liver and gut. After the salmon entered the marine stage, lipogenic gene expression in the liver was significantly decreased, as evident by the markedly lower expression (2.2- to 3.3-fold) of the master regulator of lipid metabolism SREBP1, a fivefold decrease in expression of fatty acid synthase and a two- to threefold decrease in rate-limiting enzymes in LC-PUFA synthesis (i.e., Δ5FAD, Δ6FADa; Figure 4). Liver and gut gene expression also indicated increased catabolic activity in saltwater, with upregulation of the carnitine palmitoyltransferase 1 and 2 genes, responsible for uptake of fatty acids into mitochondria for β-oxidation (Lehner & Quiroga, 2016). Finally, expression of lipid transport genes shifted from liver to gut with the transition to seawater (apolipoproteins, pathway "Fat digestion and absorption" in Figure 6). Four apolipoproteins (of 11 annotated) were differentially regulated in liver between different life stages, with a 2.4- to 5-fold decrease in saltwater compared to freshwater. In stark contrast, nine of the diet-regulated apolipoproteins in gut increased their expression in saltwater between 1.8- and 9.7-fold. The results point to an adaptive shift in lipid metabolism, with increased ability to take up lipids in the gut after Atlantic salmon migrates to sea where lipid availability is higher. Remodelling of lipid metabolism across life stages is likely the result of a combination of factors, including the direct regulatory effect of dietary fat itself, effect of salinity and smoltification-induced physiological changes influencing gene regulation. Although the relative importance of these factors is undetermined in our study, the fact that DEGs in the VO versus FO feed contrast were mostly life-stage-specific (Figure 3) supports that factors other than the diet itself contribute significantly to the freshwater and seawater metabolic phenotypes.

Interestingly, diet had a strong influence on the number and direction of gene expression changes between freshwater and saltwater (Figure 6). In gut, about twice as many DEGs (with respect to the fresh- to saltwater transition) were observed in salmon when fed

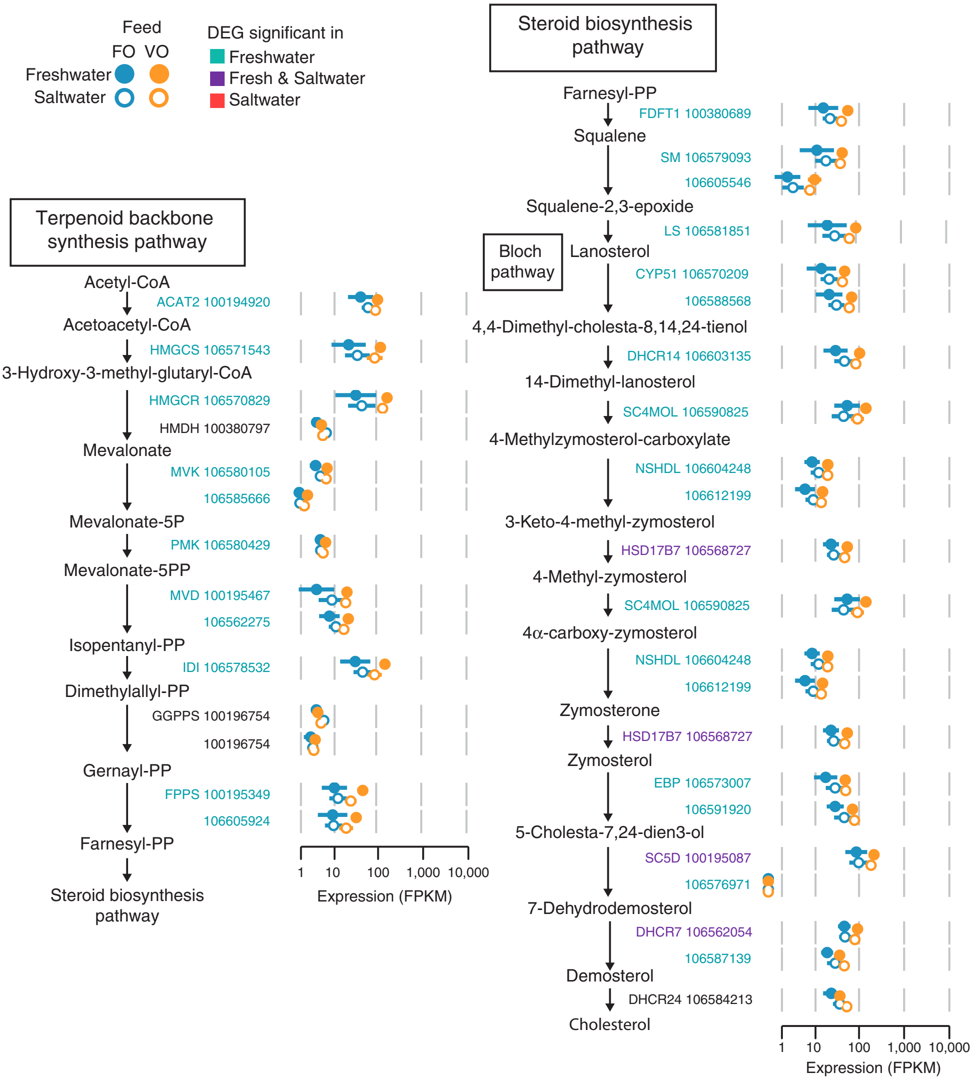


FIGURE 5 Diet and life-stage effects on cholesterol biosynthesis in salmon liver. Terpenoid backbone synthesis and steroid biosynthesis pathways with Atlantic salmon genes annotated to each catalytic step (enzyme names followed by NCBI gene numbers). Gene expression levels are shown as mean (point) and standard deviation (line) of expression in eight samples (measured in log(FPKM + 1)) from each diet (FO, VO feeds) and life-stage (freshwater, saltwater) combination. Genes significantly (FDR < 0.05) differentially expressed (DEG) between diets in a life stage are highlighted [Colour figure can be viewed at wileyonlinelibrary.com]

FO diet than VO diet (Figure 6a). In liver, the diet effect was less pronounced, with the FO group containing 46% more DEGs than the VO group (Figure 6a). This diet effect pattern was reflected in the lipid metabolism genes with 89% and 16% more DEGs in the FO group for gut and liver, respectively (Figure 6b). As this diet and life-stage interaction is a genomewide trend, and more pronounced in

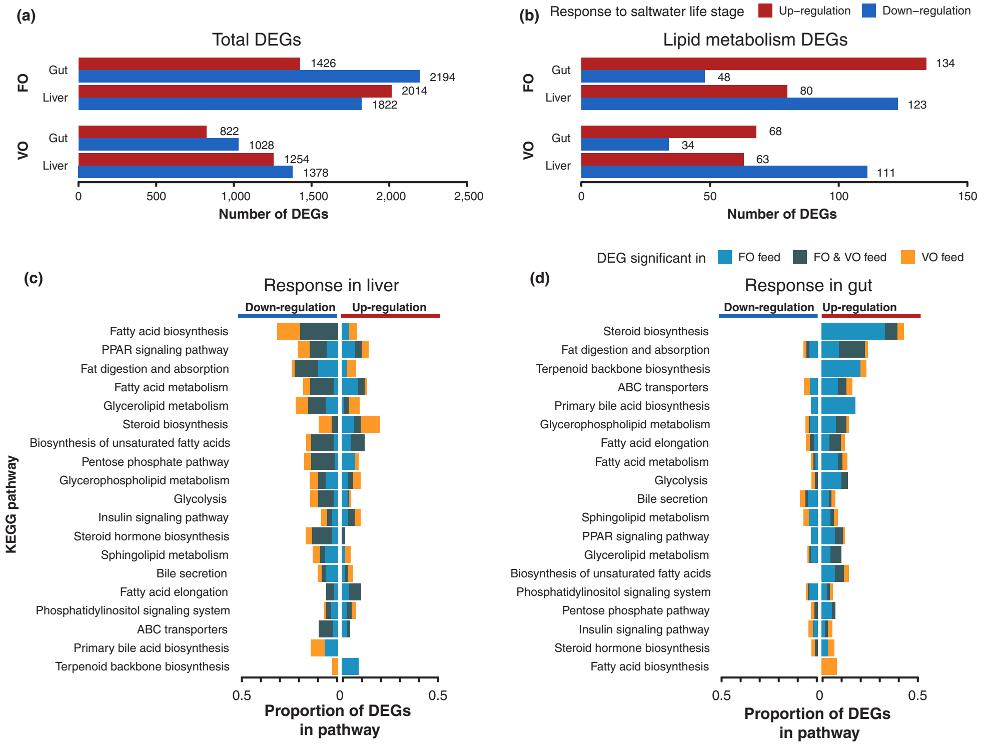


FIGURE 6 Gene regulation in response to life stage. (a) Total number of significant (FDR < 0.05) differentially expressed genes (DEGs) between freshwater and saltwater life stages in the liver and gut tissues of Atlantic salmon fed fish oil (FO) or vegetable oil (VO) diets (see Files S6 and S7 for underlying data). (b) As above, but for lipid metabolism DEGs. (c) Proportion of genes in each KEGG pathway that are DEGs in liver and (d) gut, coloured by DEG significance in only FO, only VO or both diets, and separated into up- or downregulation in saltwater samples [Colour figure can be viewed at wileyonlinelibrary.com]

gut tissue than in liver, this pattern could be related to differences in osmoregulation and adaptation to saltwater. Two studies have suggested that Atlantic salmon raised on VO-based feeds more closely resembling riverine diets adapt to saltwater sooner and better than salmon raised on FO-based diets (Bell et al., 1997; Tocher et al., 2000). Conversely, there has been evidence that VO-based diets can reduce markers for stress response upon saltwater challenge, resulting in reduced osmoregulatory capacity (Oxley et al., 2010). Regardless of the effect, it is clear that diet can modulate the smoltification process and could explain the discrepancy between diets in number of life-stage-related DEGs. Another possibility is that the different levels of fatty acids in the diets, for example DHA, affect DNA methylation and thus trigger genomewide divergence in gene regulation (Kulkarni et al., 2011).

Our results clearly demonstrate very different baseline lipid metabolic functions in pre- and postsmolt salmon, as well as life-stage-associated changes in the plasticity of lipid metabolism, for example the ability to regulate endogenous LC-PUFA synthesis as a response to changes in diet (i.e., fatty acid composition). As opportunistic carnivores, salmon tend to eat whatever the local environment provides. Thus, in freshwater, insects and amphipods provide variable, mostly low amounts of essential LC-PUFA and total fat (Jonsson & Jonsson, 2011; Sushchik, Gladyshev, Moskvichova, Makhutova, & Kalachova, 2003), favouring a metabolic function that can efficiently regulate endogenous lipid synthesis based on dietary availability (Carmona-Antonanzas, Tocher, Martinez-Rubio, & Leaver, 2014). Conversely, in marine environments, amphipods and smaller fish provide a higher, more stable source of n-3LC-PUFA and total fat (Baeza-Rojano, Hachero-Cruzado, & Guerra-García, 2014; Jonsson & Jonsson, 2011), promoting a metabolic function that allocates less energy to endogenous synthesis of essential lipids.

3.3 | Selection on gene duplicates after whole-genome duplication

Carmona-Antonanzas et al. (2014), Carmona-Antoñanzas et al. (2016) proposed that the salmonid whole-genome duplication may have adaptively increased the potential for endogenous lipid synthesis. We pursued this hypothesis by searching for distinct signatures of selection pressure on lipid metabolism genes in salmon. Specifically, we compared pathways in terms of their tendency to retain both duplicates of gene pairs, in terms of whether duplicates showed similar regulation (expression patterns across diets and environments) and in terms of total gene dosage (for the one or two genes retained of a pair) in salmon compared to pike, its closest unduplicated sister lineage.

To assess the level of Ss4R duplicate retention, we first defined 10,752 Ss4R duplicate pairs (21,504 genes) in the NCBI RefSeq annotation using the same approach as Lien et al. (2016). Of the 1,421 annotated lipid metabolism genes, 867 (61%) were retained as duplicated genes after Ss4R (Figure 7a; in contrast to 47% of the 45,127 salmon genes assigned to ortholog groups). Moreover, our results showed large variation in the proportion of retained duplicates in each lipid metabolism pathway (Figure 7), with the most extreme case being “fat digestion and absorption” with 80% retained duplicates and “steroid hormone biosynthesis” with only 27% retained Ss4R duplicates.

The regulatory conservation of the duplicates was then estimated by calculating co-expression correlation between Ss4R duplicates from RNA-Seq data representing a time-course of dynamic changes in gene expression and lipid metabolism function in liver. Fish in the same feeding trial were switched from VO to FO feed and vice versa, in both fresh- and saltwater conditions (see Figure 8 for details). In total, 38 sampling time points (20 in freshwater and 18 in saltwater) from the feed switch experiment were used. Pathway-level analyses showed that regulatory conservation was not associated with duplicate retention (Figure 7). For example, the “*biosynthesis of unsaturated fatty acids”* pathway had significantly fewer duplicates retained than expected by chance (*p*-value <.0234), but a significant overrepresentation of duplicate pairs that display highly similar regulation (*p*-value <.0142 and <.0361 in freshwater and saltwater, respectively). Interestingly, the “insulin signalling pathway” also showed higher-than-expected duplicate coregulation. This pathway has been shown to be

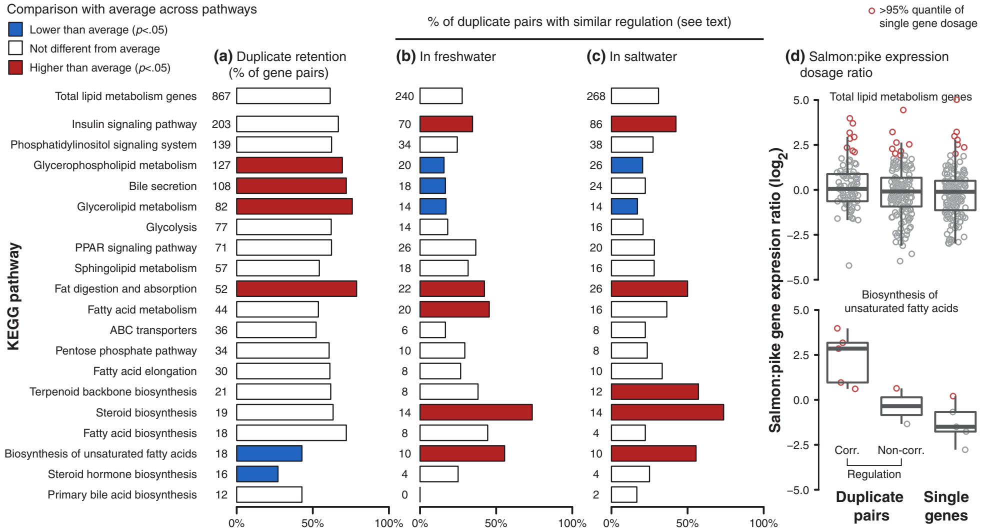
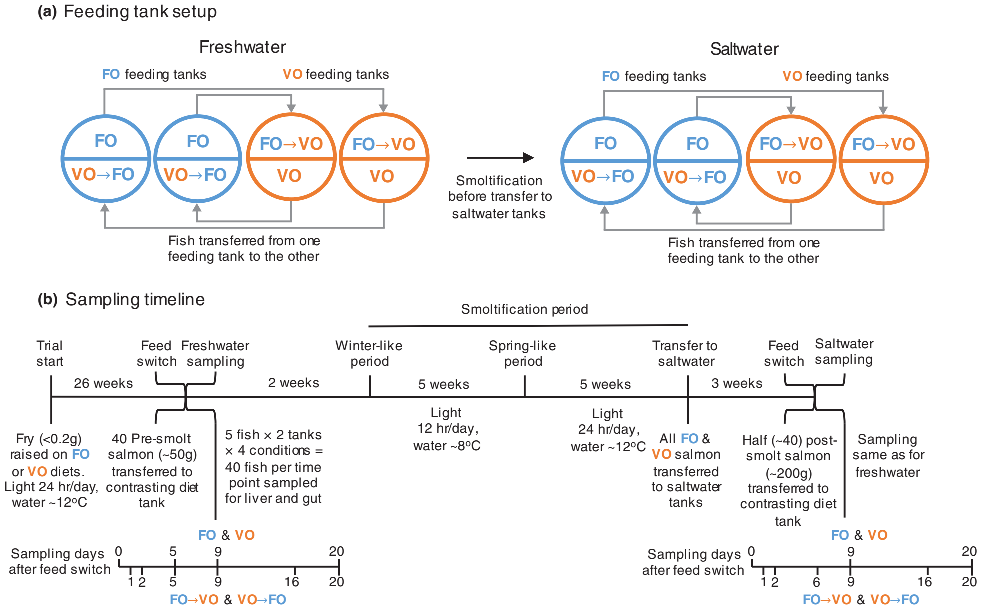


FIGURE 7 Gene duplication in lipid metabolism pathways. For the total list of lipid metabolism genes in Atlantic salmon, and sets of genes belonging to different KEGG pathways: (a) number and percentage of genes with a duplicate homolog from the Ss4R duplication. (b) Number and percentage of duplicate genes with correlated liver expression response to feed in freshwater and (c) saltwater (correlation ≥.6, *p*-value <3.306e−3, using 19 time points from feed trial for each water condition). Fisher's exact test was used to detect pathways with significant enrichment compared to all gene (*p*-value <.05) (d) Log2 gene dosage ratios (salmon/pike) in liver from fish in freshwater, where the ratio is computed between expression in the salmon duplicates (FPKM, sum of the two duplicates) and the expression of the corresponding pike ortholog. Ratios were computed for all lipid metabolism genes and genes in the pathway “biosynthesis of unsaturated fatty acids.” For comparison, ratios were also computed for genes without retained duplicates, that is with a 1:1 orthology between salmon and pike. Duplicates were grouped into correlated (corr.) or noncorrelated (noncorr.) based on saltwater correlation result in (c). Dosage ratios (points) greater than the 95% quantile of single gene dosages are marked in red [Colour figure can be viewed at wileyonlinelibrary.com]

important in regulating uptake and transport of FAs in adipose tissue, liver and muscle of Atlantic salmon (Sánchez-Gurmaches et al., 2011). Other pathways showing signatures of increased duplicate coregulation were "terpenoid backbone biosynthesis," "steroid biosynthesis," "fat digestion and absorption" and "fatty acid metabolism" (Figure 7b,c). Overall, the distinct differences in duplicate retention and conservation of regulatory mechanisms across the lipid metabolism pathways suggest differences in selective pressures shaping duplicate evolution following Ss4R. Moreover, the pathways with highly conserved duplicate coregulation were also those that were most responsive to dietary differences in fatty acid composition (Figure 3).


(a) Feeding tank setup
Freshwater
FO feeding tanks
VO feeding tanks
FO
VO→FO
FO
VO→FO
FO→VO
VO
FO→VO
VO
Fish transferred from one feeding tank to the other
Smoltification before transfer to saltwater tanks
Saltwater
FO feeding tanks
VO feeding tanks
FO
VO→FO
FO
VO→FO
FO→VO
VO
FO→VO
VO
Fish transferred from one feeding tank to the other
(b) Sampling timeline
Smoltification period
Trial start
Feed switch
Freshwater sampling
Winter-like period
Spring-like period
Transfer to saltwater
Feed switch
Saltwater sampling
26 weeks
2 weeks
5 weeks
5 weeks
3 weeks
Fry (<0.2g) raised on FO or VO diets. Light 24 hr/day, water ~12°C
40 Pre-smolt salmon (~50g) transferred to contrasting diet tank
5 fish × 2 tanks × 4 conditions = 40 fish per time point sampled for liver and gut
Light 12 hr/day, water ~8°C
Light 24 hr/day, water ~12°C
All FO & VO salmon transferred to saltwater tanks
Half (~40) post-smolt salmon (~200g) transferred to contrasting diet tank
Sampling same as for freshwater
FO & VO
Sampling days after feed switch
0 5 9 20
1 2 5 9 16 20
FO→VO & VO→FO
FO & VO
Sampling days after feed switch
0 9 20
1 2 6 9 16 20
FO→VO & VO→FO

(c) Samples sequenced for total RNA

Feed	Tissue	Freshwater sampling days 0	1	2	5	9	16	20	Saltwater sampling days 0	1	2	6	9	16	20
FO	Liver	8			8	8		8	8				8		8
	Gut	4				4		4	8				8		8
VO	Liver	8			8	8		8	8				8		8
	Gut	4				4		4	7				8		8
VO→FO	Liver		4	4	4	4	4	4		4	4	4	4	4	4
	Gut		4	4	4	4	4	4		4	4	4	4	4	4
FO→VO	Liver		4	4	4	4	4	4		4	4	4	4	4	4
	Gut		4	4	4	4	4	4		4	4	4	4	4	4

FIGURE 8 Overview of feed trial experiment. (a) Atlantic salmon fry were reared in four feeding tanks containing freshwater; two continuously fed fish oil (FO) and two vegetable oil (VO). A feed switch involved the transfer of fish from one tank to an empty partition of another tank fed the opposite diet. After smoltification, fish from FO and VO tanks were transferred to four new feeding tanks containing saltwater and the feed switch was repeated. (b) Timeline of feed trial showing fish sampling and smoltification periods. Fish were sampled before (D0) and up to 20 days after the fresh- or saltwater feed switch. (c) Total RNA was sequenced from select fish tissue samples. The number of RNA-Seq replicates is shown for each tissue, condition and time point [Colour figure can be viewed at wileyonlinelibrary.com]

Finally, to link duplicate retention and coregulation to signals of increased gene dosage following Ss4R, we used RNA-Seq data from the northern pike (*Esox lucius*), a species that belongs to the unduplicated sister lineage (see Section 2 for details). For each duplicate pair, we computed the ratio between the sum of Ss4R duplicate expression and its nonduplicated ortholog in pike and compared these ratios to those observed for salmon genes that had not retained two Ss4R duplicates. In total, 69 duplicate pairs from 18 different lipid metabolism-related pathways displayed a combined dosage increase relative to single-copy genes, of which 26 had highly conserved regulation (i.e., correlated expression; File S8). We saw no systematic effect of gene dosage when comparing the total gene expression of duplicate pairs with that of single-copy genes, nor did coregulation of duplicates associate with increased gene dosage (Figure 7d). This pattern was also true for most individual lipid pathways (Appendix S1: Figures S4, S5), except for "*biosynthesis of unsaturated fatty acids,*" "*fatty acid metabolism*" and "*fatty acid elongation.*" These three pathways showed a link between coregulation of duplicated genes and higher total gene dosage (Figure 7d; Appendix S1: Figures S4, S5). Underlying this link were three genes with coregulated dosage effects shared between all three pathways; trifunctional enzyme alpha subunit b (hadhab), elovl6 and the previously identified elovl5 (Carmona-Antonanzas et al., 2014, 2016). Only elovl5 is known to be directly involved in core PUFA biosynthesis. Hadhab is involved in mitochondrial β-oxidation/elongation, and elovl6 is involved in elongation of saturated and monounsaturated fatty acids (Bond, Miyazaki, O'Neill, Ding, & Ntambi, 2016). Although we do not see a general trend of increased gene dosage effects on lipid metabolism genes after whole-genome duplication, it is likely that an increased dosage of elovl5 and the 68 other duplicate pairs has affected the function of lipid metabolism in salmon.

4 | CONCLUSION

Atlantic salmon needs great plasticity of physiology and behaviour to adapt for migration between freshwater and sea. By analysing transcriptomic changes through the transition from fresh- to saltwater and the associated increase in dietary lipids, we identified an overall remodelling of lipid metabolism, with liver reflecting higher lipid metabolic plasticity and higher capacity of endogenous synthesis of LC-PUFAs in freshwater, while gut lipid uptake genes become more active in saltwater. These results indicate adaptive optimization of the Atlantic salmon lipid metabolism to account for life-stage-specific dietary availability. Moreover, we found signatures of pathway-specific selection pressure on gene duplicates, including a gene dosage increase in three genes involved in fatty acid metabolism. This illustrates possible adaptive consequences of the salmonid whole-genome duplication for the evolution of lipid metabolism. Future studies should attempt to decipher how the life-stage-related metabolic reprogramming is controlled (e.g., through epigenetic mechanisms). Understanding this will have important implications for understanding evolution of genome regulatory processes in anadromous salmonids and potentially have economically important implications for Atlantic salmon aquaculture.

ACKNOWLEDGEMENTS

This work was financed by the projects DigiSal NFR 248792 and GenoSysFat NFR 244164. We also thank anonymous reviewers for comments on the earlier version of the manuscript as well as Dominic Nanton and Cargill Innovation Center Dirdal for producing the fish trial feeds.

DATA ACCESSIBILITY

- Supplementary files have been deposited to datadryad.org under the accession: https://doi.org/10.5061/dryad.j4h65.
- All gene expression results can be accessed through the interactive shiny Web server: https://goo.gl/8Ap89a.
- Lipid metabolism gene annotation can be accessed from https://goo.gl/VVUVWr.
- Raw RNA-Seq data have been deposited into European Nucleotide Archive (ENA) under the project Accession no. PRJEB24480.

AUTHOR CONTRIBUTIONS

S.R.S., J.O.V., A.G., and J.S.T. conceived of the study and designed the feeding trial experiment. S.R.S., J.O.V., and A.G. carried out orthogroup prediction and lipid metabolism annotation. T.N.H., Y.J., J.S.T., and J.O.V. carried out the feeding trial. T.N.H. and Y.J. prepared the samples for RNA sequencing. M.T., S.L., and M.L. provided input on the experimental design and helped interpreting the results from the transcriptomic and gene duplicate analysis. G.G. and T.R.H. performed the bioinformatic analyses. T.N.H., G.G., S.R.S., and T.R.H. wrote the manuscript. All authors reviewed the final manuscript draft.

ORCID

Gareth Gillard http://orcid.org/0000-0001-9533-3227
Thomas N. Harvey http://orcid.org/0000-0003-4882-2188
Arne Gjuvsland http://orcid.org/0000-0002-4391-3411
Yang Jin http://orcid.org/0000-0001-5597-8397
Magny Thomassen http://orcid.org/0000-0003-1415-4652
Michael Leaver http://orcid.org/0000-0002-3155-0844
Jacob S. Torgersen http://orcid.org/0000-0002-6758-2979
Torgeir R. Hvidsten http://orcid.org/0000-0001-6097-2539
Jon Olav Vik http://orcid.org/0000-0002-7778-4515
Simen R. Sandve http://orcid.org/0000-0003-4989-5311

REFERENCES

Allendorf, F. W., & Thorgaard, G. H. (1984). Tetraploidy and the evolution of Salmonid Fishes. In B. J. Turner (Ed.), *Evolutionary genetics of* *fishes* (pp. 1–53). Boston, MA: Springer US. https://doi.org/10.1007/978-1-4684-4652-4_1

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq–A Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), 166–169. https://doi.org/10.1093/bioinformatics/btu638

Baeza-Rojano, E., Hachero-Cruzado, I., & Guerra-García, J. M. (2014). Nutritional analysis of freshwater and marine amphipods from the Strait of Gibraltar and potential aquaculture applications. *Journal of Sea Research*, *85*, 29–36. https://doi.org/10.1016/J.SEARES.2013.09.007

Bell, J. G., Tocher, D. R., Farndale, B. M., Cox, D. I., McKinney, R. W., & Sargent, J. R. (1997). The effect of dietary lipid on polyunsaturated fatty acid metabolism in Atlantic salmon (*Salmo salar*) undergoing parr-smolt transformation. *Lipids*, *32*(5), 515–525. https://doi.org/10.1007/s11745-997-0066-4

Berthelot, C., Brunet, F., Chalopin, D., Juanchich, A., Bernard, M., Noël, B., . . . Guiguen, Y. (2014). The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nature Communications*, *5*, 3657. https://doi.org/10.1038/ncomms4657

Bond, L. M., Miyazaki, M., O'Neill, L. M., Ding, F., & Ntambi, J. M. (2016). Fatty acid desaturation and elongation in mammals. *Biochemistry of lipids, lipoproteins and membranes* (pp. 185–208). Amsterdam, The Netherlands: Elsevier. https://doi.org/10.1016/b978-0-444-63438-2.00006-7

Brown, A. J., & Sharpe, L. J. (2016). Cholesterol synthesis. In N. D. Ridgway & R. S. McLeod (Eds.), *Biochemistry of lipids, lipoproteins and membranes* (pp. 327–358). Amsterdam, The Netherlands: Elsevier. https://doi.org/10.1016/b978-0-444-63438-2.00011-0

Carmona-Antonanzas, G., Tocher, D. R., Martinez-Rubio, L., & Leaver, M. J. (2014). Conservation of lipid metabolic gene transcriptional regulatory networks in fish and mammals. *Gene*, *534*(1), 1–9. https://doi.org/10.1016/j.gene.2013.10.040

Carmona-Antoñanzas, G., Zheng, X., Tocher, D. R., & Leaver, M. J. (2016). Regulatory divergence of homeologous Atlantic salmon elovl5 genes following the salmonid-specific whole-genome duplication. *Gene*, *591*(1), 34–42. https://doi.org/10.1016/j.gene.2016.06.056

Ciftci, O. N., Przybylski, R., & Rudzińska, M. (2012). Lipid components of flax, perilla, and chia seeds. *European Journal of Lipid Science and Technology*, *114*(7), 794–800. https://doi.org/10.1002/ejlt.201100207

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., . . . Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Emms, D. M., Kelly, S., Alexeyenko, A., Tamas, I., Liu, G., Sonnhammer, E., . . . Kellis, M. (2015). ORTHOFINDER: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, *16*(1), 157. https://doi.org/10.1186/s13059-015-0721-2

Glatz, J. A N. F. C., Luiken, J. J. F. P., & Bonen, A. (2010). Membrane fatty acid transporters as regulators of lipid metabolism: Implications for metabolic disease. *Physiological Reviews*, *90*, 367–417. https://doi.org/10.1152/physrev.00003.2009

Jonsson, B., & Jonsson, N. (2011). *Ecology of Atlantic Salmon and Brown Trout—Habitat as a template for life histories*. *Fish and Fisheries Series* (Vol. 33). Dordrecht: Springer.

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066. https://doi.org/10.1093/NAR/GKF436

Kennedy, S. R., Leaver, M. J., Campbell, P. J., Zheng, X., Dick, J. R., & Tocher, D. R. (2006). Influence of dietary oil content and conjugated linoleic acid (CLA) on lipid metabolism enzyme activities and gene expression in tissues of Atlantic salmon (*Salmo salar* L.). *Lipids*, *41*(5), 423–436. https://doi.org/10.1007/s11745-006-5116-4

Kim, J.-H., Leong, J. S., Koop, B. F., & Devlin, R. H. (2016). Multi-tissue transcriptome profiles for coho salmon (*Oncorhynchus kisutch*), a species undergoing rediploidization following whole-genome duplication. *Marine Genomics*, *25*, 33–37. https://doi.org/10.1016/j.margen.2015.11.008

Kortner, T. M., Björkhem, I., Krasnov, A., Timmerhaus, G., & Krogdahl, Å. (2014). Dietary cholesterol supplementation to a plant-based diet suppresses the complete pathway of cholesterol synthesis and induces bile acid production in Atlantic salmon (*Salmo salar* L.). *British Journal of Nutrition*, *111*(12), 2089–2103. https://doi.org/10.1017/S0007114514000373

Kulkarni, A., Dangat, K., Kale, A., Sable, P., Chavan-Gautam, P., & Joshi, S. (2011). Effects of altered maternal folic acid, vitamin B12 and docosahexaenoic acid on placental global DNA methylation patterns in wistar rats. *PLoS ONE*, *6*(3), e17706. https://doi.org/10.1371/journal.pone.0017706

Leaver, M. J., Bautista, J. M., Björnsson, B. T., Jönsson, E., Krey, G., Tocher, D. R., & Torstensen, B. E. (2008). Towards fish lipid nutrigenomics: Current state and prospects for fin-fish aquaculture. *Reviews in Fisheries Science*, *16*(April), 73–94. https://doi.org/10.1080/10641260802325278

Leaver, M. J., Villeneuve, L. A., Obach, A., Jensen, L., Bron, J. E., Tocher, D. R., & Taggart, J. B. (2008). Functional genomics reveals increases in cholesterol biosynthetic genes and highly unsaturated fatty acid biosynthesis after dietary substitution of fish oil with vegetable oils in Atlantic salmon (*Salmo salar*). *BMC Genomics*, *9*, 299. https://doi.org/10.1186/1471-2164-9-299

Lehner, R., & Quiroga, A. D. (2016). Fatty acid handling in mammalian cells. In N. D. Ridgway & R. S. McLeod (Eds.), *Biochemistry of lipids, lipoproteins and membranes* (pp. 149–184). Amsterdam, The Netherlands: Elsevier. https://doi.org/10.1016/b978-0-444-63438-2.00005-5

Li, B., & Dewey, C. N. (2011). RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*(1), 323. https://doi.org/10.1186/1471-2105-12-323

Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Kent, M. P., Nome, T., . . . Davidson, W. S. (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature*, *533*(7602), 200–205. https://doi.org/10.1038/nature17164

Lorgen, M., Casadei, E., Krol, E., Douglas, A., Birnie, M. J., Ebbesson, L. O. E., . . . Martin, A. M. S. (2015). Functional divergence of type 2 deiodinase paralogs in the Atlantic salmon. *Current Biology*, *25*(7), 936–941. https://doi.org/10.1016/j.cub.2015.01.074

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

Morais, S., Pratoomyot, J., Taggart, J. B., Bron, J. E., Guy, D. R., Bell, J. G., & Tocher, D. R. (2011). Genotype-specific responses in Atlantic salmon (*Salmo salar*) subject to dietary fish oil replacement by vegetable oil: A liver transcriptomic analysis. *BMC Genomics*, *12*(1), 255. https://doi.org/10.1186/1471-2164-12-255

Oxley, A., Jolly, C., Eide, T., Jordal, A.-E. O., Svardal, A., & Olsen, R.-E. (2010). The combined impact of plant-derived dietary ingredients and acute stress on the intestinal arachidonic acid cascade in Atlantic salmon (*Salmo salar*). *The British Journal of Nutrition*, *103*(6), 851–861. https://doi.org/10.1017/S0007114509992467

Price, M. N., Dehal, P. S., Arkin, A. P., Nawrocki, E., Kolbe, D., Eddy, S., . . . Meyer, F. (2010). FASTTREE 2—Approximately maximum-likelihood trees for large alignments. *PLoS ONE*, *5*(3), e9490. https://doi.org/10.1371/journal.pone.0009490

Rimoldi, S., Benedito-Palos, L., Terova, G., & Pérez-Sánchez, J. (2016). Wide-targeted gene expression infers tissue-specific molecular signatures of lipid metabolism in fed and fasted fish. *Reviews in Fish Biology and Fisheries*, *26*(1), 93–108. https://doi.org/10.1007/s11160-015-9408-8

Robertson, F. M., Gundappa, M. K., Grammes, F., Hvidsten, T. R., Redmond, A. K., Martin, S. A. M., . . . Macqueen, D. J. (2017). Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification. *Genome Biology*, https://doi.org/doi:10.1101/098582

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). EDGER: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Ruyter, B., Røsjø, C., Måsøval, K., Einen, O., & Thomassen, M. S. (2000). Influence of dietary n-3 fatty acids on the desaturation and elongation of [1- 14 C] 18: 2 n-6 and [1- 14 C] 18: 3 n-3 in Atlantic salmon hepatocytes. *Fish Physiology and Biochemistry*, *23*(2), 151–158.

Sánchez-Gurmaches, J., Østbye, T.-K., Navarro, I., Torgersen, J., Hevrøy, E. M., Ruyter, B., & Torstensen, B. E. (2011). In vivo and in vitro insulin and fasting control of the transmembrane fatty acid transport proteins in Atlantic salmon (*Salmo salar*). *American Journal of Physiology. Regulatory, Integrative and Comparative Physiology*, *301*(4), R947–R957. https://doi.org/10.1152/ajpregu.00289.2011

Sheridan, M. A. (1989). Alterations in lipid metabolism accompanying smoltification and seawater adaptation of salmonid fish. *Aquaculture*, *82*(1–4), 191–203. https://doi.org/10.1016/0044-8486(89)90408-0

Stefansson, S. O., Björnsson, B. T., Ebbesson, L. O., & McCormick, S. D. (2008). Smoltification. *Fish Larval Physiology*, *639–681*, https://doi.org/DOI: 10.1111/j.1095-8649.2009.02440_2.x

Storebakken, T. (2002). Atlantic salmon, *Salmo salar*. In C. D. Webster, & C. Lim (Eds.), *Nutrient requirements and feeding of finfish for aquaculture* (pp. 79–102). Wallingford: CABI. https://doi.org/10.1079/9780851995199.0079

Sushchik, N. N., Gladyshev, M. I., Moskvichova, A. V., Makhutova, O. N., & Kalachova, G. S. (2003). Comparison of fatty acid composition in major lipid classes of the dominant benthic invertebrates of the Yenisei river. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, *134*(1), 111–122. https://doi.org/10.1016/S1096-4959(02)00191-4

Tocher, D. R. (2003). Metabolism and functions of lipids and fatty acids in teleost fish. *Reviews in Fisheries Science*, *11*(2), 107–184. https://doi.org/10.1080/713610925

Tocher, D. R., Bell, J. G., Dick, J. R., Henderson, R. J., McGhee, F., Michell, D., & Morris, P. C. (2000). Polyunsaturated fatty acid metabolism in Atlantic salmon (*Salmo salar*) undergoing parr-smolt transformation and the effects of dietary linseed and rapeseed oils. *Fish Physiology and Biochemistry*, *23*(1), 59–73. https://doi.org/10.1023/A:1007807201093

Tocher, D. R., Bell, J. G., MacGlaughlin, P., McGhee, F., & Dick, J. R. (2001). Hepatocyte fatty acid desaturation and polyunsaturated fatty acid composition of liver in salmonids: Effects of dietary vegetable oil. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, *130*(2), 257–270. https://doi.org/10.1016/S1096-4959(01)00429-8

Tocher, D. R., Fonseca-Madrigal, J., Bell, J. G., Dick, J. R., Henderson, R. J., & Sargent, J. R. (2002). Effects of diets containing linseed oil on fatty acid desaturation and oxidation in hepatocytes and intestinal enterocytes in Atlantic salmon (*Salmo salar*). *Fish Physiology and Biochemistry*, *26*(2), 157–170. https://doi.org/10.1023/A:1025416731014

Varadharajan, S., Sandve, S. R., Tørresen, O. K., Lien, S., Vollestad, L. A., Jentoft, S., . . . Jakobsen, K. S. (2017). The grayling genome reveals selection on gene expression regulation after whole genome duplication. *bioRxiv*. https://doi.org/10.1101/153270

Verleyen, T., Forcades, M., Verhe, R., Dewettinck, K., Huyghebaert, A., & De Greyt, W. (2002). Analysis of free and esterified sterols in vegetable oils. *Journal of the American Oil Chemists' Society*, *79*(2), 117–122. https://doi.org/10.1007/s11746-002-0444-3

Zheng, X., Torstensen, B. E., Tocher, D. R., Dick, J. R., Henderson, R. J., & Bell, J. G. (2005). Environmental and dietary influences on highly unsaturated fatty acid biosynthesis and expression of fatty acyl desaturase and elongase genes in liver of Atlantic salmon (*Salmo salar*). *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids*, *1734*(1), 13–24. https://doi.org/10.1016/j.bbalip.2005.01.006

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Gillard G, Harvey TN, Gjuvsland A, et al. Life-stage-associated remodelling of lipid metabolism regulation in Atlantic salmon. *Mol Ecol*. 2018;27:1200–1213. https://doi.org/10.1111/mec.14533

Paper II

The Grayling Genome Reveals Selection on Gene Expression Regulation after Whole-Genome Duplication

Srinidhi Varadharajan[1,†], Simen R. Sandve[2,*,†], Gareth B. Gillard[3], Ole K. Tørresen[1], Teshome D. Mulugeta[2], Torgeir R. Hvidsten[3,4], Sigbjørn Lien[2], Leif Asbjørn Vøllestad[1], Sissel Jentoft[1], Alexander J. Nederbragt[1,5], and Kjetill S. Jakobsen[1,*]

[1]Centre for Ecological and Evolutionary Synthesis (CEES), Department of Biosciences, University of Oslo, Norway

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

[3]Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Ås, Norway

[4]Umeå Plant Science Centre, Department of Plant Physiology, Umeå University, Sweden

[5]Biomedical Informatics Research Group, Department of Informatics, University of Oslo, Norway

[†]These authors contributed equally to this work.

*Corresponding authors: E-mails: k.s.jakobsen@ibv.uio.no; simen.sandve@nmbu.no.

Accepted: September 12, 2018

Data deposition: The raw sequencing reads data have been deposited at ENA under accession PRJEB21333. The Atlantic salmon liver expression data are available at ENA or NCBI under sample accessions: SAMEA104483623, SAMEA104483624, SAMEA104483627, and SAMEA104483628.

Abstract

Whole-genome duplication (WGD) has been a major evolutionary driver of increased genomic complexity in vertebrates. One such event occurred in the salmonid family ∼80 Ma (Ss4R) giving rise to a plethora of structural and regulatory duplicate-driven divergence, making salmonids an exemplary system to investigate the evolutionary consequences of WGD. Here, we present a draft genome assembly of European grayling (*Thymallus thymallus*) and use this in a comparative framework to study evolution of gene regulation following WGD. Among the Ss4R duplicates identified in European grayling and Atlantic salmon (*Salmo salar*), one-third reflect nonneutral tissue expression evolution, with strong purifying selection, maintained over ∼50 Myr. Of these, the majority reflect conserved tissue regulation under strong selective constraints related to brain and neural-related functions, as well as higher-order protein–protein interactions. A small subset of the duplicates have evolved tissue regulatory expression divergence in a common ancestor, which have been subsequently conserved in both lineages, suggestive of adaptive divergence following WGD. These candidates for adaptive tissue expression divergence have elevated rates of protein coding- and promoter-sequence evolution and are enriched for immune- and lipid metabolism ontology terms. Lastly, lineage-specific duplicate divergence points toward underlying differences in adaptive pressures on expression regulation in the nonanadromous grayling versus the anadromous Atlantic salmon. Our findings enhance our understanding of the role of WGD in genome evolution and highlight cases of regulatory divergence of Ss4R duplicates, possibly related to a niche shift in early salmonid evolution.

Key words: *Thymallus thymallus*, genome assembly, salmonid, WGD, rediploidization, lineage-specific ohnolog resolution.

Introduction

Whole-genome duplication (WGD) through spontaneous doubling of all chromosomes (autopolyploidization) has played a vital role in the evolution of vertebrate genome complexity (Van de Peer et al. 2009). However, the role of selection in shaping novel adaptations from the redundancy that arises from WGD is not well understood. The idea that functional redundancy arising from gene duplication sparks evolution of novel traits and adaptations was pioneered by Ohno (1970). Duplicate genes that escape loss or pseudogenization are known to acquire novel regulation and expression divergence (Lynch and Conery 2000; Zhang 2003; Conant and Wolfe 2008). Functional genomic studies over the past decade have demonstrated that large-scale duplications lead to the rewiring of regulatory networks through divergence of spatial and temporal expression patterns (Osborn et al. 2003;

De Smet et al. 2017). As changes in gene regulation are known to be important in the evolution of phenotypic diversity and complex trait variation (Carroll 2000; Wray 2003), these post-WGD shifts in expression regulation may provide a substrate for adaptive evolution. Several studies have investigated the genome-wide consequences of WGD on gene expression evolution in vertebrates (Sémon and Wolfe 2008; Kassahn et al. 2009; Berthelot et al. 2014; Li et al. 2015; Acharya and Ghosh 2016; Lien et al. 2016; Robertson et al. 2017) and have revealed that a large proportion of gene duplicates have evolved substantial regulatory divergence of which, in most cases, one copy retains ancestral-like regulation (consistent with Ohno's model of regulatory neofunctionalization). However, to what extent this divergence in expression is linked to adaptation remains to be understood. A major factor contributing to this knowledge gap is the lack of studies that integrate expression data from multiple species sharing the same WGD (Hermansen et al. 2016). Such studies would allow us to distinguish neutral evolutionary divergence in regulation from regulatory changes representing adaptive divergence and those maintained by purifying selection.

Salmonids have emerged as a model for studying consequences of autopolyploidization in vertebrates, owing to their relatively young WGD event (Ss4R, <100 Ma) (Ohno 1970; Alexandrou et al. 2013) and ongoing rediploidization (Macqueen and Johnston 2014; Lien et al. 2016; Limborg et al. 2016; Robertson et al. 2017). Directly following autopolyploidization, duplicated chromosomes pair randomly with any of their homologous counterparts resulting in an increased risk of formation of multivalents and consequently production of nonviable aneuploid gametes. Restoring bivalent chromosome pairing is therefore a critical step toward a functional genome post-WGD (Wolfe 2001). This can be achieved through, for example, structural rearrangements that suppress recombination, block multivalent formation, and drive the process of returning to a functional diploid state (i.e., rediploidization). As the mutational process is stochastic, rediploidization occurs independently for different chromosomes. As a result, the divergence of gene duplicates resulting from WGD (referred to as ohnologs) is also achieved independently for different chromosomes and hence occurs at different rates in various genomic regions. Recent studies on genome evolution subsequent to Ss4R have shown that the rediploidization process temporally overlaps with species radiation, resulting in lineage-specific ohnolog resolution (LORe) that may fuel differentiation of genome structure and function (Macqueen and Johnston 2014; Robertson et al. 2017). In fact, due to the delayed rediploidization, only 75% of the duplicated genome diverged before the basal split in the salmonid family ∼60 Ma (henceforth referred to as ancestral ohnolog resolution, AORe). Consequently, ∼25% of the Ss4R duplicates have experienced independent rediploidization histories after the basal salmonid divergence resulting in the Salmoninae and Thymallinae clades. Interestingly, the species within these two clades have also evolved widely different genome structures, ecology, physiology, and life history adaptations (Hendry and Stearns 2004). In contrast to the *Thymallus* lineage, the species in the subfamily Salmoninae have fewer and highly derived chromosomes resulting from large-scale chromosomal translocations and fusions (supplementary fig. S1, Supplementary Material online), display extreme phenotypic plasticity, and have evolved the capability of migrating between fresh and saltwater habitats (referred to as anadromy) (Nygren et al. 1971; Hartley 1987; Phillips and Ráb 2001; Alexandrou et al. 2013; Ocalewicz et al. 2013). This unique combination of both shared and lineage-specific rediploidization histories, and striking differences in genome structure and adaptations, provides an ideal study system for addressing key questions about the evolutionary consequences of WGD.

To gain deeper insights into how selection has shaped the evolution of gene duplicates post-WGD, we have sequenced, assembled, and annotated the genome of the European grayling (*Thymallus thymallus* Linnaeus, 1758), a species representative of an early diverging nonanadromous salmonid lineage, Thymallinae. We use this novel genomic resource in a comparative phylogenomic framework with the genome of Atlantic salmon (*Salmo salar*), of the Salmoninae lineage, to address the consequences of Ss4R WGD on lineage-specific rediploidization and selection on ohnolog gene expression regulation.

Our results reveal signatures of adaptive regulatory divergence of ohnologs, strong selective constraints on expression evolution in brain and neural-related genes, and lineage-specific ohnolog divergence. Moreover, diverse biological processes correspond to differences in evolutionary constraints during the 88–100 Myr of evolution post-WGD, pointing toward underlying differences in adaptive pressures in nonanadromous grayling and anadromous Atlantic salmon.

Materials and Methods

Sampling and Sequencing

A male grayling specimen was sampled outside of its spawning season (October 2012) from the River Glomma at Evenstad, Norway (61°25′0.1″N 11°9′49.7″E). The fish was humanely sacrificed, and various tissue samples were immediately extracted and conserved for later DNA and RNA analyses. Fin clips were stored on 96% ethanol for DNA sequencing. Tissues from the muscle, the gonad, the liver, the head kidney, the spleen, the brain, the eye, the gill, and the heart were stored in RNAlater for RNA extraction.

The DNA was extracted from fin clips using a standard high-salt DNA extraction protocol. A paired-end library with an insert size ∼180 bp (150 bp read length) and mate pair libraries of insert size ∼3 and 6 kb (100 bp read length) were sequenced using the Illumina HiSeq2000 platform (supplementary table S1, Supplementary Material online). Total RNA was extracted from the different tissue samples using the RNeasy mini kit (Qiagen) following the manufacturer's instructions. The library construction and sequencing were carried out using Illumina TruSeq RNA Preparation kit on Illumina HiSeq2000 (supplementary table S2, Supplementary Material online). All the library preparation and sequencing were performed at the McGill University and the Génome Québec Innovation Centre.

Genome Assembly and Validation

The sequences were checked for their quality, and adapter trimming was performed using cutadapt (version 1.0) (Martin 2011). A de novo assembly was generated with Allpaths-LG (release R48777) (Gnerre et al. 2011) using the 180-bp paired-end library and the mate pair (3 and 6 kb) libraries. Assembly polishing was carried out using pilon (version 1.9) (Walker et al. 2014). The high copy number of mitochondrial DNA often leads to high read coverage and thus misassembly. The mitochondrial genome sequence in the assembly was thus reassembled by extracting the reads that mapped to the grayling (*Thymallus thymallus*) mtDNA sequence (GenBank accession number: NC_012928), followed by a variant calling step using Genome Analysis Toolkit (GATK) (version 3.4-46) (Van der Auwera et al. 2013). The consensus mtDNA sequence thus obtained was added back to the assembly.

To identify and correct possibly erroneous grayling scaffolds, we aligned the scaffolds against a repeat masked version of the Atlantic salmon genome (Lien et al. 2016) using megablast (*e*-value threshold 1e-250). Stringent filtering of the aligned scaffolds (representing 1.3 Gb of the 1.4-Gb assembly) identified 13 likely chimeric scaffolds mapping to two or more salmon chromosomes (supplementary file 1, Supplementary Material online), which were then selectively "broken" between, apparently, incorrectly linked contigs.

Transcriptome Assembly

The RNA-Seq data from all the tissue samples were quality checked using FastQC (version 0.9.2). The sequences were assembled using the following two methods. Firstly, a de novo assembly was performed using the Trinity (version 2.0.6) (Grabherr et al. 2011) pipeline with default parameters coupled with in silico normalization. This resulted in 730,471 assembled transcript sequences with a mean length of 713 bases. RSEM protocol-based abundance estimation within the Trinity package was performed where the RNA-Seq reads were first aligned back to the assembled transcripts using Bowtie2 (Faust and Hall 2012), followed by calculation of various estimates including normalized expression values such as FPKM (fragments per kilobase million). A script provided with Trinity was then used to filter transcripts based on FPKM, retaining only those transcripts with a FPKM of at least one.

Secondly, reference-guided RNA assembly was performed by aligning the RNA reads to the genome assembly using STAR (version 2.4.1b) (Dobin et al. 2013). Cufflinks (version 2.1.1) (Trapnell et al. 2010; Dobin et al. 2013) and TransDecoder (Haas et al. 2013) were used for transcript prediction and ORF (open reading frame) prediction, respectively. The resulting transcripts were filtered and retained based on homology against zebrafish and stickleback proteins, using BlastP and PFAM (1e-05). The de novo method resulted in 134,368 transcripts and the reference-based approach followed by filtering resulting in 55,346 transcripts.

Genome Annotation

A de novo repeat library was constructed using RepeatModeler with default parameters. Any sequence in the de novo library matching a known gene was removed using BlastX against the UniProt database. CENSOR and TEclass were used for classification of sequences that were not classified by RepeatModeler. Gene models were predicted using an automatic annotation pipeline involving MAKER (version 2.31.8), in a two-pass iterative approach (as described in https://github.com/sujaikumar/assemblage/blob/master/README-annotation.md). Firstly, ab initio gene predictions were generated using GeneMark ES (version 2.3e) (Lomsadze et al. 2005) and SNAP (version 20131129) (Korf 2004) trained on core eukaryotic gene data set (CEGMA). The first round of MAKER was then run using the following data as input: ab initio gene models, the UniProt database as protein evidence, the de novo identified repeat library and the de novo and reference guided transcriptome assemblies, as well as the transcript sequences from the recent Atlantic salmon annotation (Lien et al. 2016). The second pass involved additional data from training AUGUSTUS (Stanke et al. 2008) and SNAP models on the generated MAKER predictions.

Putative functions were added to the gene models using BlastP against the UniProt database (*e*-value 1e-5), and the domain annotations were added using InterProScan (version 5.4-47) (Quevillon et al. 2005). Using the MAKER standard filtering approach, the resulting set of genes was first filtered using the threshold of AED (Annotation Edit Distance), retaining gene models with AED score <1 and PFAM domain annotation. AED is a quality score given by MAKER that ranges from 0 to 1 and indicates the concordance between predicted gene model and the evidence provided, where an AED of 0 indicates that the gene models completely conforms to the evidence. Further, for the genes with AED score of 1 and no domain annotations, a more conservative BLAST search was performed against UniProt proteins and Atlantic salmon proteins with an *e*-value cut off of 1e-20. The genes with hits to either of these databases were also retained. The completeness of the annotations was again assessed using

CEGMA (Parra et al. 2007) and benchmarking universal single-copy ortholog (BUSCO) (Simão et al. 2015).

Analysis of Orthologous Groups

We used orthofinder (version 0.2.8, *e*-value threshold at 1e-05) (Emms and Kelly 2015) to identify orthologous gene groups (i.e., orthogroup). As input to orthofinder, we used the MAKER-derived *T. thymallus* gene models as well as protein sequences from three additional salmonid species (Atlantic salmon, rainbow trout, and coho salmon), four non-salmonid teleost species (*Esox lucius*, *Danio rerio, Gasterosteus aculeatus*, and *Oryzias latipes*), and two mammalian outgroups (*Homo sapiens* and *Mus musculus*). Rainbow trout protein annotations were taken from https://www.genoscope.cns.fr/trout/. Atlantic salmon (Annotation Release 100), *Esox lucius* (Annotation Release 101) data were downloaded from NCBI ftp server (ftp://ftp.ncbi.nlm.nih.gov/genomes/). The transcriptome data for Coho salmon were obtained from NCBI (GDQG00000000.1) and translated using TransDecoder. All other annotations were downloaded from ENSEMBL.

Each set of orthogroup proteins was then aligned using MAFFT(v7.130) (Katoh et al. 2002) using default settings, and the resulting alignments were then used to infer maximum-likelihood gene trees using FastTree (v2.1.8) (Price et al. 2010) (figs. 1*a* and *b*). As we were only interested in gene trees containing information on Ss4R duplicates, complex orthogroup gene trees (i.e., containing 2R or 3R duplicates of salmonid genes) were subdivided into the smallest possible subtrees. To this end, we developed an algorithm to extract all clans (defined as unrooted monophyletic clade) from each unrooted tree (Wilkinson et al. 2007) with two monophyletic salmonid tips as well as nonsalmonid outgroups resulting in a final set of 20,342 gene trees. In total, 31,291 grayling genes were assigned to a clan (fig. 1 and supplementary fig. S2, Supplementary Material online). We then identified homeology in the Atlantic salmon genome by integrating all-versus-all protein BLAST alignments with a priori information of Ss4R synteny as described by Lien et al. (2016). Using the homeology information, we inferred a set of high-confidence ohnologs originating from Ss4R. The scaffold length distribution and number of genes per scaffold containing the inferred Ss4R genes are plotted in supplementary figure S13, Supplementary Material online. The clans were grouped based on the gene tree topology into duplicates representing LORe and those with ancestrally diverged duplicates (AORe). The LORe regions were further categorized into two (duplicated or collapsed) based on the number of corresponding *T. thymallus* orthologs. These data were plotted on Atlantic salmon chromosomes using circos plot generated using OmicCircos (https://bioconductor.org/packages/release/bioc/html/OmicCircos.html). The LORe and AORe ohnologs with two copies in each species are hereafter referred to as ohnolog-tetrads (see supplementary fig. S14, Supplementary Material online, for the summary of the above steps).

Expression Divergence and Conservation

The grayling RNA-Seq reads from each of the eight tissues (liver, muscle, spleen, heart, head kidney, eye, brain, and gills) were mapped to the genome assembly using STAR (version 2.4.1b). The reads uniquely mapping to the gene features were quantified using htseq-count (Anders et al. 2015). The CPM (counts per million) value, here used as a proxy for expression, was then calculated using edgeR (Robinson et al. 2010). Similar CPM data sets were obtained from Atlantic salmon RNA-Seq data reported by Lien et al. (2016).

Filtering of ortholog groups (i.e., clans) was performed prior to analyses of expression evolution of Ss4R ohnologs: 1) We only considered Ss4R duplicates that were retained in both Atlantic salmon and grayling, and 2) the Ss4R duplicates were classified into AORe or LORe, based on topologies of the ortholog group gene trees, only gene pairs with non-zero CPM value were considered. This filtering resulted in a set of 5,070 duplicate pairs from both Atlantic salmon and grayling (ohnolog-tetrads) (summarized in supplementary fig. S14, Supplementary Material online). The gene expression values from the gene duplicates in the ohnolog-tetrads were clustered using hclust function in R, using Pearson correlation into eight tissue-dominated clusters. The expression pattern in the eight clusters of the genes in ohnolog-tetrads was used to further classify them into one of the ohnolog expression evolution categories (see table 2). The ohnolog-tetrads were further filtered based on expected topologies under LORe and AORe scenarios (see supplementary fig. S14, Supplementary Material online, for summary). Heatmaps of expression counts were plotted using pheatmap package in R (https://CRAN.R-project.org/package=pheatmap). To quantify the breadth of expression (i.e., the number of tissues a gene is expressed in), we calculated the tissue specificity index Tau (Yanai et al. 2005) for all the genes in ohnolog-tetrads, where a τ value approaching 1 indicates higher tissue specificity while 0 indicates ubiquitous expression.

Expression Comparison in Liver

Utilizing independent liver tissue samples, we compared differential expression in liver tissue gene expression among ohnologs of grayling and Atlantic salmon with their ohnolog-tetrad tissue expression evolution categories. The liver samples from four grayling individuals were sampled in the river Gudbrandsdalslågen (61°18′53.09″N 10°18′1.53″E). The samples were from two males (370,375 mm) and two females (330,360 mm). The fish was euthanized and dissected immediately after capture, and the liver was stored in RNAlater. Total RNA was extracted and 100-bp single-end read libraries were generated for two individuals and

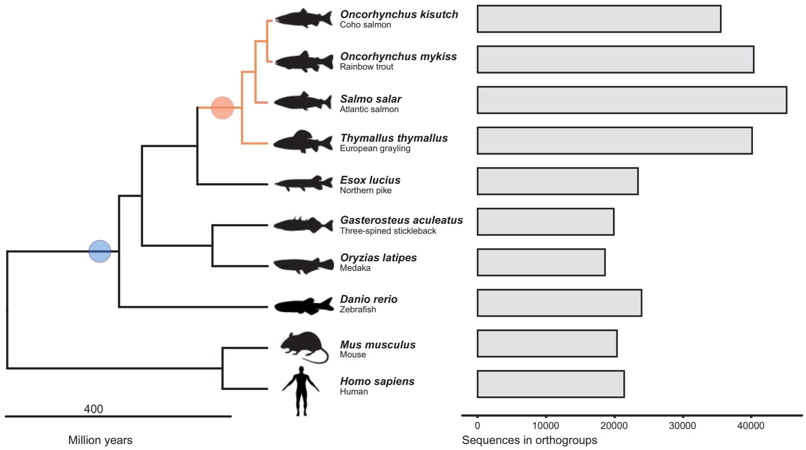


Fig. 1.—Species and genes in ortholog groups. Left: Phylogenetic relationship of species used for constructing ortholog groups and gene trees. The blue circle indicates the 3R-WGD event, while the Ss4R event is indicated with an orange circle. Right: Number of genes assigned to ortholog groups in each of the species used in the analysis.

sequenced using the Illumina HiSeq4000 platform. For the other two individuals, 150-bp paired-end read libraries were generated and sequenced using the Illumina HiSeq2500 platform. RNA-Seq data for an additional four Atlantic salmon liver tissue samples were obtained from a feeding experiment (Gillard et al. 2018). Presmolt salmon were raised on fish oil-based diets under freshwater conditions.

The RNA-Seq read data were quality processed using CutAdapt (Martin 2011) before alignment we aligned the reads to grayling or Atlantic salmon (ICSASG_v2; Lien et al. 2016) genomes, respectively, using STAR (Dobin et al. 2013). RSEM (Li and Dewey 2011) expected counts were generated for gene features. EdgeR (Robinson et al. 2010) was used to generate normalized library sizes of samples (TMM normalization), followed by a differential expression analysis using the exact test method between the gene expression of both the grayling and Atlantic salmon ohnologs in each ohnolog-tetrad. The fold change (log2 scaled) and significance of differential expression (false discovery rate-corrected *P*-values) were produced for grayling and Atlantic salmon duplicates, as well as relative counts in the form of CPM.

Sequence Evolution

To estimate coding sequence evolution rates, we converted amino acid alignments to codon alignments using pal2nal (Suyama et al. 2006). The “seqinr” R package (http://seqinr.r-forge.r-project.org/) was used to calculate pairwise d*N* and d*S* values for all sequences in each alignment using the “kaks” function. For in-depth analyses of branch-specific sequence evolution of the cystic fibrosis transmembrane conductance regulator (*CFTR*) genes, we used the codeml in PAML (version 4.7a) (Yang 1997). To assess whether sequences in the *CFTR* gene tree evolved under similar selection pressure, we contrasted a fixed d*N*/d*S* ratio (1-ratio) model and a free-ratio model of codon evolution. A likelihood ratio test was conducted to assess whether a free-ratio model was a significantly better fit to the data. Branch-specific d*N*/d*S* values were extracted from the maximum likelihood results for the free ratios model.

The two Pacific salmon genes in the *CFTR* tree (fig. 5) correspond to a gene from rainbow trout and another from Coho salmon. A BLAT (Kent 2002) search of *CFTR* gene against the rainbow trout assembly (https://www.genoscope.cns.fr/trout/) resulted in hits on three different scaffolds, with one complete hit and two other partial hits on unplaced scaffolds. Additionally, Coho salmon data are based on a set of genes inferred from transcriptome data. Therefore, the presence of a single copy in the tree for the two species is likely an assembly artifact.

Genome-Wide Identification of Transcription Factor-Binding Sites

A total of 13,544 metazoan transcription factor protein sequences together with their binding site represented as position-specific scoring matrices (PSSMs referred to as motifs) were collected from transcription factor-binding profile databases such as CISBP, JASPAR, 3D-footprint, UniPROBE, HumanTF, HOCOMOCO, HumanTF2, and TRANSFAC.

DNA sequences from upstream promoter regions of Atlantic salmon (−1,000/+200 bp from TSS) were extracted. A first-order Markov model was created from the entire set of upstream promoter regions using the fasta-get-markov program in the MEME Suite (Bailey et al. 2009). This background model was used to convert frequency matrices into log-odds score matrices. We performed a genome-wide transcription factor-binding sites prediction in the Atlantic salmon genome using the PSSM collection and the Finding Individual Motif Occurrences (FIMO) (Grant et al. 2011) tool in the MEME Suite (*P*-value = 0.0001 and FDR = 0.2).

Motif similarity between Atlantic salmon ohnolog promoters was scored using the "Jaccard coefficient." The promoter Jaccard coefficient is defined as

$$J(A,B) \;= \frac{A \cap B}{|A| \;+\; |B| \;-\; |A \cap B|},$$

where *A* and *B* represents the type of motifs that were present in promoters of the *A* and *B* ohnolog copies. If *A* and *B* are empty, we set $J(A, B) = 0$ where $0 \leq J(A, B) \leq 1$.

Gene Ontology Analysis

The gene ontology (GO) term enrichment analysis was performed using the "elim" algorithm implemented in the "topGO" R package (http://www.bioconductor.org/packages/2.12/bioc/html/topGO.html), with a significance threshold of 0.05 using the genes from all ohnolog-tetrad categories as the background. GO terms were assigned to salmon genes using Blast2GO (Conesa et al. 2005).

Results

Genome Assembly and Annotation

We sequenced the genome of a wild-caught male grayling individual, sampled from the Norwegian river Glomma (61°25′0.1″N 11°9′49.7″E), using the Illumina HiSeq 2000 platform (supplementary tables S1 and S2, Supplementary Material online). De novo assembly was performed using ALLPATHS-LG (Gnerre et al. 2011), followed by assembly correction using Pilon (Walker et al. 2014), resulting in 24,343 scaffolds with an N50 of 284 kb and a total size of 1.468 Gb (table 1). The scaffolds represent ∼85% of the k-mer-based genome size estimate of ∼1.8 Gb. The *C*-values estimated previously for European grayling are 2.1 pg (http://www.genomesize.com/) and 1.9 pg (Hartley 1987). To annotate gene structures, we used RNA-Seq data from nine tissues extracted from the sequenced individual. We constructed transcriptome assemblies using both de novo and reference-based methods. Repeat masking with a repeat library constructed using a combination of homology and de novo-based methods identified and masked ∼600 Mb (∼40%) of the assembly and was dominated by class1 DNA transposable elements (supplementary table S3 and a repeat landscape in supplementary fig. S2, Supplementary Material online). Finally, the transcriptome assemblies, the de novo-identified repeats along with the UniProt proteins (UniProt Consortium 2015), and Atlantic salmon coding sequences (Lien et al. 2016) were utilized in the MAKER annotation pipeline, predicting a total of 117,944 gene models, of which 48,753 protein-coding genes were retained based on AED score, homology with UniProt and Atlantic salmon proteins or presence of known domains. Assembly completeness was assessed at the gene level by looking for conserved genes using CEGMA and BUSCO. The assembly contains 236 (95.16%) out of 248 conserved eukaryotic genes (CEGs) with 200 (80.65%) complete CEGs. Of the 4,584 BUSCO (database: Actinopterygii, odb9), 4,102 complete (89.5%) and 179 (3.9%%) fragmented genes were found in the assembly (table 1).

Divergent Rediploidization Rates among the Salmonid Lineages

Previous studies have suggested that up to 25% of the genome of the most recent common salmonid ancestor was still tetraploid when the grayling and Atlantic salmon lineages diverged (Lien et al. 2016; Robertson et al. 2017). To test this hypothesis, we used a phylogenomic approach to characterize rediploidization following Ss4R in grayling. We inferred 23,782 groups of orthologous genes (i.e., ortholog groups or orthogroups) using gene models from *Homo sapiens* (human), *Mus musculus* (mouse), *Danio rerio* (zebrafish), *Gasterosteus aculeatus* (three-spined stickleback), *Oryzias latipes* (medaka), *Esox lucius* (northern pike), *Salmo salar* (Atlantic salmon), *Oncorhynchus mykiss* (rainbow trout), and *Oncorhynchus kisutch* (coho salmon) (fig. 1). These orthogroups were used to infer gene trees. In total, 20,342 gene trees contained WGD events older than Ss4R (Ts3R or 2R) and were further subdivided into smaller subgroups (i.e., unrooted monophyletic clade termed as clans, see Materials and Methods for details and supplementary fig. S3, Supplementary Material online). To identify orthogroups with retained Ss4R duplicates, we relied on the high-quality reference genome of Atlantic salmon (Lien et al. 2016). A synteny-aware blast approach (Lien et al. 2016) was first used to identify Ss4R duplicate/ohnolog pairs in the Atlantic salmon genome, and this information was used to identify a total of 8,527 gene trees containing high-confidence ohnologs originating from Ss4R. Finally, gene trees were classified based on the tree topology into duplicates conforming to LORe and those with ancestrally diverged duplicates following the topology expected under AORe (fig. 2*a*). In total, 3,367 gene trees correspond to LORe regions (2,403 with a single copy in grayling) and 5,160 correspond to an AORe-like topology. These data were cross-checked with the LORe coordinates suggested by Robertson et al. (2017), and genes with LORe-type topologies from non-LORe regions of the genome

Table 1
Grayling Genome Assembly Statistics

Assembly Statistics		Assembly Validation	
Total size of scaffolds (bp)	1,468,519,221	Complete CEGMA[a] genes	80.65% (200/248)
Number of scaffolds	24,369	Partial CEGMA genes	95.16% (236/248)
Scaffold N50 (bp)	283,328	Complete BUSCOs[b]	4,102 (89.5%)
Longest scaffold (bp)	2,502,076	Complete Duplicated BUSCOs	1,724 (37.6%)
Total size of contigs (bp)	1,278,330,545	Fragmented BUSCOS	179 (3.9%)
Number of contigs	216,549	Missing BUSCOS	303 (6.6%)
Contig N50 (bp)	11,206	Total BUSCOS searched	4,584

[a]Based on 248 highly CEGS.
[b]Based on a set of 4,584 Actinopterygii odb9 BUSCOs.

were discarded. The final set (henceforth referred to as ohnolog-tetrads) consisted of 5,475 gene trees containing Ss4R duplicates from both species (4,735 AORe, 740 LORe). In addition, 482 ortholog sets contained Ss4R duplicates in Atlantic salmon but not in grayling.

To identify regions of ancestral and lineage-specific rediploidization in the grayling genome, we assigned genes from gene trees that contained Ss4R duplicates to genomic positions on the Atlantic salmon chromosomes (fig. 2*b*). In Atlantic salmon, several homeologous chromosome arms (2p-5q, 2q-12qa, 3q-6p, 4p-8q, 7q-17qb, 11qa-26, and 16qb-17qa) have previously been described as homeologous regions under delayed rediploidization (Lien et al. 2016; Robertson et al. 2017) (indicated in fig. 2*b* as red and blue ribbons). Interestingly, for the homeologous LORe regions 2q-12qa, 4p-8q, and 16qb-17qa in Atlantic salmon, we identified only one orthologous region in grayling, suggesting either loss of large duplicated blocks or sequence assembly collapse in grayling. To test the "assembly collapse" hypothesis, we mapped the grayling Illumina paired-end reads that were used for the assembly back to the grayling genome sequence using BWA-MEM (Li 2013) and determined the mapped read depth for each of the grayling genes. Single-copy grayling genes in LORe regions consistently displayed read depths (∼100×) twice that of the LORe duplicates in grayling (fig. 2*c* and supplementary fig. S4*a*, Supplementary Material online), indicating assembly collapse rather than loss of large chromosomal regions. Additionally, the single nucleotide polymorphism (SNP) density of the scaffolds in these regions computed using FreeBayes (Garrison and Marth 2012) (quality filter of 30) displayed values that were on an average twice that of the background SNP density, albeit with a much wider distribution (fig. 2*d* and supplementary fig. S4*b*, Supplementary Material online).

Ohnolog Tissue Gene Expression Regulation

To investigate the regulatory divergence in tissue expression following Ss4R, we exploited tissue expression atlases of Atlantic salmon and grayling in a coexpression analysis. Individual genes from 5,070 "expressed" ohnolog-tetrads (20,280 genes in total) were assigned to eight "tissue-dominant" expression clusters (Materials and Methods, supplementary fig. S5, Supplementary Material online). These coexpression clusters were used to identify ohnolog-tetrads conforming to expectations of expression patterns under five evolutionary scenarios (see table 2, fig. 3 and supplementary fig. S6, Supplementary Material online): Ancestral ohnolog divergence followed by independent purifying selection in both species (I), conserved tissue regulation in both species (II), lineage-specific regulatory divergence of one duplicate (III and IV). In addition, a fifth (V) scenario where regulation among duplicates is shared within species but different between species is expected to be common in genomic regions with LORe. Further, the ohnolog-tetrads where three or all four of the duplicates were in different tissue-expression clusters were grouped into a sixth "unclassified" category.

After applying a gene tree topology-based filtering criterion to the 5,070 ohnolog-tetrads (see Materials and Methods), 509 conforming to the expectations of LORe and 3,480 conforming to AORe gene tree topologies were retained for further analyses. Of the five evolutionary scenarios, conserved tissue regulation was the most common (∼25%), followed by species-specific divergence of a single duplicate (∼11% in Atlantic salmon and ∼15% in grayling). Categories I and V were the least common categories (table 2), and as expected, category V was significantly enriched in LORe regions (Fisher's exact test, two-sided, P-value < 0.0005).

To assess the directionality of the expression divergence relative to the presumed ancestral state, we compared tissue expression of the ohnolog-tetrads with that of the corresponding orthologs in northern pike (fig. 3). Previous studies have shown that genome-wide tissue-specific expression divergence among WGD ohnologs in teleosts mostly evolves through asymmetric divergence in tissue regulation (Lien et al. 2016; Sandve et al. 2018). The predominant expression pattern thus reflects one ohnolog copy retaining more regulatory similarity with unduplicated orthologs (regulatory neofunctionalization), with very small proportion (<1%) of

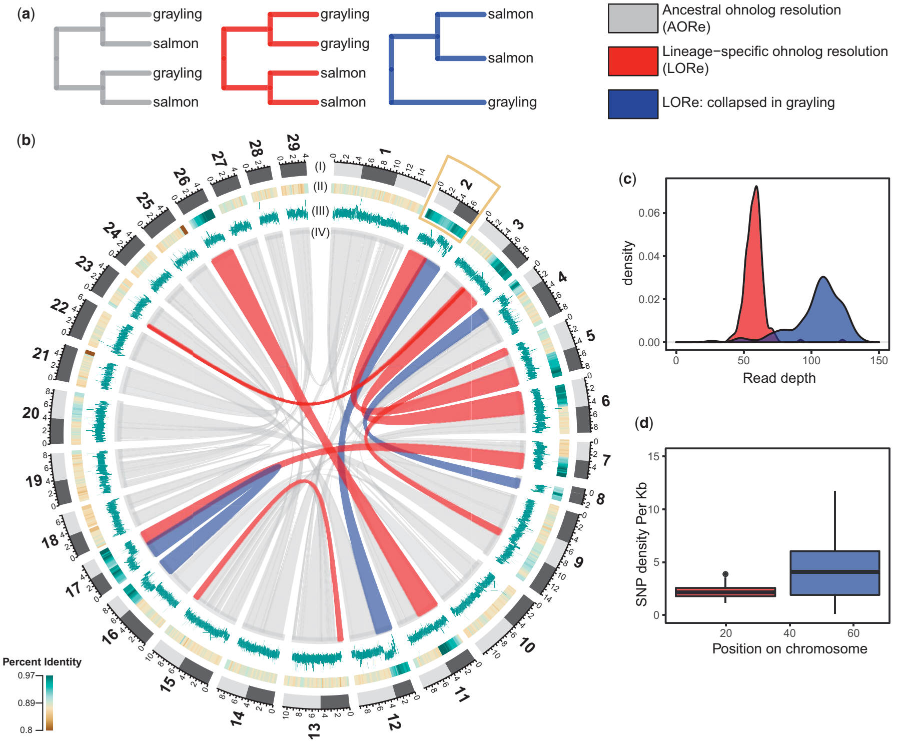


Fig. 2.—Rediploidization in grayling genome. (*a*) Gene tree topologies corresponding to the different models of ohnolog resolution (ancestral divergence of ohnologs [AORe; gray] and lineage-specific divergence of ohnologs (LORe in red and LORe-like regions, with repeat collapse in grayling, in blue). (*b*) Circos plot: Outer track (I) represents the 29 chromosomes of Atlantic salmon with chromosome arms indicated using light and dark gray. (II) Percent identity between duplicated genomic regions in Atlantic salmon with darker green representing higher percent identity (see color scale). (III) Average number of reads mapped to grayling genes in the corresponding regions. (IV) The gray ribbons represent the ancestrally diverged gene duplicate pairs (AORe), while the red ribbons represent the LORe duplicate pairs and the blue ribbons correspond to LORe regions with a collapsed assembly in grayling. The inset plots show the distribution of average depth of reads mapped to the grayling genes (*c*) and SNP density per kb (*d*) across chromosome 2 (marked with a yellow box in *b*).

ohnologs displaying characteristics of regulatory subfunctionalization (Lien et al. 2016). Under a model of subfunctionalization, the sum of expression levels of both ohnologs should correlate better to the assumed ancestral expression regulation than any of the individual ohnologs (Sandve et al. 2018). Therefore, we tested whether the divergence patterns leading to maintenance of the two ohnolog copies in the category I tetrads are associated with this atypical mode of expression divergence. Both the distribution of ohnolog tissue expression correlations (supplementary fig. S7, Supplementary Material online) and the patterns in the heatmaps (fig. 3) support the regulatory neofunctionalization pattern for all three evolutionary scenarios where we observe ohnolog divergence (categories I, III, and IV).

As different tissues are involved in different biological functions, we expect that the regulatory evolution is shaped by tissue-specific selective pressures (Gu and Su 2007). To test this, we evaluated the hypothesis that tissues are disproportionately contributing to ohnolog-tetrad divergence by comparing the "tissue-dominant" cluster distribution across all tetrads. For all evolutionary scenarios, between two and five tissue-expression clusters were significantly over- or underrepresented (Fisher's tests, two sided, Bonferroni-corrected *P*-value $\leq$ 0.05), with the conserved category being the most skewed in tissue representation with a strong bias toward brain-specific expression (supplementary table S4, Supplementary Material online). The high tissue specificity (Tau score) of genes in ohnolog-tetrads associated with these

Table 2
Classification of Tissue Expression Divergence in the Ohnolog-Tetrads

Evolutionary Scenario	AORe	LORe
I: Ancestral ohnolog divergence followed by purifying selection independently in both species	199 (5.7%)	24 (4.7%)
II: Conserved tissue regulation of all ohnologs	869 (25%)	131 (25.7%)
III: Atlantic salmon-specific divergence of one ohnolog	375 (10.8%)	70 (13.8%)
IV: Grayling-specific divergence of one ohnolog	516 (14.8%)	80 (15.7%)
V: Conserved tissue regulation among ohnologs within species but different between species	195 (5.6%)	51 (10.0%)
Unclassified (VI): Ohnolog-tetrads assigned to three or more tissue clusters	1,326 (38.1%)	153 (30.1%)
Total	3,480	509

Note.—The number and percentages of genes in each category calculated based on the total number of topology-filtered ohnolog-tetrads.

genes (supplementary fig. S8, Supplementary Material online) corroborates the observed brain-specific expression bias.

Further, we tested whether distinct ohnolog-tetrad divergence categories were coupled to patterns of protein-coding and promoter sequence evolution. Specifically, we tested the hypothesis that conserved regulation is associated with conserved protein-coding evolution. We estimated the d*N*/d*S* ratios for each duplicate pair within each species and compared the distribution of d*N*/d*S* statistics in each class with that of the "unclassified" category VI (supplementary fig. S9, Supplementary Material online). Low d*N*/d*S* ($\ll$1) indicates strong purifying selection pressure. Categories I–V show variability in among-ohnolog d*N*/d*S* ratios, with category I having significantly higher d*N*/d*S* ratio compared with the "unclassified" category (Wilcoxon rank sum, $P = 0.005$) and categories II and V having significantly lower d*N*/d*S* ratios (Wilcoxon rank sum, $P = 0.014$ and $P = 0.0017$, respectively). The ohnolog pairs showing lineage-specific expression divergence (III and IV) did not have a significantly different d*N*/d*S* ratio compared with the unclassified category (Wilcoxon rank test, $P = 0.36$ and $P = 0.26$, respectively). Further, we used the Atlantic salmon genome to annotate and compare known transcription factor motifs divergence in the promoters (from 1,000 bp upstream to 200 bp downstream of the transcription start site) of ohnologs. Under the assumption that expression divergence is, at least partly, driven by changes in transcription factor-binding motifs located in proximal promoters, we tested whether ohnolog regulatory divergence in salmon (scenario I and III) was associated with divergence of promoter motifs. Indeed, the results add validation to the different expression evolution classifications (supplementary fig. S10, Supplementary Material online), with categories I and III having significantly less similar promoter motif content compared with ohnolog-tetrads with conserved tissue expression regulation in salmon (II, IV, and V) (Wilcoxon test all contrasts between I/III and II/IV/V, $P < 0.04$–0.002).

To evaluate whether the ohnologs in different classes were associated with distinct biological functions, we performed GO term enrichment tests. The ohnolog-tetrads of category II under strict selective constraints show highly brain-specific expression and are enriched for GO functions related to behavior and neural functions. In contrast, genes in category I, which represents ohnologs that underwent divergence in gene regulation following WGD, are associated with functions related to lipid metabolism, development, and immune system (supplementary file 2, Supplementary Material online).

Highly connected genes in protein–protein interaction (PPI) networks are often placed under strong constraints to maintain stoichiometry (Freeling and Thomas 2006; Sémon and Wolfe 2007). To test whether the strong constraints on the ohnolog-tetrads with conserved tissue expression (II) are associated with having higher PPIs, we extracted all the zebrafish genes from the gene trees corresponding to the ohnologs in the expression divergence categories and queried them against the STRING database (Szklarczyk et al. 2017) (version 10.5). Only associations with a score of above 7.0, suggesting high-confidence associations, were retained. As expected, we found that category II genes were indeed enriched for PPI (enrichment *P*-value = 1.05e-05) in comparison to the genes in the other classes (I, III, and IV) with diverged expression (enrichment *P*-value = 0.79).

Evolution of Gene Expression Levels Following WGD

The coexpression analyses leverage gene expression variation between tissues to classify ohnologs according to regulatory divergence. However, it is important to note that it does not explicitly test for significant changes in gene expression levels. To assess the divergence of ohnolog expression levels, we generated an RNA-Seq data set from additional liver samples from Atlantic salmon ($n = 4$) and grayling ($n = 4$). We tested for differences in liver expression levels between ohnolog pairs within both species and calculated absolute differences in fold change (FC) and the statistical significance (FDR-adjusted *P*-values) for these tests. Of the 2,467 ohnolog-tetrads in categories I–V (table 2), 54% (1,349) showed significant ($\mathrm{FDR} < 10^{-3}$) fold change differences ($\mathrm{FC} > 2$) in liver expression in at least one species, with 19% (467) in both species, 18% (455) in grayling only, and 17% (427) in salmon only.

We then focused on the subset of ohnolog-tetrads where at least one ohnolog was assigned to an expression cluster displaying dominant expression in liver. From this subset of

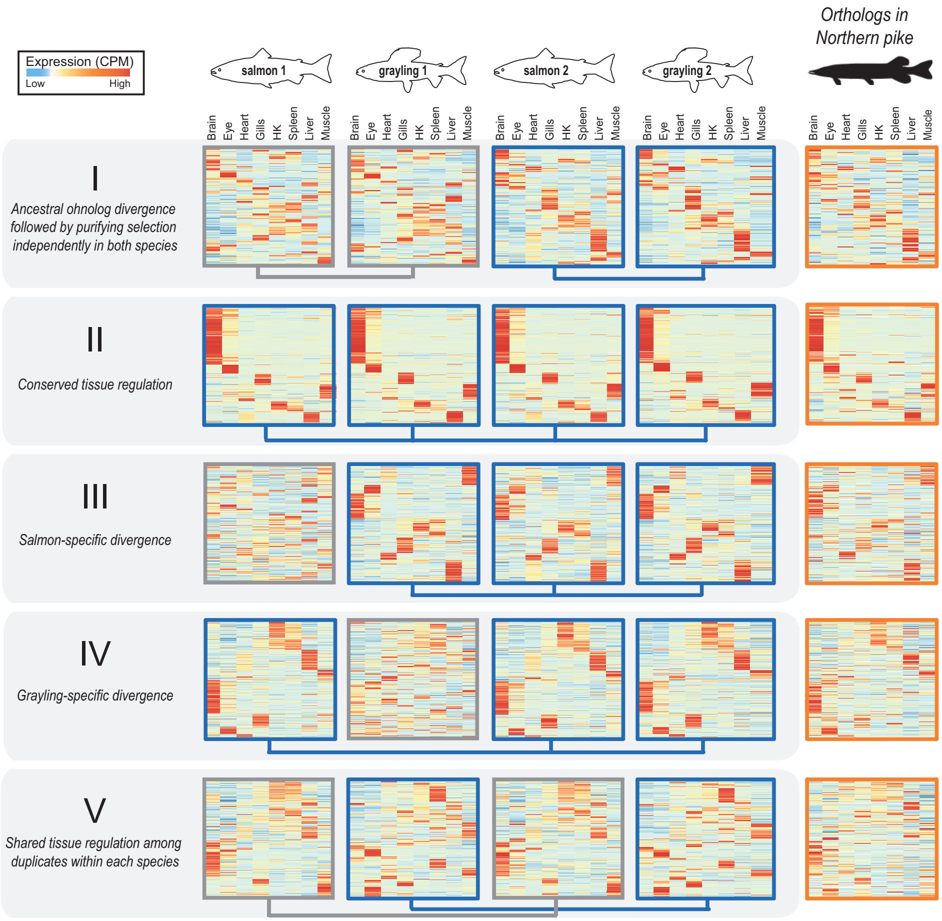


FIG. 3.—Selection on tissue expression regulation after whole-genome duplication. Heatmaps show clustering of tissue expression of the ohnolog-tetrads for each of the five evolutionary scenarios of tissue expression regulation following Ss4R WGD (see table 2). Within each category, the first four heatmaps represent one ohnolog-tetrad (i.e., four genes: salmon1, grayling1, salmon2, and grayling2) that were ordered based on similarity of expression profiles with the corresponding orthologs in northern pike (the fifth heatmap depicted on the right). Darker red corresponds to the highest expression level observed for one gene, and the darker blue to the lowest (scaled CPM). Connecting blue lines below the heatmaps indicate duplicates belonging to the same tissue clusters (conserved expression pattern). (An extended figure with ohnolog-tetrads subdivided into LORe and AORe in supplementary fig. S6, Supplementary Material online.)

552 ohnolog-tetrads, 80% (442) showed significant (FDR $< 10^{-3}$) fold change differences (FC > 2) in liver expression in at least one species, 37% (204) both species, 25% (136) grayling only, and 18% (102) salmon only (supplementary table S5, Supplementary Material online). As tissue-dominance is the main factor in the analyses of tissue-regulatory evolution, we expected that the different evolutionary scenarios (fig. 3) should be associated with enrichments in certain patterns of expression level divergence, or alternatively, a lack thereof. We indeed found that the ohnolog-tetrads reflecting ancestral divergence followed by purifying selection in both species (scenario I) were

significantly enriched in ohnologs being expressed at different levels in both species compared with other scenarios (fig. 4*a*, Fisher's test *P*-value = 1.74×10^{-4}). Conversely, those ohnolog pairs that show shared tissue regulatory patterns (scenarios II and V) were significantly enriched in ohnologs with no expression level divergence (*P*-values = 0.0117 and 6.79×10^{-3}). Finally, ohnologs with tissue regulatory divergence in one species (scenario III and IV in fig. 3) also had the most pronounced enrichment of expression level divergence for that species (fig. 4*a*, *P*-values = 3.32×10^{-7} and 2.03×10^{-6}). Three examples of putative liver-specific expression gains showing high correspondence between tissue regulation and expression level evolution are highlighted in figure 4*c*–*h*.

Further, we assessed whether liver-specific expression level differences between ohnologs were associated with changes in transcription factor-binding motif presence in promoters. We partitioned the ohnologs into three categories; differentially expressed in both species (likely diverged in expression in an ancestor of all salmonids), species-specific expression level divergence (we only used salmon-specific cases as we had no promoter motif data for grayling), and no significant difference in expression level. The lowest promoter motif similarity was found among ohnologs where both species showed strong expression level divergence, followed by the salmon-specific and then no expression divergence (fig. 4*b*).

Selection on Chloride Ion Transporter Regulation

The most apparent difference in biology between grayling and Atlantic salmon is the ability of Atlantic salmon to migrate between freshwater and saltwater (anadromy), a trait that grayling has not evolved. A key feature in saltwater acclimation involves remodeling gill physiology to enable efficient ion secretion and maintenance of osmotic homeostasis (Evans et al. 2005). We therefore specifically investigated the ohnolog divergence in gill gene expression regulation for key ion transport-associated genes that perform critical function in the process of chloride ion secretion in sea water (Mackie et al. 2007; Nilsen et al. 2007). The $Na^+/K^+/2Cl^-$ cotransporter 1 (*NKCC1a*) gene showed an extreme gill-dominated expression of one of the ohnologs in salmon, while no such gill-specific expression was observed for the corresponding grayling ohnologs (supplementary fig. S11, Supplementary Material online). A particularly striking difference in expression pattern was observed for the cystic fibrosis transmembrane conductance regulator gene (*CFTR*; an ABC transporter-class conducting chloride ion transport), which exhibited grayling-specific regulatory divergence (Category IV). From the tissue expression profiles of this tetrad (fig. 5*a*), it was evident that the divergence of tissue regulation in grayling was associated with a loss of gill tissue expression specificity compared with Atlantic salmon. To determine whether the grayling *CFTR* duplicate with diverged expression also had signatures of coding sequence divergence, we computed branch-specific d*N*/d*S*. Notably, the grayling *CFTR* displaying diverged expression regulation also displays a 2-fold increase in d*N*/d*S* compared with its Ss4R duplicate copy with conserved expression regulation (fig. 5*b*). Na^+/K^+-ATPase subunit genes were not well represented in the annotation and orthogroups and hence not included in the analysis.

Discussion

A major limitation in previous studies of evolution of gene regulation following WGD in vertebrates has been the inability to distinguish between neutral and adaptive divergence (Hermansen et al. 2016). Here, we leverage gene expression data from two salmonid species and a close outgroup species (northern pike) in a comparative approach to identify shared expression evolution patterns following WGD in lineages evolving independently for ~50 Myr. This allows us to identify evolutionarily long-term conservation of novel expression "phenotypes" arising after WGD—the hallmark of novel adaptive functions.

Although regulatory divergence of Ss4R ohnologs is widespread (Lien et al. 2016; Gillard et al. 2018), we show that ohnolog regulatory tissue divergence shared among species separated by ~50 Myr of evolution is comparably rare (table 2).

Ohnolog-tetrads that include genes with liver- and gill-biased expression contribute disproportionately to signals that are consistent with adaptive expression evolution in a salmonid ancestor (supplementary table S4, Supplementary Material online). The genes predominantly expressed in liver have been shown to have a strong association with phyletic age (Kryuchkova-Mostacci and Robinson-Rechavi 2015), while also being associated with particularly fast expression evolution when compared with other tissues (Duret and Mouchiroud 2000; Khaitovich et al. 2006). However, in contrast to our results, this latter pattern is associated with signatures of neutral evolution rather than adaptive evolution of novel regulation in mammals (Kryuchkova-Mostacci and Robinson-Rechavi 2015). Future studies should therefore look into the forces that drive regulatory evolution in liver-centric genes, using broader comparative data sets, detailed characterization of the evolutionary turnover of liver-specific regulatory elements, and use of phylogenetic methods that are able to detect shifts in selection on gene regulation (Sandve et al. 2018)

Salmonids are suggested to have evolved from a pike-like ancestor, a relatively stationary ambush predator (Craig 2008). Under this assumption, early salmonid evolution must have involved adaptation to new pelagic and/or riverine habitats. Adaptations to new environments and evolution of different life history strategies are known to be associated with strong selective pressure on immune-related genes (Star et al. 2011; Haase et al. 2013; Solbakken et al. 2017).

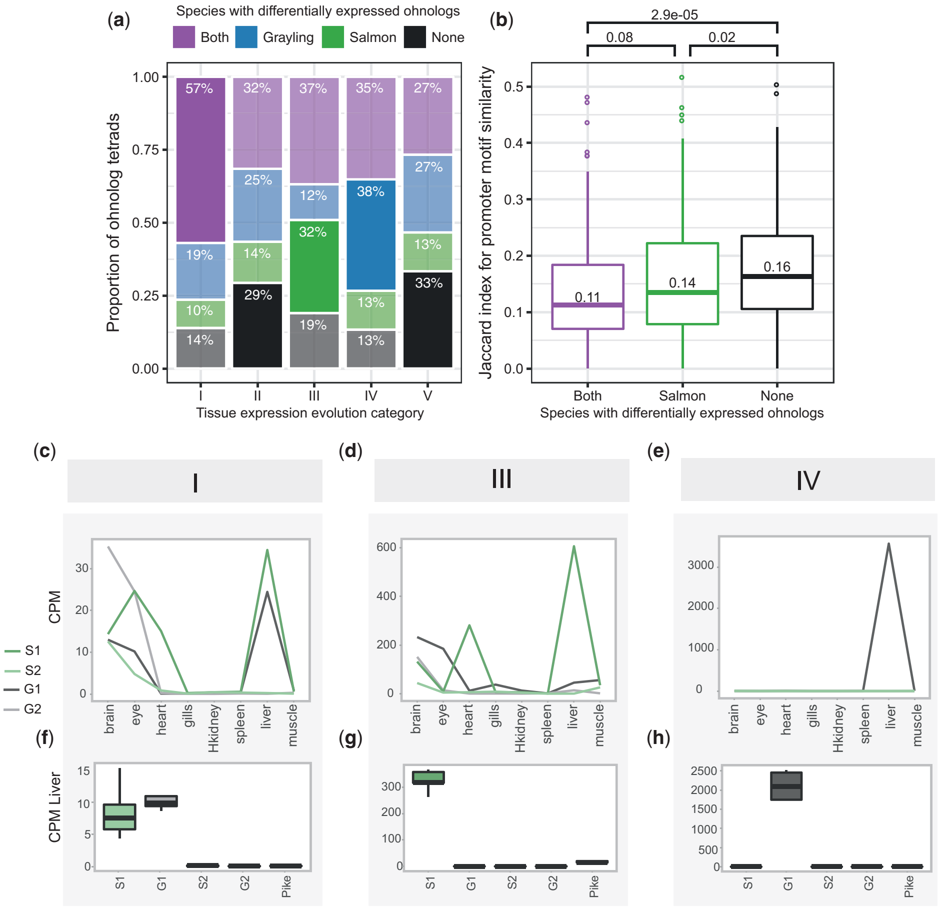


Fig. 4.—Expression level evolution of the ohnolog-tetrads. (*a*, *b*) Ohnolog-tetrads were tested using liver expression data for cases of highly significant differential expression (FDR-adjusted *P*-value $<10^{-3}$, absolute FC >2) between ohnologs of both species (purple), grayling ohnologs only (blue), Atlantic salmon ohnologs only (green), and no ohnologs of either species (black). Ohnolog tetrads shown in the figure had at least one ohnolog assigned to a tissue expression cluster displaying dominant expression in liver. The differential expression outcome expected to be the highest for each category is highlighted opaque, while the remaining outcomes are transparent. (*a*) The number and proportions of cases are given per tissue expression evolution category (see table 2). The differential expression outcome expected to be the highest for each category is highlighted opaque, while the rest are transparent. (*b*) Jaccard index score distributions for Atlantic salmon ohnolog promoter motif similarity, separated by differential expression outcome. *P*-values from pairwise comparisons testing for lower Jaccard index using the Wilcoxon test are indicated—as well as the median scores. The expression levels, in terms of CPM, from the tissue atlas data (*c*–*e*) and the corresponding data from the liver expression data are plotted (boxplots in *f*–*h*) for a selected example with a liver-specific gain of expression in each of the categories I, III, and IV. The examples indicated in (*c*–*h*) include ohnologs of ephrin type-B receptor 2-like (category I), contactin-1a-like gene (category III), and an E3 ubiquitin-protein ligase-like gene (category IV). The ohnologs in Atlantic salmon and grayling are represented as S1, S2 and G1 and G2, respectively.

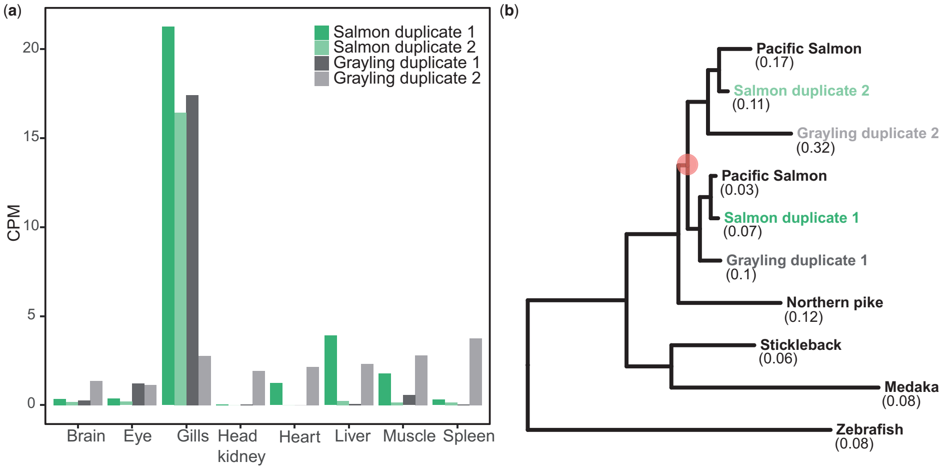


FIG. 5.—Divergent selection on cystic fibrosis transmembrane conductance regulator. (*a*) Expression values, in terms of CPM, of the *CFTR* ohnologs in Atlantic salmon and grayling across eight tissues. (*b*) *CFTR* gene tree. The orange circle represents the Ss4R duplication. Branch-specific d*N*/d*S* values of the tip nodes are given in parentheses.

In line with this, we see an overrepresentation of immune-related genes among ohnologs that diverged in the common ancestor of salmon and grayling but have been under purifying selection in both species after speciation (category I, table 2 and supplementary file 2, Supplementary Material online). Furthermore, pikes are generally piscivorous throughout their life span, while salmonids depend more on aquatic and terrestrial invertebrate prey with significantly lower input of lipids, especially in early life (Carmona-Antoñanzas et al. 2013). Interestingly, duplicates with shared ancestral divergence (category I), which are candidates for adaptive divergence in regulation, are enriched for genes involved in lipid-homeostasis, metabolism, and energy storage (glycogen)-related functions (GO test results in supplementary file 2, Supplementary Material online).

The regulatory divergence of metabolism-related genes and its association with corresponding shifts in the prey nutrient profile have been previously described in other fish (McGirr and Martin 2017). In this study, we do find individual candidate genes for tissue remodeling of metabolism function, such as the ATP-binding cassette transporter gene linked to cholesterol metabolism (*ABCA1*, supplementary fig. S12, Supplementary Material online). However, in order to interpret these results from a perspective of gene regulation evolution related to novel lifestyle and diet adaptations in salmonid ancestor, a comprehensive analysis using, for example, liver coexpression network comparisons and controlled experiments with dietary modifications would be necessary.

Taken together, our results suggest a role of Ss4R ohnologs in adaptive evolution of novel gene expression regulation, possibly related to new pathogenic pressures in a new type of habitat, and optimization of lipid-homeostasis and glycogen metabolism-related functions in response to evolution of a more active pelagic/riverine life with limited lipid resources.

Purifying selection to maintain ancestral tissue regulation of ohnologs in both salmonid species was the most commonly observed fate of ohnolog expression evolution (category II, table 2 and fig. 3). These ohnologs were predominantly brain-specific and enriched for predicted PPIs. Several other studies in vertebrates have found similar results, with strong purifying selection on sequence and expression evolution in brain-dominant genes, as well as high retention probability following large-scale genome duplication (Khaitovich et al. 2005; Chan et al. 2009; Zheng-Bradley et al. 2010; Guschanski et al. 2017; Roux et al. 2017). As neuron function-related genes are involved in complex networks of signaling cascades and higher-order PPIs, this pattern is believed to be driven by either direct selection for maintaining novel gene copies due to dosage balance (relative or absolute) or indirectly through selection against toxic effects of misfolding or misinteracting proteins (Roux et al. 2017).

Recent analyses of salmonid genomes have revealed ~25% LORe between Atlantic salmon and grayling (Lien et al. 2016; Robertson et al. 2017). Here, we find a set of LORe regions, corresponding to whole chromosome arms in Atlantic salmon, detected as single copy genes in grayling as a result of collapse during the assembly process. This strongly suggests that these sequences are in fact present as near-identical duplicated regions in the grayling genome. The larger chromosome arm-sized regions virtually indistinguishable at the sequence level (~10% in total, i.e., blue ribbons in fig. 2*b*) are likely still recombining or have only ceased to do so in the recent evolutionary past. Large-scale chromosomal rearrangements often follow genome duplication to block or hinder recombination among duplicated regions (Comai 2005; Lien et al. 2016). The difference we observe in the rediploidization history between Atlantic salmon and grayling is thus likely linked to the distinctly different chromosome evolution in these species (supplementary fig. S1, Supplementary Material online) (Qumsiyeh 1994).

The genomic footprints of LORe also extend to ohnolog regulatory divergence. LORe regions showed a strong enrichment of species-specific tissue-specific expression pattern (category V, in table 2 and supplementary table S5, Supplementary Material online), as expected under lineage-specific rediploidization and subsequent regulatory divergence. However, we also find a small proportion (~5%) of genes in AORe regions of the genome that reflect conserved tissue regulation among ohnologs within species but different regulation between species (category V). This observation is more difficult to explain, but it is likely real as the coding- and promoter sequence evolution analyses show that these ohnologs are biased toward high similarity in the coding and promoter sequences within each species (supplementary figs. S9 and S10, Supplementary Material online). Possible explanations for this observation could be a result of nonhomologous gene conversion (Hastings 2010) or evolution of species-specific tissue regulatory networks.

Finally, one fundamental difference between European grayling and Atlantic salmon is that Atlantic salmon has the ability to migrate between fresh- and seawater (anadromy). We demonstrate differences between European grayling and Atlantic salmon gill gene expression regulation for ohnologs of two key genes, *NKCC1* and *CFTR*, involved in chloride ion homeostasis (Marshall and Singer 2002; Nilsen et al. 2007). *NKCC1* is involved in entry of chloride ions into the basolateral membrane and is known to be regulated during migration to saltwater. Our finding that only Atlantic salmon displays a strong gill expression bias for one Ss4R copy of *NKCC1a* (supplementary fig. S11, Supplementary Material online) is congruent with adaptive evolution of novel gill expression regulation to ensure efficient ion transport in gills in anadromous salmonids.

As for the *CFTR* ohnologs, the results point toward an ancestral gill expression dominance in a salmonid ancestor, followed by a grayling lineage-specific loss of both tissue expression specificity and gill expression dominance, accompanied by increased accumulation of nonsynonymous substitutions (figs. 5*a* and *b*). Atlantic salmon on the other hand has retained both copies as "gill specific." The diverged expression of the *CFTR* gene copy in European grayling could be related to evolution of novel function, with the elevated d*N*/d*S* reflecting a mixture of positive selection on some codons and relaxed purifying selection on others. However, a more parsimonious model of *CFTR* evolution in European grayling would be that there has been a relaxation of purifying selection pressure to maintain both *CFTR* copies in the nonanadromous species. We thus propose that maintaining two functional *CFTR* genes could be an adaptive trait in anadromous salmonids to improve their ability to remove excess chloride ions and maintain ion homeostasis in the sea.

Conclusions

We present the first genome assembly of European grayling and use it for comparative studies with the reference genome assembly of Atlantic salmon. We show that this draft genome assembly is a highly valuable resource for gene-based analyses of salmonids and their relatives. Our comparative genome and transcriptome analysis between Atlantic salmon and grayling provides novel insights into evolutionary fates of ohnologs subsequent to WGD and associations between signatures of selection pressures on gene duplicate regulation and the evolution of salmonid traits, including anadromy. The key candidate genes potentially involved in differences in lifestyle, dietary adaptations, and anadromy between salmon and grayling should be further followed up in future evolutionary and experimental studies. Hence, the genome resource of grayling opens up new exciting avenues for utilizing salmonids as a model system to understand the evolutionary consequences of WGD in vertebrates.

Availability of Data

The Illumina reads have been deposited at ENA under the project accession: PRJEB21333. The genome assembly and annotation data are available at https://doi.org/10.6084/m9.figshare.c.3808162. Atlantic salmon liver expression data are available at ENA or NCBI under sample accessions: SAMEA104483623, SAMEA104483624, SAMEA104483627, and SAMEA104483628.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This work was supported by funding from University of Oslo to the SAK project "Building a marine genome hub" and the Strategic Research Initiative, Center for Computational Inference in Evolutionary Life Science (CELS) to K.S.J. S.R.S. was funded partly by the Norwegian Research Council (project NFR 244164). We thank Kim M. Bærum for sampling of grayling and Marianne H.S. Hansen for excellent technical assistance. Sample preparation, library construction, and sequencing were carried out at the Norwegian Sequencing Centre (NSC), Norway and McGill University, and Génome Québec Innovation Centre, Canada. The computational work was performed on the Abel Supercomputing Cluster (Norwegian Metacenter for High-Performance Computing [NOTUR] and the University of Oslo), operated by the Research Computing Services group at USIT, the University of Oslo IT-department. We thank Geir O. Storvik and Kjetil L. Voje for helpful discussions. We greatly appreciate Daniel J. Macqueen, Marine S. Brieuc, and Monica H. Solbakken for critical reading of the manuscript.

Author Contributions

K.S.J., L.A.V., S.J., and S.L. conceived and planned the project and generation of the data. S.R.S. and S.V. performed all the analyses with help from A.J.N. and O.K.T. Differential expression analysis on the liver data set was performed by G.B.G. and T.R.H., and the promoter motif analysis was prepared by T.D.M. S.R.S., S.V., and A.J.N. drafted the manuscript. All authors read and commented on the manuscript.

Literature Cited

Acharya D, Ghosh TC. 2016. Global analysis of human duplicated genes reveals the relative importance of whole-genome duplicates originated in the early vertebrate evolution. BMC Genomics 17:71.

Alexandrou MA, Swartz BA, Matzke NJ, Oakley TH. 2013. Genome duplication and multiple evolutionary origins of complex migratory behavior in Salmonidae. Mol Phylogenet Evol. 69(3):514–523.

Anders S, Pyl PT, Huber W. 2015. HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics 31(2):166–169.

Bailey TL, et al. 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 37(Web Server issue):W202–W208.

Berthelot C, et al. 2014. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nat Commun. 5:3657.

Carmona-Antoñanzas G, Tocher DR, Taggart JB, Leaver MJ. 2013. An evolutionary perspective on Elovl5 fatty acid elongase: comparison of Northern pike and duplicated paralogs from Atlantic salmon. BMC Evol Biol. 13:85.

Carroll SB. 2000. Endless forms: the evolution of gene regulation and morphological diversity. Cell 101(6):577–580.

Chan ET, et al. 2009. Conservation of core gene expression in vertebrate tissues. J Biol. 8(3):33.

Comai L. 2005. The advantages and disadvantages of being polyploid. Nat Rev Genet. 6(11):836–846.

Conant GC, Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet. 9(12):938–950.

Conesa A, et al. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21(18):3674–3676.

Craig JF. 2008. A short review of pike ecology. Hydrobiologia 601(1):5–16.

De Smet R, Sabaghian E, Li Z, Saeys Y, Van de Peer Y. 2017. Coordinated functional divergence of genes after genome duplication in *Arabidopsis thaliana*. Plant Cell 29(11):2786–2800.

Dobin A, et al. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29(1):15–21.

Duret L, Mouchiroud D. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol Biol Evol. 17(1):68–74.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16:157.

Evans DH, Piermarini PM, Choe KP. 2005. The multifunctional fish gill: dominant site of gas exchange, osmoregulation, acid-base regulation, and excretion of nitrogenous waste. Physiol Rev. 85(1):97–177.

Faust GG, Hall IM. 2012. YAHA: fast and flexible long-read alignment with optimal breakpoint detection. Bioinformatics 28(19):2417–2424.

Freeling M, Thomas BC. 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res. 16(7):805–814.

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. arXiv [q-Bio.GN]. Available from: http://arxiv.org/abs/1207.3907.

Gillard G, et al. 2018. Life-stage-associated remodelling of lipid metabolism regulation in Atlantic salmon. Mol Ecol. 27(5):1200–1213.

Gnerre S, et al. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A. 108(4):1513–1518.

Grabherr MG, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 29(7):644–652.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. Bioinformatics 27(7):1017–1018.

Gu X, Su Z. 2007. Tissue-driven hypothesis of genomic evolution and sequence-expression correlations. Proc Natl Acad Sci U S A. 104(8):2779–2784.

Guschanski K, Warnefors M, Kaessmann H. 2017. The evolution of duplicate gene expression in mammalian organs. Genome Res. 27(9):1461–1474.

Haas BJ, et al. 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 8(8):1494–1512.

Haase D, et al. 2013. Absence of major histocompatibility complex class II mediated immunity in pipefish, *Syngnathus typhle*: evidence from deep transcriptome sequencing. Biol Lett. 9(2):20130044.

Hartley SE. 1987. The chromosomes of salmonid fishes. Biol Rev Camb Philos Soc. 62(3):197–214.

Hastings PJ. 2010. Mechanisms of ectopic gene conversion. Genes 1(3):427–439.

Hendry AP, Stearns SC. 2004. Evolution illuminated: salmon and their relatives. Oxford University Press.

Hermansen RA, Hvidsten TR, Sandve SR, Liberles DA. 2016. Extracting functional trends from whole genome duplication events using comparative genomics. Biol Proced Online. 18:11.

Kassahn KS, Dang VT, Wilkins SJ, Perkins AC, Ragan MA. 2009. Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates. Genome Res. 19(8):1404–1418.

Katoh K, Misawa K, Kuma K-I, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30(14):3059–3066.

Kent WJ. 2002. BLAT—the BLAST-like alignment tool. Genome Res. 12(4):656–664.

Khaitovich P, Enard W, Lachmann M, Pääbo S. 2006. Evolution of primate gene expression. Nat Rev Genet. 7(9):693–702.

Khaitovich P, et al. 2005. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. Science 309(5742):1850–1854.

Korf I. 2004. Gene finding in novel genomes. BMC Bioinformatics 5:59.

Kryuchkova-Mostacci N, Robinson-Rechavi M. 2015. Tissue-specific evolution of protein coding genes in human and mouse. PLoS One 10(6):e0131673.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12:323.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv [q-Bio.GN]. Available from: http://arxiv.org/abs/1303.3997.

Li J-T, et al. 2015. The fate of recent duplicated genes following a fourth-round whole genome duplication in a tetraploid fish, common carp (*Cyprinus carpio*). Sci Rep. 5(1):8199.

Lien S, et al. 2016. The Atlantic salmon genome provides insights into rediploidization. Nature 533(7602):200–205.

Limborg MT, Seeb LW, Seeb JE. 2016. Sorting duplicated loci disentangles complexities of polyploid genomes masked by genotyping by sequencing. Mol Ecol. 25(10):2117–2129.

Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 33(20):6494–6506.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science 290(5494):1151–1155.

Mackie PM, Gharbi K, Ballantyne JS, McCormick SD, Wright PA. 2007. $Na^+/K^+/2Cl^-$ cotransporter and *CFTR* gill expression after seawater transfer in smolts (0^+) of different Atlantic salmon (*Salmo salar*) families. Aquaculture 272(1–4):625–635.

Macqueen DJ, Johnston IA. 2014. A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. Proc Biol Sci. 281(1778):20132881.

Marshall WS, Singer TD. 2002. Cystic fibrosis transmembrane conductance regulator in teleost fish. Biochim Biophys Acta Biomembr. 1566(1–2):16–27.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 17(1):10.

McGirr JA, Martin CH. 2017. Parallel evolution of gene expression between trophic specialists despite divergent genotypes and morphologies. Evol Lett. 2:62–75.

Nilsen TO, et al. 2007. Differential expression of gill Na^+, K^+ -ATPase α- and β-subunits, Na^+,K^+, 2Cl- cotransporter and *CFTR* anion channel in juvenile anadromous and landlocked Atlantic salmon *Salmo salar*. J Exp Biol. 210(16):2885–2896.

Nygren A, Nilsson B, Jahnke M. 1971. Cytological studies in *Thymallus thymallus* and *Coregonus albula*. Hereditas 67(2):269–274.

Ocalewicz K, et al. 2013. Pericentromeric location of the telomeric DNA sequences on the European grayling chromosomes. Genetica 141(10–12):409–416.

Ohno S. 1970. Evolution by gene duplication. New York: Springer-Verlag.

Osborn TC, et al. 2003. Understanding mechanisms of novel gene expression in polyploids. Trends Genet. 19(3):141–147.

Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23(9):1061–1067.

Phillips R, Ráb P. 2001. Chromosome evolution in the Salmonidae (*Pisces*): an update. Biol Rev Camb Philos Soc. 76(1):1–25.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2-approximately maximum-likelihood trees for large alignments. PLoS One 5(3):e9490.

Quevillon E, et al. 2005. InterProScan: protein domains identifier. Nucleic Acids Res. 33(Web Server issue):W116–W120.

Qumsiyeh MB. 1994. Evolution of number and morphology of mammalian chromosomes. J Hered. 85(6):455–465.

Robertson FM, et al. 2017. Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification. Genome Biol. 18(1):111.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1):139–140.

Roux J, Liu J, Robinson-Rechavi M. 2017. Selective constraints on coding sequences of nervous system genes are a major determinant of duplicate gene retention in vertebrates. Mol Biol Evol. 34(11):2773–2791.

Sandve SR, Rohlfs RV, Hvidsten TR. 2018. Subfunctionalization versus neofunctionalization after whole-genome duplication. Nat Genet. 50(7):908–909.

Sémon M, Wolfe KH. 2007. Consequences of genome duplication. Curr Opin Genet Dev. 17(6):505–512.

Sémon M, Wolfe KH. 2008. Prefrential subfunctionalization of slow-evolving genes after allopolyploidization in *Xenopus laevis*. Proc Natl Acad Sci U S A. 105(24):8333–8338.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31(19):3210–3212.

Solbakken MH, Voje KL, Jakobsen KS, Jentoft S. 2017. Linking species habitat and past palaeoclimatic events to evolution of the teleost innate immune system. Proc Biol Sci. 284(1853):20162810.

Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 24(5):637–644.

Star B, et al. 2011. The genome sequence of Atlantic cod reveals a unique immune system. Nature 477(7363):207–211.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34(Web Server issue):W609–W612.

Szklarczyk D, et al. 2017. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res. 45(D1):D362–D368.

Trapnell C, et al. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 28(5):511–515.

UniProt Consortium. 2015. UniProt: a hub for protein information. Nucleic Acids Res. 43(D1):D204–D212.

Van de Peer Y, Maere S, Meyer A. 2009. The evolutionary significance of ancient genome duplications. Nat Rev Genet. 10(10):725–732.

Van der Auwera GA, et al. 2013. From FastQ data to high-confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 43: 11.10.1–11.10.33.

Walker BJ, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9(11):e112963.

Wilkinson M, McInerney JO, Hirt RP, Foster PG, Embley TM. 2007. Of clades and clans: terms for phylogenetic relationships in unrooted trees. Trends Ecol Evol. 22(3):114–115.

Wolfe KH. 2001. Yesterday's polyploids and the mystery of diploidization. Nat Rev Genet. 2(5):333–341.

Wray GA, et al. 2003. The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol. 20(9):1377–1419.

Yanai I, et al. 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics 21(5):650–659.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13(5):555–556.

Zhang J. 2003. Evolution by gene duplication: an update. Trends Ecol Evol. 18(6):292–298.

Zheng-Bradley X, Rung J, Parkinson H, Brazma A. 2010. Large scale comparison of global gene expression patterns in human and mouse. Genome Biol. 11(12):R124.

Associate editor: Yves Van De Peer

Paper III

Gene regulatory evolution following salmonid whole genome duplication

Gareth B. Gillard[1], Rori V. Rohlfs[2], Ben F. Koop[3], Eric B. Rondeau[3], Simen R. Sandve[4*], and Torgeir R. Hvidsten[1*]

1 Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Norway.

2 Department of Biology, San Francisco State University, USA.

3 Department of Biology, University of Victoria, Canada.

4 Center for Integrative Genetics, Department of Animal and Aquacultural Sciences, Faculty of Biosciences, Norwegian University of Life Sciences, Norway.

*Corresponding authors: simen.sandve@nmbu.no, torgeir.r.hvidsten@nmbu.no

Introduction

Whole genome duplications have played a major role in increasing genomic complexity and fueling the eukaryotic lineages with novel genetic material. Such sudden doubling of genetic material provides a rare opportunity for evolution to shape novel gene functions out of a vast space of gene redundancy, eventually resulting in evolution of novel phenotypes and species (Ohno 1970; Van de Peer et al. 2017). In vertebrates, ancient WGDs ancestral to all vertebrates ~500 million years ago (Mya) and at the base of teleosts ~350 Mya (Hoegg et al. 2004; Amores et al. 2011) have been associated with the evolution of novel traits leading to genetic and phenotypic diversity and speciation (Holland et al. 1994; Meyer and Van de Peer 2005; Van de Peer et al. 2009; Van de Peer et al. 2017; Lynch and Conery 2000; Sémon and Wolfe 2007; Volff 2005). However, the contribution of duplicated genes arising from WGDs to the evolution of novel phenotypes, both at the molecular and organismal level, remains poorly understood.

The impact of vertebrate WGD on gene expression regulation has received considerable attention in recent years (Lien et al. 2016; Varadharajan et al. 2018; Berthelot et al. 2014; Braasch et al. 2016; Robertson et al. 2017). These studies have revealed widespread divergence in tissue regulation after WGD, but also that the majority of duplicate pairs evolve in an asymmetric manner - with one copy retaining an ancestral like regulation, while the other evolves a novel regulatory phenotype (Sandve et al. 2018). This pattern is in accordance with the classical model made famous by Ohno in the 1970s (Ohno 1970), whereby one duplicate copy is released from selective constraints, followed by accumulation of novel 'functions' which can give rise to novel adaptive phenotypes. Unfortunately, due to methodological limitations and limitations in available transcriptomics datasets, preceding analyses of regulatory evolution following WGD has not been able to distinguish between neutral and likely adaptive shifts in gene expression regulation (Sandve et al. 2018).

A statistical framework is needed to distinguish neutral and adaptive shifts in gene expression. Ornstein-Uhlenbeck processes have been proposed to model gene expression evolution, enabling robust hypothesis testing (Bedford and Hartl 2009). An OU processes is an extension of a brownian motion random walk process, with an additional assumption that a stabilizing force pulls the process back towards a particular value. This model framework has been shown to be suitable for modelling evolution of gene expression, which is subject to stabilizing selection, and has been used to detect selection on gene expression across mammalian evolution (Bedford and Hartl 2009; Kalinka et al. 2010; Perry et al. 2012; Chen et al. 2019). Here, we use a recently developed OU-based model of transcriptome evolution (Rohlfs and Nielsen 2015; Rohlfs et al. 2014) to test the hypothesis that WGD sparks novel adaptive shifts in gene expression regulation. Using the salmonid WGD 80-100 Mya (Berthelot et al. 2014; Macqueen and Johnston 2014) as a model system, our results supports that WGD results in a major shift in selective constraints and drive evolution of novel gene expression levels in liver. Interestingly, the majority of expression level shifts represents 'down tuning' of one or both duplicates, while only a small proportion displayed the hallmarks of a novel and likely adaptive expression level increase. This suggests that gene dosage stoichiometry is a strong selective force in early evolution following WGD in vertebrates.

Results

Identifying gene orthologs after whole genome duplication

We compared expression evolution in seven species, including three non-salmonid teleosts (*Danio rerio*/zebrafish, *Oryzias latipes*/medaka and *Esox lucius*/northern pike) and four salmonids (*Salmo salar*/Atlantic salmon, *Salvelinus alpinus*/Arctic char, *Oncorhynchus kisutch/*coho salmon and *Oncorhynchus mykiss*/rainbow trout) (Figure 1A).

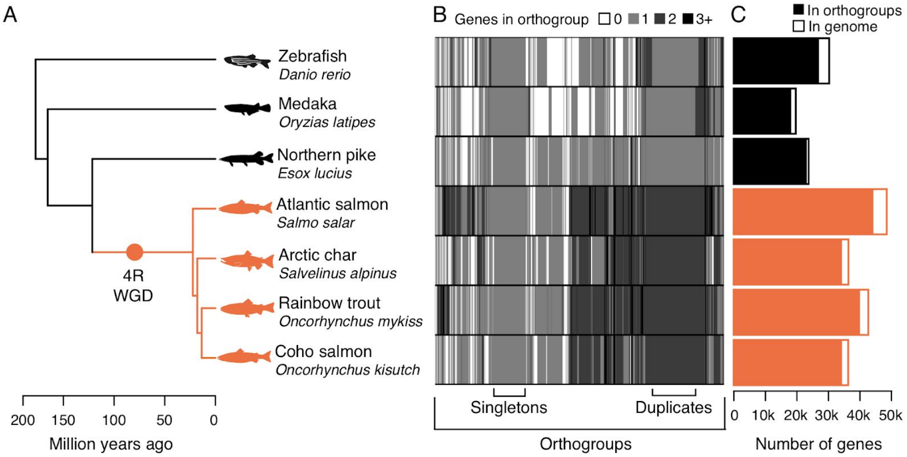


Figure 1. Orthologs identified in studied species *(A) Evolutionary timeline of the species used in the test for gene regulatory evolution (Rabosky et al. 2018). The estimated time of the salmonid (4R) WGD is shown. (B) A heatmap showing the number of genes per species in each orthogroup. Singleton orthogroups have one gene copy in all species, and duplicate orthogroups have one gene copy in the outgroup species and two gene copies in salmonid species. (C) The number of protein coding genes in the genome and in orthogroups per species.*

To this end, we identified 20,734 gene ortholog groups (orthogroups) that comprised the majority of the genes from each species (Figure 1C). By clustering the orthogroups by the number of genes from each species, we identified 2,232 (11%) singleton orthogroups that contained only one gene copy from each species (referred to as singletons), and 2,887 (14%) duplicate orthogroups that contained only one gene copy from the non-salmonid species and two gene copies from the salmonid species (Figure 1B).

Normalizing expression for comparative transcriptomics

We supplemented existing RNA-Seq data from liver samples (four replicates) in Atlantic salmon and northern pike (Gillard et al. 2018) with corresponding data for zebrafish, medaka, Arctic char, coho salmon, and rainbow trout. Reads were mapped to the respective genomes and gene expression were estimated as Transcripts Per Million (TPM). However, comparative analyses across species with widely different number of genes cannot be based on TPM values. A naive expectation is that this will lead to lower gene expression values in species with higher numbers of genes, since the total number of reads is divided among more genes. However, this expectation is further complicated by the fact that the studied species also vary in the fraction of genes that are expressed in liver and the distribution of genes expressed at different levels (Figure 2A). To allow for comparison of gene expression across species, we therefore devised an approach where we first normalized expression data within each species (Figure 2C, between replicates) and then between species (Figure 2E). Between species normalization was done by computing scaling factors based on the expression of singletons, which we assumed to be expressed at similar levels across species (see methods for details). Species normalization factors differed quite substantially from the naive gene number expectation (Figure 2D), but within species normalization had by far the largest effect on expression values.

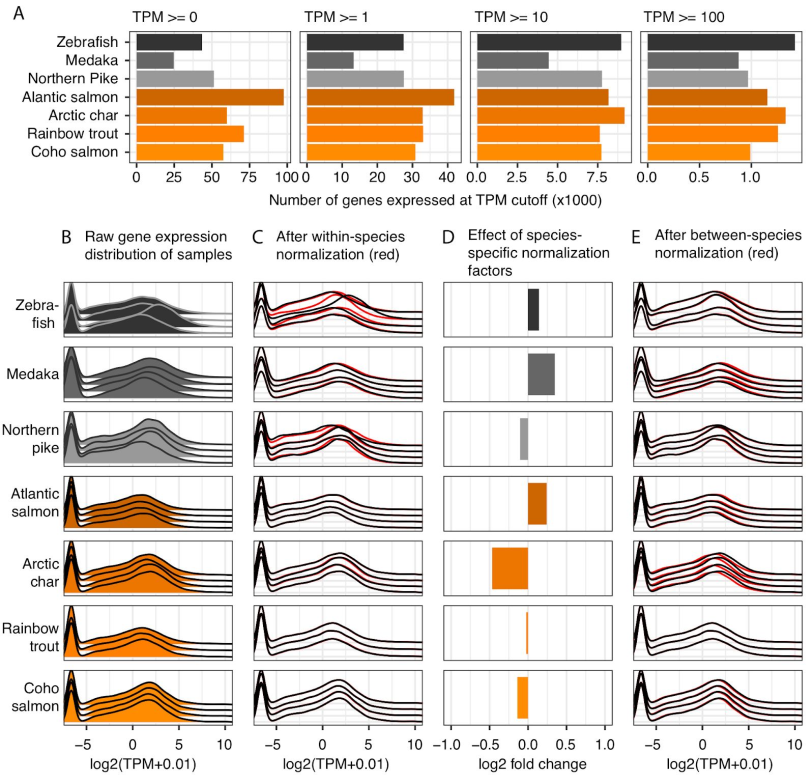


Figure 2: Within and between species expression normalization. *(A) For each species, the number of genes that are expressed at different expression levels measured in Transcripts Per Million (TPM). (B) The distribution of unnormalized gene expression values measured as log2 (TPM+0.01), for each replicate and species. (C) The distribution of gene expression values before (grey lines) and after (red lines) within-species normalization (WSN). (D) The between-species normalization factors multiplied to within-species normalized expression values (log2-scale). (E) The distribution of gene expression before (grey lines) and after (red lines) between-species normalization (BSN) using the normalization factors.*

Detecting shifts in expression levels following WGD

The Ornstein-Uhlenbeck (OU) process has previously been used to model the changes in gene expression over time (Chen et al. 2019). The OU process is a stochastic process that models accumulation of random changes in expression over time (random walk), but unlike a Brownian motion process, an OU process assumes that expression level evolution is constrained within biologically relevant bounds (Figure 3A). Due to stabilizing selection, variation in expression increases non-linearly over time. In the OU process formulation, the θ parameter represents the optimal level that expression varies around. The σ^2 parameter represents the rate of variation over time, while the α parameter represents the strength of selection pressure towards an optimal level (θ), restricting the variation (Figure 3A). To assess if expression variation in the orthologs from our studied species fits the OU process assumptions, we plotted the mean squared expression distance between pairs of species against their evolutionary distance (in sequence substitutions). Expression distance levels off as evolutionary distance increases (Figure 3B, Supplementary figure 1) in agreement with the OU process assumptions (rather than a Brownian motion process) (Bedford and Hartl 2009), demonstrating that an OU-model is appropriate for studying expression evolution in our species.

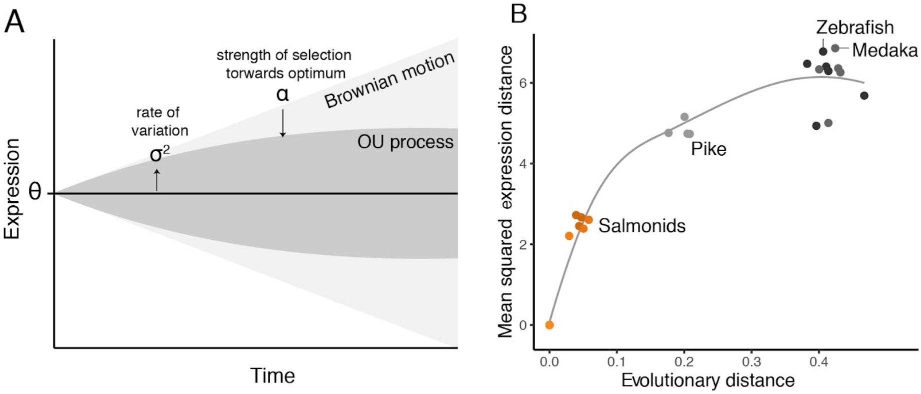


Figure 3: Modelling gene expression divergence over time. *(A) Diagram of the Ornstein-Uhlenbeck (OU) process compared to a Brownian motion-based model. Parameters in the OU model include an optimal expression level θ, the rate of variation σ^2, and the strength of selection α towards the optimal level. (B) The mean squared expression distance is plotted against the evolutionary distance (sequence substitutions) for each pair of species. The points are coloured and labeled based on the oldest species in the pair.*

To test for statistically significant shifts in expression after the 4R WGD, we used the Expression Variance and Evolution (EVE) model (Rohlfs and Nielsen 2015; Rohlfs et al. 2014), which builds upon the OU process by adding a new parameter β that represents the ratio of within-species expression variance over between-species variance. The number of species (i.e. tips in gene trees) correlate with the statistical power to detect expression level shifts. This makes it difficult to directly compare the expression shift test statistics between retained 4R copies and genes that have returned to singleton copies. To overcome this obstacle, we subsetted each duplicate orthogroup, with retained 4R duplicates, so that each test was only done on one of the two monophyletic 'duplicate' clades (Figure 4). The outgroup orthologs (i.e. genes in the species without the 4R WGD) remain the same for both subsets. For each set of orthogroups, the EVE model parameters were optimised to best fit the expression data and maximum likelihood values were calculated. Our null hypothesis (H_{null}) is that expression has not shifted after WGD from a single optimal level ($\theta_{\text{ ancestral}}$) and that expression differences is explained by species and evolutionary variance under the model of consistent stabilizing selection. Our alternative hypothesis (H_{alt}) is that expression has shifted for salmonid orthologs after WGD from a pre-4R WGD optimal level ($\theta_{\text{ ancestral}}$) to a new post-4R WGD optimal level ($\theta_{\text{ derived}}$). We then perform a Likelihood Ratio Test (LRT) and reject H_{null} ($\theta_{\text{ ancestral}} = \theta_{\text{ derived}}$) and accept H_{alt} ($\theta_{\text{ ancestral}} \neq \theta_{\text{ derived}}$) if the LRT score is greater than the upper 95% quantile of the χ^2 distribution with one degree of freedom (Figure 4). Orthologs with a significant shift in expression were further classified into those with a shift up and those with a shift down in expression compared to the pre-WGD level. The direction of the shift was determined by comparing the difference between $\theta_{\text{ derived}}$ and $\theta_{\text{ ancestral}}$. As controls, we used a corresponding setup to tested for shifts from pre-4R WGD levels in singleton orthogroups, and to test for shifts in individual outgroup species in singleton orthogroups.

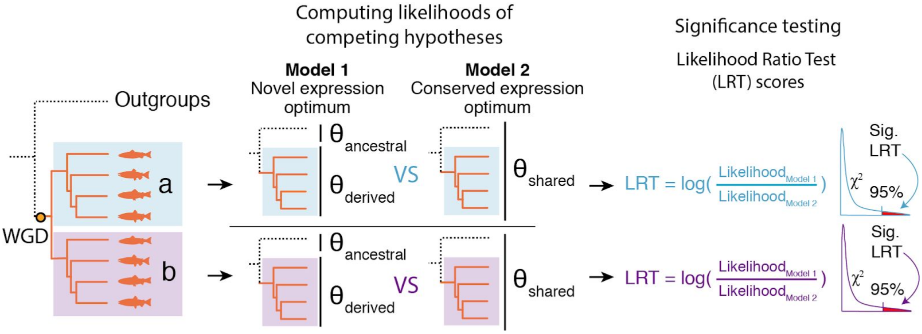

Computing likelihoods of competing hypotheses
Significance testing
Likelihood Ratio Test (LRT) scores
Model 1
Novel expression optimum
Model 2
Conserved expression optimum
Outgroups
a
b
WGD
θ ancestral
θ derived
VS
θ shared
LRT = log(Likelihood Model 1 / Likelihood Model 2)
Sig. LRT
χ²
95%

Figure 4: Design for detecting significant shifts in gene expression after the 4R WGD*. For each gene duplicate (a and b), we compare two models. Under model 1, the expression optimum of the duplicate ($\theta_{\text{ derived}}$) is shifted compared to the ancestral expression optimum ($\theta_{\text{ ancestral}}$), while under model 2 there is not shift in expression. A Likelihood Ratio Test is then used to compare the two models and the statistical significance is determined using the* χ^2 distribution with one degree of freedom.

The EVE analysis revealed proportionately more salmonid duplicates with a shift in expression level optimum compared to salmonid singletons or teleost outgroup singletons (Figure 5A-B). Twentysix percent of orthogroups with retained Ss4R duplicates have a significant shift in expression level optimum for at least one of the copies. In comparison, the same proportion was 16% for salmonid singletons and 6-10% for non-salmonid singletons. The large majority (81%) of duplicates with a significant expression level shift displayed decreased derived expression levels. In comparison, the same numbers were only 38% and 40-60% of salmonid singletons and outgroup lineages, respectively (Figure 5A). Using independent Atlantic salmon RNA-Seq expression data from a large feeding trial (Gillard et al. 2018) and a tissue atlas (Lien et al. 2016) confirmed shifts in expression consistent with what we detect in this study (Figure 6).

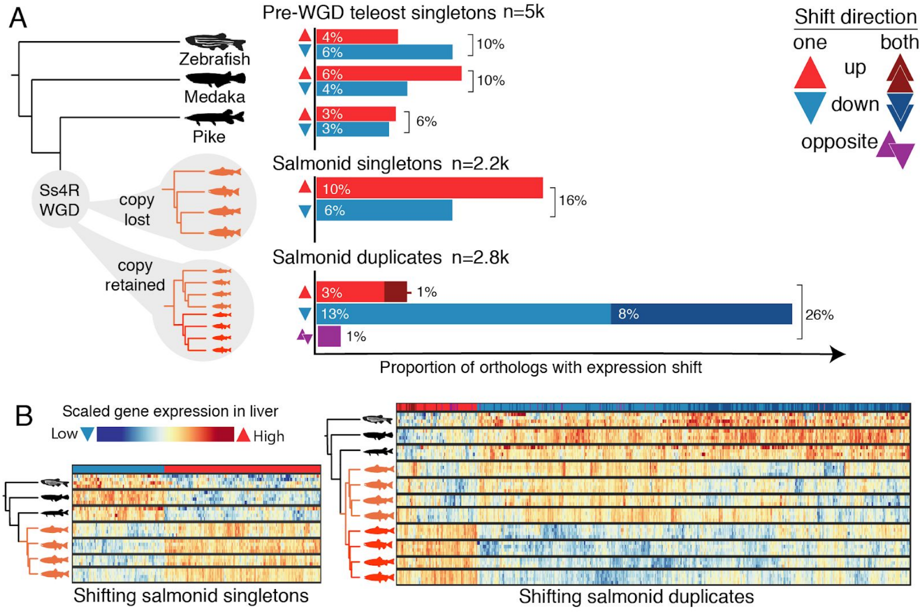


Figure 5: Gene expression shifts after WGD. *(A) The proportions of duplicate, singleton, or outgroup singletons that were detected to have a significant shift up (red, up-triangle) or down (blue, down-triangle) in expression. Salmonid duplicates can have one (lighter colors) or both copies (darker colors) shifted in expression, or have one copy shifted up and one down (purple, up-down-triangles) (B) Heatmaps of expression levels of salmonid singleton and duplicates with a significant shift in expression. Gene expression is column-scaled (gene-scaled) with the highest values in red and the lowest values in blue.*

Both the rate at which genes evolves new expression levels (Berthelot et al. 2018) and the power to detect gene expression shifts is associated with gene expression levels. If duplicate and singleton genes have vastly different average expression levels, this could explain the different propensity to evolve increased or decreased expression. To account for expression level differences, we compared the proportions of salmonid duplicates and singletons genes (both with and without significant expression shifts) with different expression levels (divided into quartiles) (Supplementary figure 2). The results fits well with the general idea that we have little power to detect decrease in expression levels for already lowly expressed genes, and that the power to detect any shift increase with expression level. More importantly, this analyses also clearly demonstrates that the distinct differences between duplicates and singletons in how their expression levels evolve (Figure 5A) is not due to systematic biases in expression levels.

In half of cases where duplicates had shifted in expression optimum, one of the duplicates had a shift down in expression level. This pattern could be explained by one duplicate being redundant and evolving under relaxed selection pressure - i.e. on the path towards pseudogenization. To test this idea we analysed the difference in co-expression network centrality of duplicates with different expression evolution patterns. Our expectation was that pseudogenized genes have become decoupled from their ancestral liver gene regulatory network and would display decreased centrality compared to their duplicate counterpart under selective constraints. Indeed, using a co-expression network inferred from RNA-Seq data from 209 Atlantic salmon liver samples (Gillard et al. 2018), the duplicate with a significantly decreased expression level also had lower network centrality (Figure 6A).

Under a model of pseudogenization we also would expect that loss of co-expression network centrality would hold if co-expression networks are built using tissue expression data. However, the analyses using RNA-seq data from 15 tissues from Atlantic salmon showed quite the opposite result. Duplicated genes that have evolved lower expression levels in liver instead become more integrated into the tissue co-expression network (Figure 6) and more tissue specific (Supplementary figure 3A).

Visual inspection of tissue expression profiles of these genes provide a more nuanced picture, revealing that most genes with derived lower expression in liver show similar expression profiles across most tissues (Supplementary Figure 3A). This pattern is completely reversed for duplicates with a derived increase in expression level optimum in liver; as expected these are more integrated into the liver co-expression network, but perhaps unexpectedly, less integrated in the tissue network (Figure 6) with a broader tissue expression (lower tissue specificity, Supplementary figure 3B). This was corroborated by visual inspection of the clustered tissue expression profiles (Supplementary Figure 3) which suggests that that only a small minority of duplicates with increased expression levels have a strong tissue expression bias towards liver. Finally, the change in centrality is unlikely to be explained by expression levels only, as there was low correlation between expression levels and co-expression network centrality in both liver (r^2=0.053) and tissue atlas networks (r^2=0.0018).

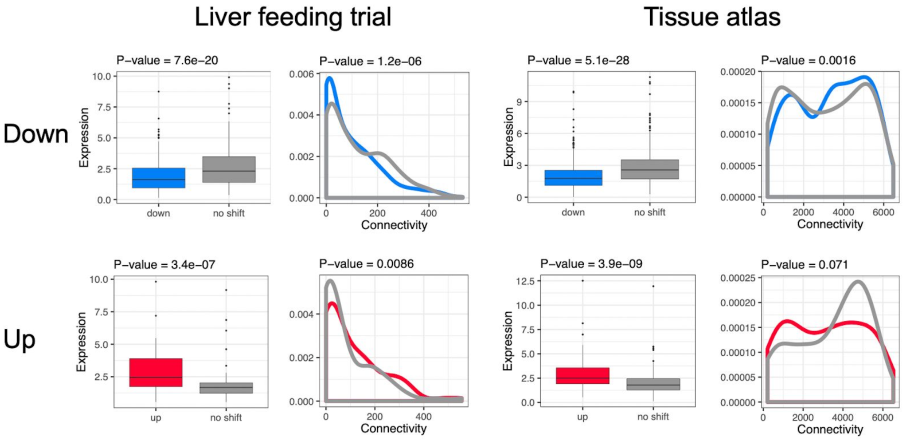


Figure 6: Co-expression network analysis of gene duplicated with shifted expression. *Analysis of orthogroups with one copy having a significant down/up shift in expression compared to the ancestral expression level (rows) in two independent data sets (columns). Box plots compare the expression in liver of the two copies, one shifted (up/down) and one remaining at ancestral levels (no shift). For the liver feeding trial, the average expression levels across all samples were used while for the tissue atlas the expression levels in the liver sample were used. Line plots shows the distribution of network centrality scores (network degree, equivalent to the number of neighbors in the network but computed in a weighted network) for the copy shifted up/down and for the copy remaining at ancestral levels.*

Divergent expression levels between duplicated genes can evolve through changes in cis-regulatory elements. Tests for enrichment of transcription factor (TF) binding sites motifs identified 71 and 43 significantly enriched ($p < 0.0001$) motifs associated with evolution of lower expression and higher expression levels, respectively, following WGD (Supplementary tables 1 and 2, https://salmobase.org/apps/SalMotifDB/). Among the top enriched motifs associated with increased liver gene expression we find both general transcription factor TFIIIA, and more liver specific motifs such as POU3F2 and GFI1B.

Functional enrichment of genes with similar expression evolution

To assess if duplicates with a significant shift in expression were associated with specific functions we performed KEGG pathway and Gene Ontology (GO) term enrichment tests (Supplementary tables 3 and 4). Duplicates shifted down, as well as singletons with shifts up or down, were primarily associated with housekeeping roles such as ribosome- and spliceosome-associated activities and metabolic processes. Duplicates with a shift up in expression level were enriched in functions relevant to lipid- and fatty acid metabolism-related functions. Three of these genes were elongase genes, elovl1, elovl5 and elovl6 (Figure 7).

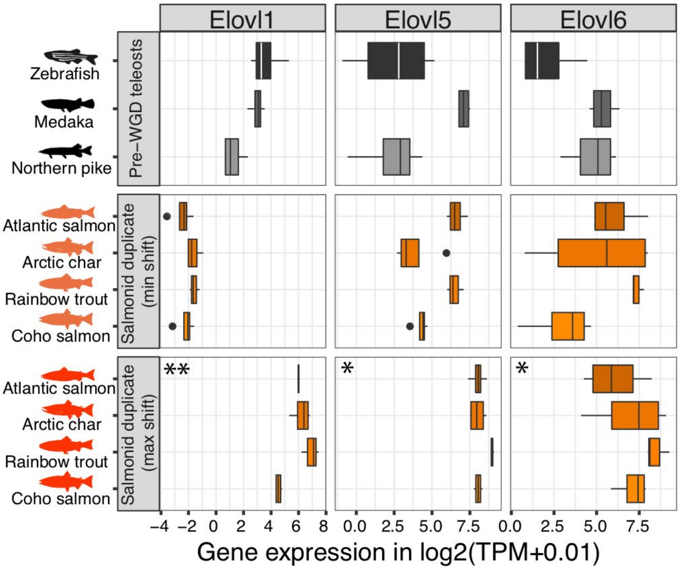


Figure 7: Expression levels of three fatty acid elongation genes; elovl1, elovl5, and elovl6, with an expression shift in one salmonid duplicate copy. Significance of shift is shown for p-value < 0.05 (*) and < 0.01 (**).

Discussion

Adaptive shifts in liver gene expression following WGD

In this paper we attempt to understand the evolutionary consequences of WGD on adaptive evolution of gene expression levels in liver. We use an OU-process to model gene expression evolution over a phylogeny, and then test two competing hypotheses: (1) WGD does result in a shift in expression level in a salmonid ancestor versus (2) there is no-shift in expression level. Although significant shifts in expression levels under this OU-process model is assumed to be adaptive, it is inherently difficult to exclude neutral evolution. Hence, in this study we have restricted our analyses to expression shifts that arose between the WGD and the divergence between anadromous salmonid species 20 million years ago. Since we are only focussing on these phylogenetically conserved patterns across all salmonid species we are more confident that we are identifying true adaptive expression evolution. However, it is also possible that this has biassed our results towards genes with specific functions.

Our analyses of gene expression evolution following WGD in salmonids reveal that putative adaptive down tuning of expression levels of one or both 4R duplicates is most common. This pattern could be consistent with strong selection for stoichiometric balance for genes involved in protein-protein interactions (Freeling and Thomas 2006) but could also be explained by rapid pseudogenization. Our co-expression network analyses of 4R duplicates did indeed show that 4R duplicates evolving lower expression levels also lost co-expression network centrality in liver (Figure 6), which would be expected under a pseudogenization model. However, only 21% of these 4R down tuned duplicates had consistently lower gene expression across all tissues, and moreover, they showed increased tissue specificity and increased tissue co-expression network centrality compared to duplicates with conserved expression levels (Figure 6, Supplementary figure 3). In line with the stoichiometric balance model, we see KEGG and GO-term enrichments (Supplementary table 3 and 4) for translation and splicing processes, which typically is performed by large interacting multiprotein complexes.

Another functional group highly enriched in the set of down tuned genes is genes associated with energy production (Supplementary table 3 and 4). This association could be linked to selection against oxidative stress from increased production of reactive oxygen species (Schoenfelder and Fox 2015). Taken together our results go against that pseudogenization is the major driver of evolutionary down tuning of gene duplicate expression levels. Instead it seems likely that strong selection on some form of gene product dosage balance has been important post 4R WGD for genes in liver.

Putative adaptive increase in gene expression levels were rare (4%) in 4R duplicates, with only a few KEGG-pathways and GO-terms showing strong enrichment (Supplementary table 3 and 4). The most enriched KEGG-pathways and GO-terms were linked to lipid metabolism, specifically three genes associated to elongation of fatty acids (Figure 7). This result is in accordance with previous findings that 4R duplicates have been targets for selection on lipid metabolism related gene expression divergence following the divergence of the salmonid lineage (Carmona-Antoñanzas et al. 2013; Gillard et al. 2018). Given the highly similar function of the three elongase genes it is plausible that they are under regulatory control of similar TFs that regulate lipid metabolism (Carmona-Antoñanzas et al. 2016). However, we did not find any enrichment of typical lipid metabolism TFs in these fatty acid elongase promoters (data not shown), nor in the entire set of 4R duplicates that had evolved increased liver expression levels (Supplementary table 1).

Methodological considerations

Several methods have been proposed to study duplicate expression divergence after WGD, and can be divided into methods that compare duplicates based on (a) expression levels (´on-off´), (b) differential expression (DE), (c) correlation and (d) co-expression clusters (Hermansen et al. 2016). Early attempts at applying these methods to understand expression evolution after WGDs used data from single species, while more recent studies utilized outgroups without the WGD as a proxy for the ancestral pre-WGD expression state (Braasch et al. 2016; Lien et al. 2016; Varadharajan et al. 2018). A general problem with all these methods have been that they lack a formal statistical framework to test specific hypotheses of adaptive versus neutral evolutions, and to analyse expression data from several species in a phylogenetic context (Sandve et al. 2018). Here we apply the Expression Variance and Evolution (EVE) model (Rohlfs and Nielsen 2015; Rohlfs et al. 2014) that uses the Ornstein-Uhlenbeck (OU) process to model expression level evolution in a phylogeny. Although the OU-process has previously been used to detect selection on gene expression levels (Bedford and Hartl 2009; Kalinka et al. 2010; Perry et al. 2012; Chen et al. 2019), the application of this framework to gene expression evolution after WGD present several challenges discussed below.

Normalization becomes extremely important when comparing expression levels across species with and without a recent WGD (Figure 1B). However, somewhat surprisingly, we found that gene numbers were not the main determinant for varying expression distributions across species as the number of expressed genes, and the number of genes expressed at different levels, also varied (Figure 2A). Faced with this complex picture, we decided to calculate normalisation factors based on singletons, with the assumption that these genes will have maintained similar expression levels across species. This normalization procedure resulted in similar expression distributions across species (Figure 2E) and the resulting normalized expression values were used as the basis for all downstream analysis.

Several factors can confound the statistical power of the EVE test, including the number of species used in the test and the expression level of the tested genes. When testing for significant expression shifts in duplicate orthogroups, we tested each duplicate clade against the outgroup species separately by removing its counterpart duplicate clade (Figure 4). Using the same number of species in the tests allow us to perform a fair comparison of the fraction of significant shifts in duplicate and singleton groups (Figure 5). We also compared these results to the number of shifts in single outgroup species, but here we were unable to design a test that completely rules out differences in the statistical power due to different numbers of species in the contrasted groups. Also, tests based on single species are much more vulnerable to expression differences associated with experimental conditions that, despite our attempts, cannot be controlled completely for species separated by 200 million years of evolution.

Expression levels can also affect the statistical power of our tests. Obviously, low expression in the pre-4R WGD ancestral state leaves little room for lowering the expression even further, but expression level also influence the statistical power to detect shifts up. Indeed, no significant shifts in either direction were observed for the orthogroups with ancestral expression levels in the lower quartile (Supplementary figure 2). However, although expression level clearly influence the statistical power of the EVE tests, the influence is similar for duplicate and singleton orthogroups and as such does not confound our conclusion that WGD spark adaptive expression evolution.

The current implementation of EVE requires orthogroups to be complete with exactly one, or in the case of duplicate groups, two genes from each species. This greatly limits the number of orthogroups we could test for adaptive shifts in expression (Figure 1B). Indeed, grayling and Danube salmon, with especially fragmented genome assemblies, was excluded from this study because they would have limited the number of complete orthogroups even further. This stresses the importance of high quality genome assemblies but also highlights the need for further method development. A method that can handle incomplete orthogroups will be needed to extend the scope of this type of analysis to, simultaneously, more genes and species. Finally it would be interesting to apply phylogenetic methods such as EVE to study other aspects of expression evolution such as shifts in expression profiles across developmental gradients, stress responses or different tissues. It is currently an open question how such data could be encoded to fit the current statistical framework, however one possible approach is to test for stage/time/tissue specificity using e.g. the tau statistic (Supplementary figure 3). Here we show that expression levels in liver meet the assumptions of an OU process (Figure 3B), however, whether this is true for other expression-derived measures remains to be seen.

Methods

Ortholog classification

Protein sequences were obtained from thirteen species in total for ortholog detection. Nine species were from the teleost lineage; *Danio rerio* (zebrafish), *Oryzias latipes* (medaka), *Gasterosteus aculeatus* (three-spined stickleback), *Esox lucius* (Northern pike), including five salmonid species; *Thymallus thymallus* (grayling), *Hucho hucho* (Danube salmon), *Salmo salar* (Atlantic salmon), *Salvelinus alpinus* (Arctic char), *Oncorhynchus mykiss* (rainbow trout), *Oncorhynchus kisutch* (coho salmon). One non-teleost fish; *Lepisosteus oculatus* (Spotted gar), and two mammals; *Homo sapiens* (human) and *Mus musculus* (house mouse), were outgroup species for the teleosts. Human, mouse, zebrafish, medaka and stickleback proteins fasta files were obtained from ENSEMBL (release 89). Proteins were obtained from NCBI RefSeq assemblies for species; Atlantic salmon (GCF_000233375.1), rainbow trout (GCF_002163495.1), spotted gar (GCF_000242695.1), coho salmon (GCF_002021735.1), and northern pike (GCF_000721915.3). Grayling proteins were obtained from its genome paper (Varadharajan et al. 2018). Genes from all species were assigned to gene ortholog groups (orthogroups) based on the sequence similarity of the single longest protein per gene, using the Orthofinder (v0.2.8) protocol (Emms and Kelly 2015). Within each orthogroup, protein sequences were aligned using MAFFT (v7.130) (Katoh et al. 2002) and maximum likelihood trees were estimated from this alignment using FastTree (v2.1.8) (Price et al. 2010).

We pruned complex ortholog gene trees containing duplication nodes older than the deepest species split by identifying monophyletic unrooted “clans” in ortholog trees that contained more than three salmonid/pike tips. Identification of clans and extraction of “clan trees” used in further analyses were done with an algorithm implemented in R available from github (github.com/srsand/Phylogenomics/blob/master/clanfinder.R). With the resulting clan orthogroups, the CDS sequence of each protein in the group was retrieved (NCBI/ENSEMBL) and orthogroup CDS sequence alignments and trees were generated. The clan orthogroups and their CDS trees were used in the following analysis. For expression analysis, species without replicated liver expression data was dropped from the clan orthogroups and tree tips.

RNA-sequencing of liver tissues

Liver tissue samples were collected for zebrafish, medaka, pike, rainbow trout, Arctic char, and coho salmon. Samples were taken in replicates of four, or three in the case of rainbow trout. Each fish was raised under standard rearing conditions for the species. Total RNA was extracted from the liver samples using the RNeasy Plus Universal Kit (QIAGEN). Quality was determined on a 2100 Bioanalyzer using the RNA 6000 Nano Kit (Agilent). Concentration was determined using a Nanodrop 8000 spectrophotometer (Thermo Scientific). cDNA libraries were prepared using the TruSeq Stranded mRNA HT Sample Prep Kit (Illumina). Library mean length was determined by running on a 2100 Bioanalyzer using the DNA 1000 Kit (Agilent) and library concentration was determined with the Qbit BR Kit (Thermo Scientific). Single-end sequencing of sample libraries was completed on an Illumina HiSeq 2500 with 100-bp reads. For Atlantic salmon, RNA-Seq data was obtained from a feeding trial using the samples from individuals in freshwater fed a marine based diet (Gillard et al. 2018).

Generating expression data

To generate gene expression data, RNA-Seq reads were mapped to the reference genomes (previously mentioned), with gene annotations, of their respective species using the STAR aligner with default settings (Dobin et al. 2013). RSEM (Li and Dewey 2011) was used to generate estimated values of gene read counts and Transcripts Per Million reads (TPM) that are normalized for gene average transcript length and the total number of reads from the sample.

Normalization, within and between species

The TMM method for count normalization, from the R package edgeR, was used to generate normalization factors to normalize gene expression data (Robinson et al. 2010). First the replicate samples from the same species were normalized between each other. Then, to account for potential expression differences between species, species specific normalization factors were calculated using mean expression values from only singleton orthogroups (single gene from each species), and replicates from each species were normalized by their respective species normalization factor. All expression values were log transformed (log2(TPM+0.01)) prior to testing for expression shifts.

Testing for shifts in gene expression

The EVE program (Rohlfs and Nielsen 2015) was used to test for shifts in gene expression levels on the salmonid gene branches for singleton and duplicate orthogroups. For singleton and each of the split duplicate trees, the salmonid gene expression was compared with the ancestral expression from pre-4R WGD outgroup species. The null hypothesis is that gene expression has not changed on the salmonid branch, while the alternative hypothesis is that expression of salmonid genes has diverged towards a new expression optimum after the WGD. EVE was given the replicated expression data for all species, and the species consensus tree produced by OrthoFinder. For every ortholog, a likelihood ratio test (LRT) score is calculated, representing the likelihood of the alternative hypothesis over the null hypothesis. LRT scores were compared to a Chi squared distribution with one degree of freedom and scores above the 95% quantile were considered to be significant. EVE reports theta estimates representing the expression optimum for the salmonid branch and the rest of the tree. The difference between salmonid theta estimate and non-salmonid theta estimate provided the magnitude and direction of the expression shift (details in Results section).

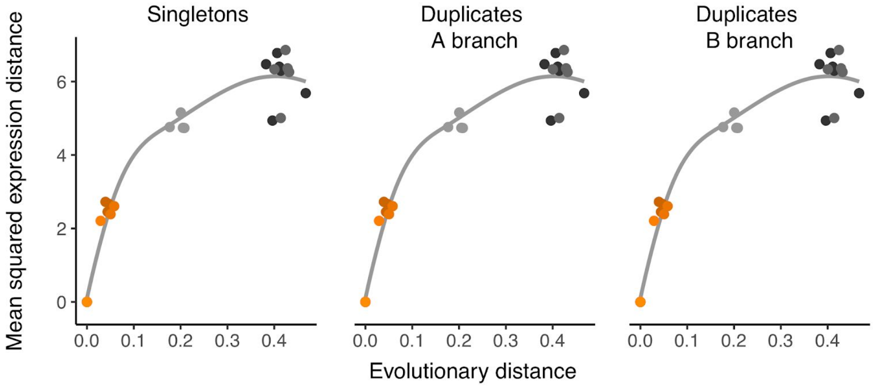
Co-expression network analysis

Gene expression data from a liver feeding trial (Gillard et al. 2018) and a tissue atlas (Lien et al. 2016) was analyzed using the *Weighted correlation network analysis (WGCNA)*-package in R (Langfelder and Horvath 2008). The soft-thresholds were determined using the *pickSoftThreshold*-function with parameters *corFnc = "bicor"* and *networkType = "signed"* and resulted in an R^2 scale free topology model fit of 0.8 at the soft-thresholding powers of 14 and 10, respectively. Modules and network connectivity were identified using the functions *blockwiseModules* and *intramodularConnectivity* with the aforementioned parameters.

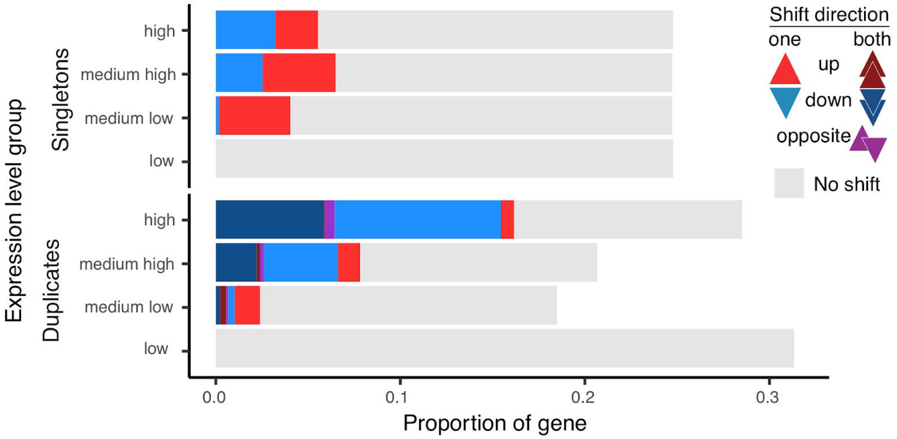
Data availability

For more information on analysis methods, availability of scripts used to generate the results in this paper, and data availability, see our git project for this paper: https://gitlab.com/garethgillard/gene-regulatory-evolution-following-salmonid-whole-genome-duplication

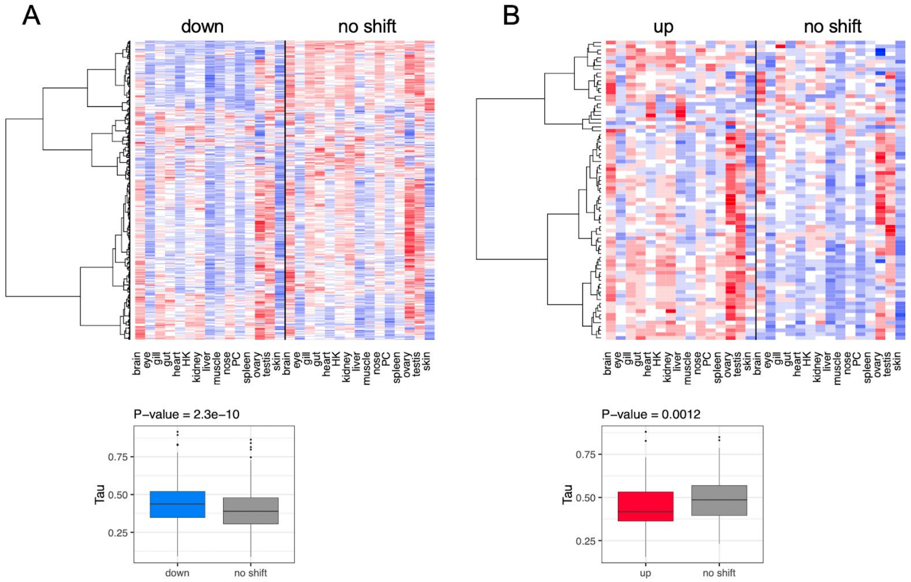
Supplementary material



Supplementary figure 1: OU process trends for different sets of orthogroups. *For the remaining sets of orthogroups tested for expression shifts (singtons and duplicate A and B copies), the mean squared expression distance is plotted against the evolutionary distance (sequence substitutions) for each pair of species. The points are coloured and labeled based on the oldest species in the pair.*



Supplementary figure 2: Shifts in singletons and duplicates for different expression levels. *The expression level groups were determined by the quartiles of the average expression of non-salmonid singletons. 'low': lower quartile, 'medium low': between lower quartile and median, 'medium high': between median and upper quartile, and 'high': upper quartile. The proportions of shifts per quartile is shown for salmonid singletons and duplicate, with the direction of the shift coloured.*



Supplementary figure 3. *Heatmap of the tissue expression of orthogroups with one copy having a significant down/up shift in expression compared to the ancestral expression level. Differences in tissue specificity (using the tau statistic) are also shown using box plots.*

Supplementary table 1: Transcription factor (TF) binding motifs enriched in duplicate orthogroups with one copy shifted up. *The table shows TF motifs where the promoters (from 1000bps upstream to 200bps downstream of transcription start sites) of shifted copies were enriched for the motif (p-value < 1x10^{-4}) and the copy without a shift did not have enrichment for any similar motif (i.e. any motifs in the SalMotifDB motif cluster, p-value > 0.05).*

TF	**Motif DB**	**TF cluster**	**P-value Shift up**	**P-value No shift**
BRN-2	TRANSFAC	POU3F2	2.56E-07	0.11
TFIIIA	TRANSFAC	TFIIIA	1.45E-06	0.62
IRF1	JASPAR	IRF-1	1.62E-06	0.34
SNAP190	TRANSFAC	SNAP190	3.31E-06	0.13
GFI1B	TRANSFAC	GFI1B	6.88E-06	0.20
AFP1	TRANSFAC	AFP1	7.50E-06	0.08
TRANSCRIPTION_FACTOR_MAFB	3D-footprint	TRANSCRIPTION_FACTOR_MAFB	1.34E-05	0.07
ZFP	TRANSFAC	ZFP	1.76E-05	0.14
CHECKPOINT_SUPPRESSOR_HOMOLOGUE	TRANSFAC	CHECKPOINT_SUPPRESSOR_HOMOLOGUE	8.83E-05	0.15
ZNF652	TRANSFAC	ZNF652	9.65E-05	0.13
NR2E3	JASPAR	NR2E3	1.14E-04	0.17
ZNF774	TRANSFAC	ZNF774	1.26E-04	0.21
BAB1	DrosophilaTF	BAB1	1.35E-04	0.33
ZNF69	TRANSFAC	ZNF69	1.94E-04	0.37
MAB-3	TRANSFAC	MAB-3	1.95E-04	0.11
HOMEOBOX_PROTEIN_MEIS1	3D-footprint	FOXL1	2.11E-04	0.56
IRF4	HUMAN.H10MO.C	BCL11A	2.48E-04	0.12
THRA	HumanTF	THRA	2.67E-04	0.14
ZNF236	TRANSFAC	ZNF236	3.16E-04	0.09
FORKHEAD_BOX_PROTEIN_O1	3D-footprint	FORKHEAD_BOX_PROTEIN_O1	3.27E-04	0.07
ZNF621	TRANSFAC	ZNF621	3.58E-04	0.28
E2F1_ELK1	HumanTF2	E2F1_ELK1	3.58E-04	0.28
FIZ1	TRANSFAC	FIZ1	3.73E-04	0.17
ZFP14	TRANSFAC	FORKHEAD_BOX_PROTEIN_M1	3.83E-04	0.12
ZNF878	TRANSFAC	ZNF878	3.90E-04	0.06
EVI-1	TRANSFAC	EVI-1	3.99E-04	0.10
ZNF655_E327G_R1	UniPROBE	ZNF655_REF_R2	4.01E-04	0.19
ZFX	CISBP	ZFP711	5.11E-04	0.06
CTCFL	TRANSFAC	CTCFL	5.17E-04	0.08
ZNF302	TRANSFAC	ZNF181	5.23E-04	0.30
GLIS2	TRANSFAC	GLIH1	5.33E-04	0.12
ENSTRUG00000003586	CISBP	GCM2	5.46E-04	0.14
HOXA3	CISBP	HOXA3	5.90E-04	0.23
CBF_(CORE_BINDING_FACTOR)	TRANSFAC	RUNX2	5.91E-04	0.21
ZBTB45	HT-SELEX2	DPF1	7.45E-04	0.12
CG4854	FlyZincFinger	CG4854	7.68E-04	0.05
RFX3	TRANSFAC	RFX3	7.70E-04	0.38
OCT-1	TRANSFAC	POU2F2	7.71E-04	0.09
CG14860	TRANSFAC	CG14860	7.84E-04	0.99
HOXB2_PITX1	HumanTF2	HOXB2_PITX1	8.62E-04	0.26
FOXN4	TRANSFAC	FOXN4	8.93E-04	0.19
GCM1_FOXO1	HumanTF2	GCM1_FOXI1	9.05E-04	0.06
ZNF197	TRANSFAC	ZNF197	9.89E-04	0.11

Supplementary table 2: Transcription factor (TF) binding motifs enriched in duplicate orthogroups with one copy shifted down. *The table shows TF motifs where the promoters (from 1000bps upstream to 200bps downstream of transcription start sites) of shifted copies were enriched for the motif (p-value < $1x10^{-4}$) and the copy without a shift did not have enrichment for any similar motif (i.e. any motifs in the SalMotifDB motif cluster, p-value > 0.05).*

TF	Motif DB	TF cluster	P-value Shift down	P-value No shift
ZFP553	TRANSFAC	ZFP553	9.66E-07	0.20
NFIL3	CISBP	NFIL3	1.08E-06	0.46
ZNF169	TRANSFAC	ZNF169	4.77E-06	0.36
ZNF160	TRANSFAC	ZNF160	5.08E-06	0.40
CUX1_PITX1	HumanTF2	CUX1_PITX1	5.35E-06	0.36
ZNF497	TRANSFAC	ZNF497	8.69E-06	0.25
FORKHEAD_BOX_PROTEIN_O1	3D-footprint	FORKHEAD_BOX_PROTEIN_O1	1.08E-05	0.26
CHECKPOINT_SUPPRESSOR_HOMOLOGUE	TRANSFAC	CHECKPOINT_SUPPRESSOR_HOMOLOGUE	1.49E-05	0.10
CT53	TRANSFAC	CT53	1.65E-05	0.09
ZBTB38	TRANSFAC	ZBTB38	2.09E-05	0.23
ZNF325	TRANSFAC	ZNF325	2.48E-05	0.29
ZNF540	TRANSFAC	ZNF540	3.54E-05	0.14
GCM	TRANSFAC	GCM	3.60E-05	0.07
TEAD4_CEBPB	HumanTF2	TEAD4_CEBPB	4.08E-05	0.24
ZNF845	TRANSFAC	ZNF845	5.97E-05	0.18
ZNF157	TRANSFAC	ZNF157	6.07E-05	0.24
HOXB13_TBX3	HumanTF2	HOXB13_TBX3	7.43E-05	0.07
ZNF37A	TRANSFAC	ZNF37A	7.88E-05	0.27
PSF	TRANSFAC	PSF	8.52E-05	0.10
ZNF846	TRANSFAC	ZNF846	9.12E-05	0.09
NHR-1	CISBP	NHR-1	1.04E-04	0.13
ZNF425	TRANSFAC	ZNF425	1.25E-04	0.06
SIGNAL_TRANSDUCER_AND_ACTIVATOR_OF_TRANSCRIPTION_1	3D-footprint	SIGNAL_TRANSDUCER_AND_ACTIVATOR_OF_TRANSCRIPTION_1	1.26E-04	0.33
ZFX	TRANSFAC	ZFY	1.27E-04	0.47
CG3919	CISBP	CG3919	1.34E-04	0.68
CG1856	FlyZincFinger	CG1856	1.45E-04	0.19
C/EBP	TRANSFAC	C/EBP	1.63E-04	0.08
E2F-4	TRANSFAC	E2F4	1.66E-04	0.18
ZNF713	TRANSFAC	ZNF713	1.85E-04	0.38
CG33557	CISBP	CG33557	1.94E-04	0.09
ZNF228	TRANSFAC	ZNF228	2.03E-04	0.16
ZNF514	TRANSFAC	ZNF514	2.16E-04	0.07
ZNF26	TRANSFAC	ZNF26	2.34E-04	0.14
CG12155	CISBP	CG12155	2.43E-04	0.14
HES7	TRANSFAC	HES7	2.81E-04	0.83
PEG3	TRANSFAC	PEG3	2.85E-04	0.60
ZNF324	TRANSFAC	ZNF324	2.86E-04	0.23
ZBTB11	TRANSFAC	ZNF257	2.92E-04	0.07
ELK1	HumanTF	ELK1	2.98E-04	0.06
ZNF426	TRANSFAC	ZNF426	3.11E-04	0.41
ETV2_ONECUT2	HumanTF2	ETV2_ONECUT2	3.37E-04	0.17
GATA-1	TRANSFAC	GATA-1	3.38E-04	0.98
SPDEF	HumanTF	SPDEF	3.44E-04	0.07
KLF4	JASPAR	KLF4	3.64E-04	0.17
T27F2.4	CISBP	T27F2.4	3.70E-04	0.21
STAF	TRANSFAC	STAF	4.00E-04	0.09

TBX1	HumanTF	TBX1	4.04E-04	0.06
SCRT2	TRANSFAC	SCRT2	4.30E-04	0.18
KLF1_R328H_R2	UniPROBE	KLF1_R328H_R2	4.46E-04	0.07
ZNF816A	TRANSFAC	ZNF816A	4.63E-04	0.05
ZNF560	TRANSFAC	ZNF560	4.83E-04	0.56
SNPC-4	JASPAR	SNPC-4	4.90E-04	0.05
ZFP1	SMILE-seq	ZFP1	5.45E-04	0.10
EGR2_REF_R1	UniPROBE	EGR2_REF_R1	5.84E-04	0.29
MAX	TRANSFAC	MAX	5.84E-04	0.43
ZNF668	TRANSFAC	ZFP668	5.86E-04	0.31
PRRXL1	TRANSFAC	OTX1	5.95E-04	0.10
ZNF132	TRANSFAC	ZNF132	5.97E-04	0.77
EVE	DrosophilaTF	EVE	6.53E-04	0.34
CEBPG_ELF1	HumanTF2	CEBPG_ELF1	7.39E-04	0.05
ZNF716	TRANSFAC	ZNF716	7.58E-04	0.06
ZNF41	TRANSFAC	ZNF41	8.06E-04	0.07
ZFP64	TRANSFAC	ZFP64	8.08E-04	0.14
ZNF579	TRANSFAC	ZNF579	8.31E-04	0.26
GCM1_MAX	HumanTF2	GCM1_MAX	8.49E-04	0.14
SMC-3	TRANSFAC	SMC-3	8.52E-04	0.16
DLIP3	TRANSFAC	DLIP3	8.60E-04	0.32
B0310.2	CISBP	B0310.2	8.93E-04	0.46
POU6F1	TRANSFAC	POU6F1	9.55E-04	0.19
VITAMIN_D3_RECEPTOR	3D-footprint	VITAMIN_D3_RECEPTOR	9.56E-04	0.25
FOXP3	TRANSFAC	FOXP3	9.99E-04	0.20

Supplementary table 3: KEGG pathway enrichment in shifted genes. *Genes were divided into groups according to whether they come from singleton and duplicate orthogroups and whether they were shifted up or down in expression level. For significantly enriched KEGG pathways, the total number of genes belonging to this pathway (Total), the number of these that had a shift in expression (Shifted) and the corresponding p-value (P-value) is shown.*

Ortholog	**Shift direction**	**KEGG Pathway**	**Total**	**Shifted**	**P-value**
Singletons	Shift up	Ubiquinone and other terpenoid-quinone biosynthesis	3	3	8.48E-04
		Selenocompound metabolism	5	3	7.33E-03
		RIG-I-like receptor signaling pathway	8	3	3.31E-02
		Spliceosome	14	4	3.71E-02
		Necroptosis	15	4	4.69E-02
		One carbon pool by folate	4	2	4.74E-02
		Non-homologous end-joining	4	2	4.74E-02
	Shift down	RNA degradation	20	5	4.15E-03
		Salmonella infection	5	2	2.83E-02
		Ribosome	23	4	3.73E-02
		Lysosome	24	4	4.28E-02
Duplicates	Shift up	Biosynthesis of unsaturated fatty acids	4	4	2.75E-06
		Fatty acid elongation	6	4	3.87E-05
		Fatty acid metabolism	9	4	2.95E-04
		Glycine, serine and threonine metabolism	5	2	1.55E-02
		Nucleotide excision repair	7	2	3.09E-02
		Glutathione metabolism	8	2	4.01E-02
		DNA replication	8	2	4.01E-02
		Neomycin, kanamycin and gentamicin biosynthesis	1	1	4.12E-02
	Shift down	Ribosome	20	18	1.26E-10
		Oxidative phosphorylation	22	17	4.18E-08
		Protein processing in endoplasmic reticulum	42	21	4.73E-05
		mRNA surveillance pathway	21	13	8.02E-05
		Protein export	9	7	5.45E-04
		RNA transport	37	17	8.83E-04

		Spliceosome	28	14	8.95E-04
		Metabolic pathways	202	62	1.47E-03
		RNA degradation	17	9	4.72E-03
		Phagosome	17	9	4.72E-03
		Regulation of actin cytoskeleton	30	13	6.59E-03
		Mitophagy - animal	18	9	7.64E-03
		Herpes simplex virus 1 infection	28	12	9.92E-03
		NOD-like receptor signaling pathway	26	11	1.49E-02
		Arginine and proline metabolism	8	5	1.51E-02
		Terpenoid backbone biosynthesis	6	4	2.31E-02
		Ribosome biogenesis in eukaryotes	15	7	2.83E-02
		Lysosome	22	9	3.38E-02
		Aminoacyl-tRNA biosynthesis	4	3	3.47E-02
		Inositol phosphate metabolism	16	7	4.11E-02
		Propanoate metabolism	7	4	4.47E-02
		SNARE interactions in vesicular transport	7	4	4.47E-02
		Phosphonate and phosphinate metabolism	2	2	4.76E-02

Supplementary table 4: Gene Ontology (GO) term enrichment in shifted genes. *Genes were divided into groups according to whether they come from singleton and duplicate orthogroups and whether they are shifted up or down in expression level. For each group, the top 20 most significantly enriched terms with more than two total genes are shown. For significantly enriched GO terms, the total number of genes with this term (Total), the number of these that had a shift in expression (Shifted) and the corresponding p-value (P-value) is shown.*

Ortholog	Shift direction	GO ID	GO description	Total	Shifted	P-value
Singletons	Shift up	GO:0008152	metabolic process	885	110	1.11E-03
		GO:0071704	organic substance metabolic process	807	100	3.03E-03
		GO:0044237	cellular metabolic process	759	94	4.88E-03
		GO:0051186	cofactor metabolic process	29	8	6.31E-03
		GO:0006732	coenzyme metabolic process	18	6	6.54E-03
		GO:0044238	primary metabolic process	756	92	1.01E-02
		GO:0006743	ubiquinone metabolic process	2	2	1.02E-02
		GO:0006744	ubiquinone biosynthetic process	2	2	1.02E-02
		GO:0098781	ncRNA transcription	2	2	1.02E-02
		GO:1901661	quinone metabolic process	2	2	1.02E-02
		GO:1901663	quinone biosynthetic process	2	2	1.02E-02
		GO:0006733	oxidoreduction coenzyme metabolic process	10	4	1.31E-02
		GO:0043170	macromolecule metabolic process	603	75	1.47E-02
		GO:0006807	nitrogen compound metabolic process	690	84	1.58E-02
		GO:1901576	organic substance biosynthetic process	298	41	1.78E-02
		GO:0044249	cellular biosynthetic process	283	39	2.03E-02
		GO:0009058	biosynthetic process	302	41	2.20E-02
		GO:0000082	G1/S transition of mitotic cell cycle	17	5	2.27E-02
		GO:0044843	cell cycle G1/S phase transition	17	5	2.27E-02
		GO:0000723	telomere maintenance	7	3	2.63E-02
	Shift down	GO:0034641	cellular nitrogen compound metabolic process	484	42	1.45E-03
		GO:0071704	organic substance metabolic process	807	60	4.73E-03
		GO:0006094	gluconeogenesis	7	3	5.50E-03
		GO:0019319	hexose biosynthetic process	7	3	5.50E-03
		GO:0046364	monosaccharide biosynthetic process	7	3	5.50E-03
		GO:0070838	divalent metal ion transport	7	3	5.50E-03
		GO:0008152	metabolic process	885	64	6.39E-03
		GO:0006139	nucleobase-containing compound metabolic process	413	35	6.72E-03
		GO:0044085	cellular component biogenesis	96	12	7.51E-03
		GO:0072511	divalent inorganic cation transport	8	3	8.44E-03
		GO:0006807	nitrogen compound metabolic process	690	52	8.50E-03
		GO:0001675	acrosome assembly	3	2	9.51E-03
		GO:0006206	pyrimidine nucleobase metabolic process	3	2	9.51E-03

		GO:0006382	adenosine to inosine editing	3	2	9.51E-03
		GO:0010927	cellular component assembly involved in morphogenesis	3	2	9.51E-03
		GO:0016553	base conversion or substitution editing	3	2	9.51E-03
		GO:0033363	secretory granule organization	3	2	9.51E-03
		GO:0006725	cellular aromatic compound metabolic process	423	35	9.98E-03
		GO:0046483	heterocycle metabolic process	424	35	1.04E-02
		GO:0022607	cellular component assembly	54	8	1.06E-02
Duplicates	Shift up	GO:0006633	fatty acid biosynthetic process	4	3	2.47E-04
		GO:0051186	cofactor metabolic process	12	4	9.61E-04
		GO:0019367	fatty acid elongation, saturated fatty acid	2	2	1.61E-03
		GO:0030497	fatty acid elongation	2	2	1.61E-03
		GO:0072330	monocarboxylic acid biosynthetic process	7	3	1.98E-03
		GO:0006631	fatty acid metabolic process	16	4	3.12E-03
		GO:0045786	negative regulation of cell cycle	26	5	3.23E-03
		GO:0008610	lipid biosynthetic process	18	4	4.93E-03
		GO:0016053	organic acid biosynthetic process	10	3	6.21E-03
		GO:0046394	carboxylic acid biosynthetic process	10	3	6.21E-03
		GO:0051188	cofactor biosynthetic process	4	2	9.16E-03
		GO:0051726	regulation of cell cycle	33	5	9.32E-03
		GO:0000075	cell cycle checkpoint	22	4	1.04E-02
		GO:0044255	cellular lipid metabolic process	48	6	1.15E-02
		GO:0032787	monocarboxylic acid metabolic process	23	4	1.22E-02
		GO:0001558	regulation of cell growth	15	3	2.03E-02
		GO:0040008	regulation of growth	15	3	2.03E-02
		GO:0051656	establishment of organelle localization	15	3	2.03E-02
		GO:0006629	lipid metabolic process	57	6	2.54E-02
		GO:0006082	organic acid metabolic process	45	5	3.28E-02
	Shift down	GO:0016071	mRNA metabolic process	88	40	4.77E-07
		GO:0006412	translation	27	17	4.50E-06
		GO:0006413	translational initiation	8	8	5.09E-06
		GO:0043043	peptide biosynthetic process	28	17	9.16E-06
		GO:0009057	macromolecule catabolic process	50	25	9.22E-06
		GO:0045333	cellular respiration	14	11	9.68E-06
		GO:0044265	cellular macromolecule catabolic process	46	23	2.12E-05
		GO:0006518	peptide metabolic process	32	18	2.18E-05
		GO:0006120	mitochondrial electron transport, NADH to ubiquinone	7	7	2.35E-05
		GO:0006119	oxidative phosphorylation	11	9	3.93E-05
		GO:0022900	electron transport chain	11	9	3.93E-05
		GO:0022904	respiratory electron transport chain	11	9	3.93E-05
		GO:0008380	RNA splicing	51	24	5.06E-05
		GO:0034622	cellular protein-containing complex assembly	34	18	6.51E-05
		GO:0000375	RNA splicing, via transesterification reactions	50	23	1.13E-04
		GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	19	12	1.19E-04
		GO:0042773	ATP synthesis coupled electron transport	10	8	1.49E-04
		GO:0042775	mitochondrial ATP synthesis coupled electron transport	10	8	1.49E-04
		GO:0006397	mRNA processing	54	24	1.57E-04
		GO:0065003	protein-containing complex assembly	36	18	1.71E-04

References

Amores, A., Catchen, J., Ferrara, A., Fontenot, Q. and Postlethwait, J.H. 2011. Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted gar, an outgroup for the teleost genome duplication. *Genetics* 188(4), pp. 799–808.

Bedford, T. and Hartl, D.L. 2009. Optimization of gene expression by natural selection. *Proceedings of the National Academy of Sciences of the United States of America* 106(4), pp. 1133–1138.

Berthelot, C., Brunet, F., Chalopin, D., Juanchich, A., Bernard, M., Noël, B., Bento, P., Da Silva, C., Labadie, K., Alberti, A., Aury, J.-M., Louis, A., Dehais, P., Bardou, P., Montfort, J., Klopp, C., Cabau, C., Gaspin, C., Thorgaard, G.H., Boussaha, M. and Guiguen, Y. 2014. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nature Communications* 5, p. 3657.

Berthelot, C., Villar, D., Horvath, J.E., Odom, D.T. and Flicek, P. 2018. Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nature ecology & evolution* 2(1), pp. 152–163.

Braasch, I., Gehrke, A.R., Smith, J.J., Kawasaki, K., Manousaki, T., Pasquier, J., Amores, A., Desvignes, T., Batzel, P., Catchen, J., Berlin, A.M., Campbell, M.S., Barrell, D., Martin, K.J., Mulley, J.F., Ravi, V., Lee, A.P., Nakamura, T., Chalopin, D., Fan, S. and Postlethwait, J.H. 2016. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nature Genetics* 48(4), pp. 427–437.

Carmona-Antoñanzas, G., Tocher, D.R., Taggart, J.B. and Leaver, M.J. 2013. An evolutionary perspective on Elovl5 fatty acid elongase: comparison of Northern pike and duplicated paralogs from Atlantic salmon. *BMC Evolutionary Biology* 13, p. 85.

Carmona-Antoñanzas, G., Zheng, X., Tocher, D.R. and Leaver, M.J. 2016. Regulatory divergence of homeologous Atlantic salmon elovl5 genes following the salmonid-specific whole-genome duplication. *Gene* 591(1), pp. 34–42.

Chen, J., Swofford, R., Johnson, J., Cummings, B.B., Rogel, N., Lindblad-Toh, K., Haerty, W., Palma, F. di and Regev, A. 2019. A quantitative framework for characterizing the evolutionary history of mammalian gene expression. *Genome Research* 29(1), pp. 53–63.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T.R. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29(1), pp. 15–21.

Emms, D.M. and Kelly, S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* 16, p. 157.

Freeling, M. and Thomas, B.C. 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Research* 16(7), pp. 805–814.

Gillard, G., Harvey, T.N., Gjuvsland, A., Jin, Y., Thomassen, M., Lien, S., Leaver, M., Torgersen, J.S., Hvidsten, T.R., Vik, J.O. and Sandve, S.R. 2018. Life-stage-associated remodelling of lipid metabolism regulation in Atlantic salmon. *Molecular Ecology* 27(5), pp. 1200–1213.

Hermansen, R.A., Hvidsten, T.R., Sandve, S.R. and Liberles, D.A. 2016. Extracting functional trends from whole genome duplication events using comparative genomics. *Biological procedures online* 18, p. 11.

Hoegg, S., Brinkmann, H., Taylor, J.S. and Meyer, A. 2004. Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. *Journal of Molecular Evolution* 59(2), pp. 190–203.

Holland, P.W., Garcia-Fernàndez, J., Williams, N.A. and Sidow, A. 1994. Gene duplications and the origins of vertebrate development. *Development (Cambridge, England). Supplement*, pp. 125–133.

Kalinka, A.T., Varga, K.M., Gerrard, D.T., Preibisch, S., Corcoran, D.L., Jarrells, J., Ohler, U., Bergman, C.M. and Tomancak, P. 2010. Gene expression divergence recapitulates the developmental hourglass model. *Nature* 468(7325), pp. 811–814.

Katoh, K., Misawa, K., Kuma, K. and Miyata, T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* 30(14), pp. 3059–3066.

Langfelder, P. and Horvath, S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9, p. 559.

Li, B. and Dewey, C.N. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12, p. 323.

Lien, S., Koop, B.F., Sandve, S.R., Miller, J.R., Kent, M.P., Nome, T., Hvidsten, T.R., Leong, J.S., Minkley, D.R., Zimin, A., Grammes, F., Grove, H., Gjuvsland, A., Walenz, B., Hermansen, R.A., von Schalburg, K., Rondeau, E.B., Di Genova, A., Samy, J.K.A., Olav Vik, J. and Davidson, W.S. 2016. The Atlantic salmon genome provides insights into rediploidization. *Nature* 533(7602), pp. 200–205.

Lynch, M. and Conery, J.S. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290(5494), pp. 1151–1155.

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. Biological Sciences / the Royal Society* 281(1778), p. 20132881.

Meyer, A. and Van de Peer, Y. 2005. From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *Bioessays: News and Reviews in Molecular, Cellular and Developmental Biology* 27(9), pp. 937–945.

Ohno, S. 1970. *Evolution by Gene Duplication*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Perry, G.H., Melsted, P., Marioni, J.C., Wang, Y., Bainer, R., Pickrell, J.K., Michelini, K., Zehr, S., Yoder, A.D., Stephens, M., Pritchard, J.K. and Gilad, Y. 2012. Comparative RNA sequencing reveals substantial genetic variation in endangered primates. *Genome Research* 22(4), pp. 602–610.

Price, M.N., Dehal, P.S. and Arkin, A.P. 2010. FastTree 2 — approximately maximum-likelihood trees for large alignments. *Plos One* 5(3), p. e9490.

Rabosky, D.L., Chang, J., Title, P.O., Cowman, P.F., Sallan, L., Friedman, M., Kaschner, K., Garilao, C., Near, T.J., Coll, M. and Alfaro, M.E. 2018. An inverse latitudinal gradient in speciation rate for marine fishes. *Nature* 559(7714), pp. 392–395.

Robertson, F.M., Gundappa, M.K., Grammes, F., Hvidsten, T.R., Redmond, A.K., Lien, S., Martin, S.A.M., Holland, P.W.H., Sandve, S.R. and Macqueen, D.J. 2017. Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification. *Genome Biology* 18(1), p. 111.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26(1), pp. 139–140.

Rohlfs, R.V., Harrigan, P. and Nielsen, R. 2014. Modeling gene expression evolution with an extended Ornstein-Uhlenbeck process accounting for within-species variation. *Molecular Biology and Evolution* 31(1), pp. 201–211.

Rohlfs, R.V. and Nielsen, R. 2015. Phylogenetic ANOVA: the expression variance and evolution model for quantitative trait evolution. *Systematic Biology* 64(5), pp. 695–708.

Sandve, S.R., Rohlfs, R.V. and Hvidsten, T.R. 2018. Subfunctionalization versus neofunctionalization after whole-genome duplication. *Nature Genetics* 50(7), pp. 908–909.

Schoenfelder, K.P. and Fox, D.T. 2015. The expanding implications of polyploidy. *The Journal of Cell Biology* 209(4), pp. 485–491.

Sémon, M. and Wolfe, K.H. 2007. Consequences of genome duplication. *Current Opinion* *in Genetics & Development* 17(6), pp. 505–512.

Van de Peer, Y., Maere, S. and Meyer, A. 2009. The evolutionary significance of ancient genome duplications. *Nature Reviews. Genetics* 10(10), pp. 725–732.

Van de Peer, Y., Mizrachi, E. and Marchal, K. 2017. The evolutionary significance of polyploidy. *Nature Reviews. Genetics* 18(7), pp. 411–424.

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.

Volff, J.N. 2005. Genome evolution and biodiversity in teleost fish. *Heredity* 94(3), pp. 280–294.

ISBN: 978-82-575-1612-3
ISSN: 1894-6402

N M B U
Norwegian University
of Life Sciences

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

103272 / ANDVORDGRAFISK.NO