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, Oslo NO-0316, Norway, [2]Centre for Integrative Genetics (CIGENE), Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, Ås NO-1432, Norway, [3]Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Ås NO-1432, Norway, [4]Umeå Plant Science Centre, Department of Plant Physiology, Umeå University, Umeå, Sweden, [5]Biomedical Informatics Research Group, Department of Informatics, University of Oslo, Oslo NO-0316, Norway

*These authors contributed equally to this work

‡Corresponding authors: simen.sandve@nmbu.no, k.s.jakobsen@ibv.uio.no

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 million years ago (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 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, one third reflect non-neutral tissue expression evolution, with strong purifying selection, maintained over ∼50 million years. Of these, 84% reflect conserved tissue regulation under strong selective constraints and are involved in brain and neural-related functions, as well as higher-order protein-protein interactions. In contrast, 16% of the duplicates have evolved regulatory divergence in a common ancestor, suggestive of adaptive divergence following WGD. These candidates for adaptive expression divergence have elevated rates of protein coding- and promoter sequence evolution, and are enriched for immune- and metabolism ontology terms. Lastly, species-specific duplicate divergence points towards underlying differences in adaptive pressures on expression regulation in the non-anadromous grayling and anadromous Atlantic salmon. Our findings enhance our understanding of the role of WGD in genome evolution and highlights cases of functional divergence of Ss4R duplicates, possibly related to a niche shift in early salmonid evolution.

Introduction

Whole genome duplication (WGD) through spontaneous doubling of all chromosomes (autopolyploidization) has played a vital role in the evolution of vertebrate genome complexity [1]. 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 made famous by Susumu Ohno [2]. Duplicate genes that escape loss or pseudogenization are known to acquire novel regulation and expression divergence [3, 4, 5]. 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 [6]. As changes in gene regulation are known to be important in the evolution of phenotypic diversity and complex trait variation [7, 8], 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 (e.g. [9, 10, 11, 12, 13, 14, 15] 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 functional data from multiple species sharing the same WGD [16]. 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, <100MYA) [2, 17] and ongoing rediploidization [18, 19, 14, 15]. 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 non-viable aneuploid gametes. Restoring bivalent chromosome pairing is therefore a critical step towards a functional genome post-WGD [20]. This can be achieved through e.g. structural rearrangements that suppress recombination, block multivalent formation, and drive the process of returning to a functional diploid state (i.e. rediploidization). Since 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 the species radiation, resulting in lineage-specific ohnolog resolution (LORe) that may fuel differentiation of genome structure and function [18, 15]. In fact, due to the delayed rediploidization, only 75% of the duplicated genome diverged before the basal split in the Salmonid family ∼60 MYA (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 [21]. In contrast to the Thymallus lineage, the species in the subfamily Salmoninae have fewer and highly derived chromosomes resulting from large-scale chromosomal rearrangements and chromosomal fusions (Supplementary Figure S1), display extreme phenotypic plasticity, and have evolved the capability of migrating between fresh and saltwater habitats (referred to as anadromy) [17, 22, 23, 24, 25]. 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*), a species representative of an early diverging non-anadromous salmonid lineage. We use this novel genomic resource in a comparative phylogenomic framework 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 are correlated to differences in evolutionary constraints during the 88-100MY of evolution post-WGD, pointing towards underlying differences in adaptive pressures in non-anadromous grayling and anadromous Atlantic salmon.

Results

Genome assembly and annotation

We sequenced the genome of a wild-caught male grayling individual, sampled from the Norwegian river Glomma, using the Illumina HiSeq 2000 platform (Supplementary Table S1 and S2). *De novo* assembly was performed using ALLPATHS-LG [26], followed by assembly correction using Pilon [27], resulting in 24,343 scaffolds with an N50 of 284 Kbp and a total size of 1.468 Gbp (Table 1). The scaffolds represent approximately 85% of the k-mer based genome size estimate of ∼1.8 Gbp. The C-values estimated previously for European grayling are 2.1pg (`http://www.genomesize.com/`) and 1.9pg [24]. 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 approximately 600Mb (∼40%) of the assembly, dominated by class1 DNA transposable elements (Supplementary Table S3 and a repeat landscape in Supplementary Figure S2). Finally, the transcriptome assemblies, the *de novo* identified repeats along with the UniProt proteins [28] and Atlantic salmon coding sequences [14] 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 (Annotation edit distance), homology with UniProt and Atlantic salmon proteins or presence of known domains. Assembly completeness was assessed at the gene level based on CEGMA and BUSCO. The assembly contains 236 (95.16%) out of 248 conserved eukaryotic genes (CEGs) with 200 (80.65%) complete CEGs. Of the 3,698 BUSCO genes

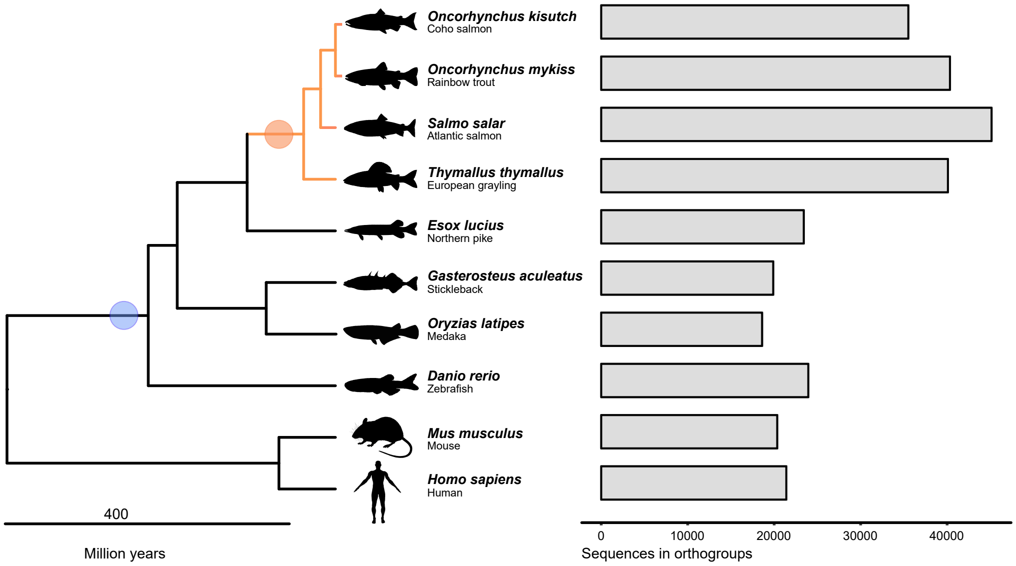

Figure 1 Species and genes in ortholog groups.
Left: phylogenetic relationship of species used for constructing ortholog groups and gene trees. The blue circle indicates the 3R-WGD event while the Ss4R event is indicated with an orange circle. Right: number of genes assigned to ortholog groups in each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

of the class Actinopterygii, 3,192 complete (86.3%) and 222 (6%) 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 [14, 15]. 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* (stickleback), *Oryzias latipes* (medaka), *Esox lucius* (northern pike), *Salmo salar* (Atlantic salmon), *Oncorhynchus mykiss* (rainbow trout) and *Oncorhynchus kisutch* (coho salmon) (Figure 1). These orthogroups were used to infer gene trees. 20,342 gene trees contained WGD events older than Ss4R (Ts3R or 2R) and were further subdivided into smaller sub-groups (i.e. clans, see Methods for details and Supplementary Figure S3). To identify orthogroups with retained Ss4R duplicates, we relied on the high-quality reference genome of Atlantic salmon [14]. A synteny-aware blast approach [14] was first used to identify Ss4R duplicate pairs/ohnolog

Table 1 Genome assembly statistics.

Assembly Statistics	
Total size of scaffolds (bp)	1,468,519,221
Number of scaffolds	24,369
Scaffold N50 (bp)	283,328
Longest scaffold (bp)	2,502,076
Total size of contigs (bp)	1,278,330,545
Number of contigs	216,549
Contig N50 (bp)	11,206
Assembly validation	
Complete CEGMA[a] genes	80.65% (200/248)
Partial CEGMA genes	95.16% (236/248)
Complete Single-Copy BUSCOs[b]	3192 (86.3%)
Complete Duplicated BUSCOs	896 (24.2%)
Fragmented BUSCOS	222 (6%)
Missing BUSCOS	284 (7.7%)
Total BUSCOS searched	3,698

[a] Based on 248 highly Conserved Eukaryotic Genes (CEGS),[b] Based on 3,698 actinopterygii-specific BUSCO genes

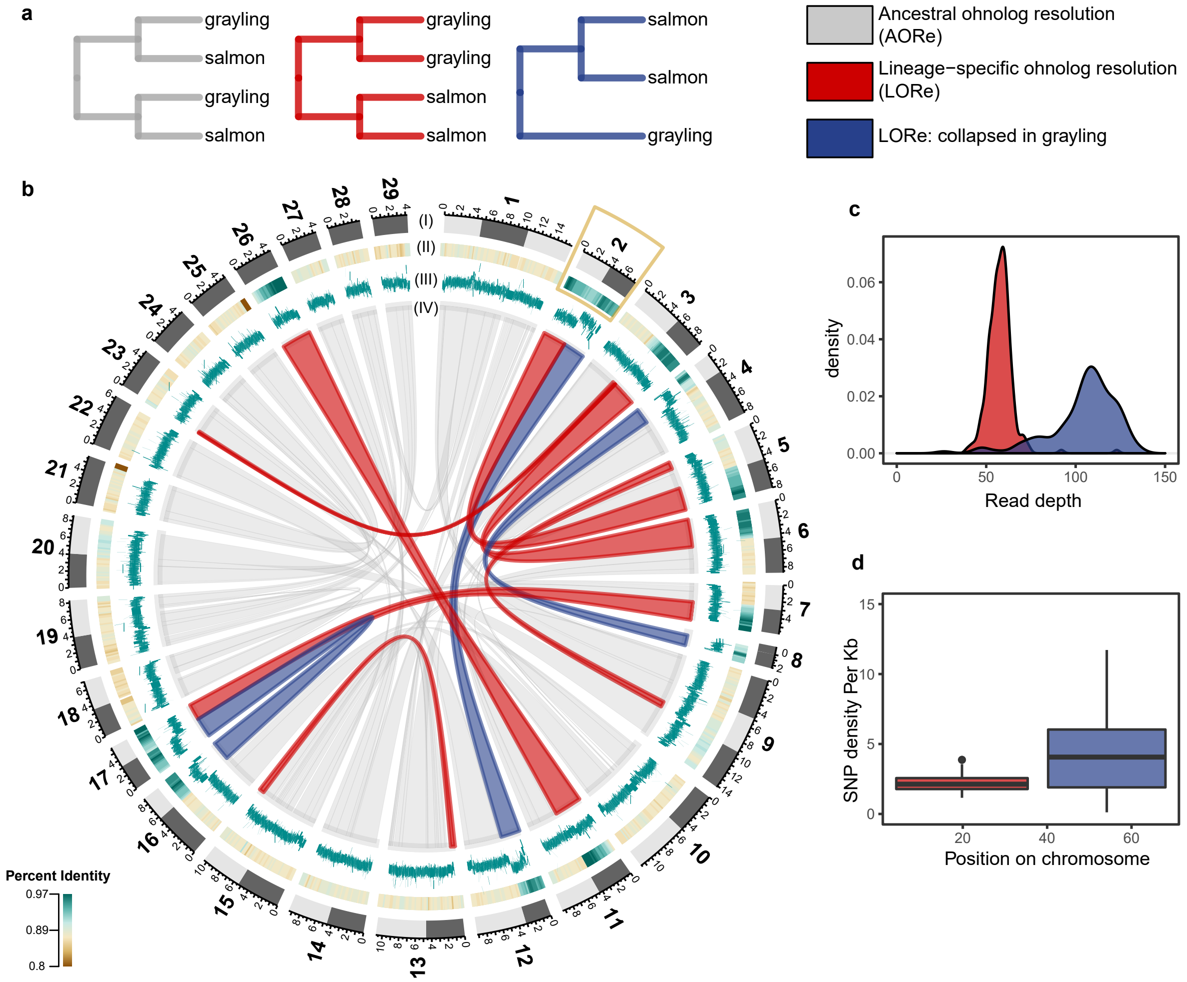

Figure 2 Rediploidization in grayling genome.
a) Gene tree topologies corresponding to the different models of ohnolog resolution (ancestral divergence of ohnologs (AORe) and lineage-specific divergence of ohnologs (LORe and LORe-like regions with repeat collapse in grayling). b) Circos plot: Outer track (I) represents the 29 chromosomes of Atlantic salmon with chromosome arms indicated using light and dark grey. (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 grey 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 plot shows 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)).

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 ancestral ohnolog resolution (AORe) (Figure 2a). 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 [15] and cases that did not conform were omitted from further analyses. The final set consisted of 5,475 gene trees containing Ss4R duplicates from both species (4,735 AORe, 740 LORe), in addition to 482 ortholog sets containing 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 (Figure 2b). In Atlantic salmon, several home-

Table 2 Classification of Expression Evolution Fates (EEF). The number of genes and percentages calculated based on the total number of topology-filtered ohnolog-tetrads.

EEF	Description	AORe	LORe
EEF1	Conserved divergence: *Duplicates in both species have evolved identical novel expression regulation*	199 (5.7%)	24 (4.7%)
EEF2	Fixed-specific divergence: *Tissue regulation among duplicates are conserved within species but different between species*	195 (5.6%)	51 (10.0%)
EEF3	Salmon-specific divergence: *One Atlantic salmon duplicate has diverged in expression regulation*	375 (10.8%)	70 (13.8%)
EEF4	Grayling-specific divergence: *One grayling duplicate has diverged in expression regulation*	516 (14.8%)	80 (15.7%)
EEF5	Conserved: *All genes in the ohnolog-tetrad have conserved tissue regulation*	869 (25%)	131 (25.7%)
EEF6	Unclassified: *Tetrads with neutral-like expression evolution*	1326 (38.1%)	153 (30.1%)
Total		3480	509

ologous chromosome arms (2p-5q, 2q-12qa, 3q-6p, 4p-8q, 7q-17qb, 11qa-26, 16qb-17qa) have previously been described as Ss4R regions under delayed rediploidization [14, 15](indicated in Figure 2b as red and blue ribbons). Interestingly, the homeologous LORe regions 2q-12qa, 4p-8q and 16qb-17qa in Atlantic salmon had only one orthologous region in grayling, suggesting either loss of large duplicated blocks or sequence assembly collapse in grayling. To assess this, we mapped the grayling Illumina paired end reads that were used for the assembly back to the grayling genome sequence using BWA-MEM [29] and determined the mapped read depth for each of the grayling genes. Single-copy grayling genes in LORe regions had consistently double read depth (∼100x) compared to the LORe duplicates in grayling (Figure 2c and Supplementary Figure S4a), indicating assembly collapse rather than loss of large chromosomal regions. Additionally, the SNP density of the scaffolds in these regions computed using FreeBayes [30] (quality filter of 30) displayed values on an average twice the background SNP density, albeit with a much wider distribution (Figure 2d and Supplementary Figure S4b).

Evolution of tissue gene expression regulation following WGD.

To investigate how selection has operated on Ss4R ohnologs, we used tissue gene expression data from Atlantic salmon and grayling in a comparative phylogenetic approach. We classified expression evolution fates (EEF) of Ss4R duplicates by clustering duplicate pairs in Atlantic salmon and grayling, hereafter referred to as ohnolog-tetrads, into eight tissue-dominant expression clusters (Methods, Supplementary Figure S5). Next, the ohnolog-tetrads were classified into five groups each representing differences in the past selection pressure on the tissue regulation of ohnolog pairs (see Table 2, Figure 3). The conserved divergence (EEF1) category represents expression divergence among ohnologs that is identical for both species. EEF1 is thus best explained by purifying selection on ancestral ohnolog expression divergence. Fixed-lineage divergence of expression (EEF2) represents cases of conserved expression regulation among duplicates within species. EEF3 and 4 include ohnolog-tetrads with species-specific expression divergence pointing to species-specific adaptive divergence or relaxed purifying selection in one of duplicate genes. Lastly, EEF5 contains ohnolog-tetrads with all genes expressed in the same tissue, thus pointing to strong purifying selection to maintain ancestral tissue-specificity. In addition to these five categories, there were ohnolog-tetrads where three, or all four of the duplicates were in different tissue-expression clusters. These were grouped into a 6th 'unclassified' EEF category assumed to be enriched in ohnolog-tetrads under neutral or nearly neutral evolution, or a result of low tissue specificity (Table 2). After applying a gene tree topology-based filtering criterion (see Methods) to the genes in the EEF categories, 3,989 ohnolog-tetrads that conformed to expectations of LORe (509) or AORe (3480) gene tree topologies were used in further analyses.

Of the six classes of EEFs, unclassified (EEF6, 30-38%) and conserved tissue regulation (EEF5, ∼25%)

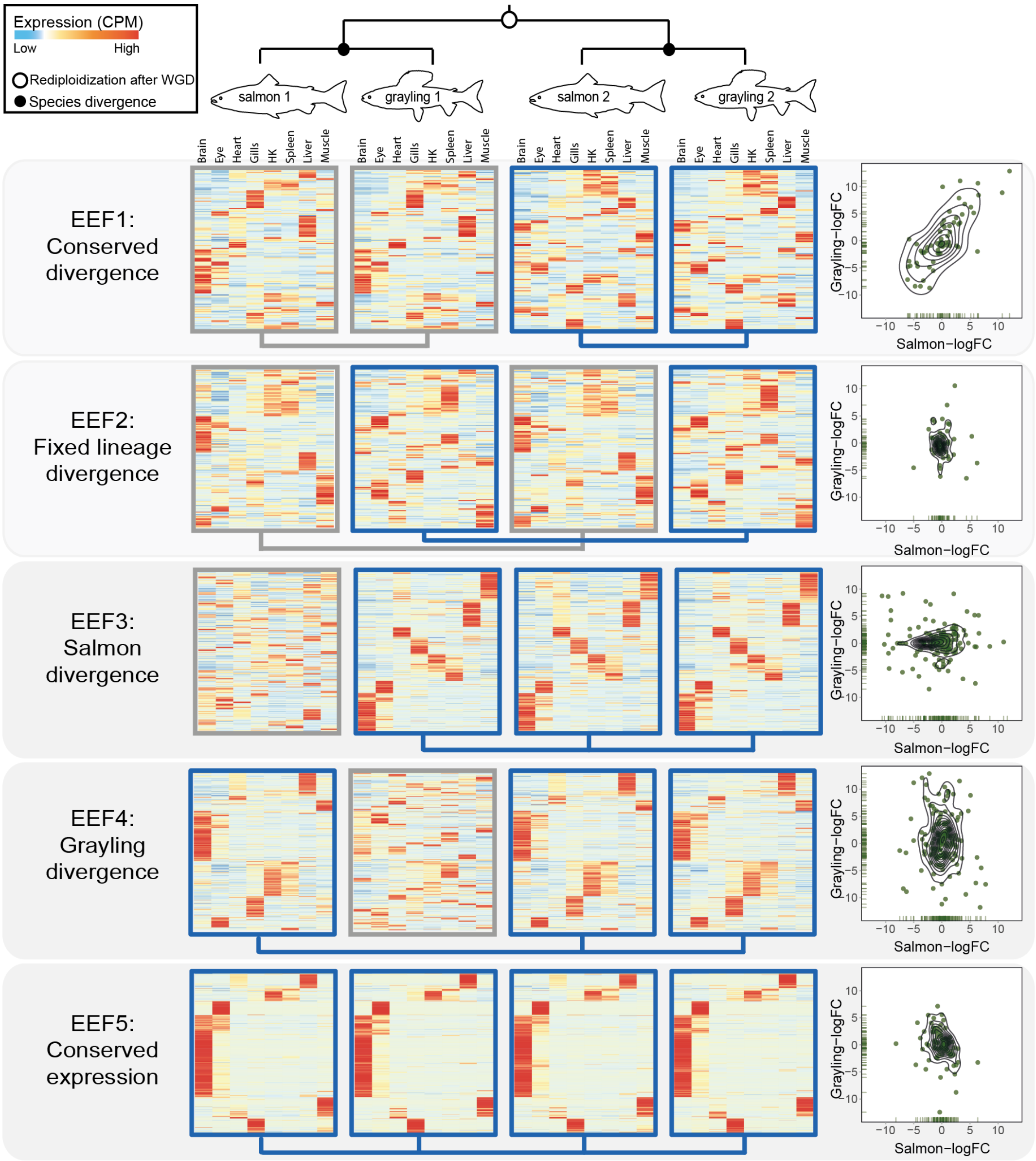

Figure 3 Selection on tissue expression regulation after whole genome duplication. Heatmaps showing clustering of expression values of the ohnolog tetrads across the five non-neutral Expression Evolution Fates (EEFs) reflecting differential selection on tissue expression regulation following Ss4R WGD (see Table 2). The phylogenetic tree (top) represents a typical AORe (ancestrally diverged) topology corresponding to the ohnolog-tetrads represented in the figure (analogous patterns were observed in LORe, see Supplementary Figure S6). Each row across the four heatmaps represents one ohnolog-tetrad (four genes), with darker red corresponding to the highest expression level observed for one gene (scaled counts per million, CPM). Connecting blue lines below the heatmaps indicate duplicates belonging to the same tissue clusters (conserved expression pattern). Each EEF has an associated scatter plot (at the right) with density lines showing, for the ohnolog-tetrads associated with a liver cluster, the log2 fold change (logFC) of gene expression, in liver, between Ss4R duplicates of Atlantic salmon against that of grayling.

were the most common, followed by species-specific divergence of one duplicate (EEF3 and 4), lineage-specific divergence of both duplicates (EEF2) and conserved divergence (EEF1; Table 2). Although the size ranking of the EEF categories were similar (Table 2), the relative EEF category fractions were significantly different between AORe and LORe (Fisher's exact test, two sided, p-value $<$ 0.0005). For example, we observed a doubling of lineage-specific expression divergence (EEF2) among LORe tetrads, consistent with these genomic regions having undergone lineage specific rediploidization.

As different tissues are involved in different biological functions, we expect that regulatory evolution is shaped by tissue-specific selective pressures [31]. To test this expectation, we evaluated the hypothesis that tissues are disproportionately represented in EEFs 1-5 compared to the tissue distribution across all tetrads. For all EEF-classes, between 2-5 tissue-expression clusters were significantly over- or under-represented (Fisher tests, two sided, Bonferroni corrected p-value<=0.05), with the conserved tissue regulation class (EEF5) being the most skewed in tissue representation with a strong bias towards brain-specific expression (Supplementary Table S4). The high tissue-specificity (Tau score) of genes in ohnolog-tetrads associated with EEF5 (Supplementary Figure S7) corroborates the observed brain-specific expression bias.

Next, we tested whether distinct evolutionary trajectories at the regulatory level (EEFs 1-5) were coupled to patterns of protein-coding and promoter sequence evolution. We estimated dN/dS ratios for each duplicate pair within each species and compared the distribution of dN/dS statistics in each EEF class with that of the neutral-like ('unclassified', EEF6) regulatory evolution (Supplementary Figure S8). Low dN/dS (<<1) indicates strong purifying selection pressure. EEF 1-5 show variability in among-ohnolog dN/dS ratio, with conserved divergence (EEF1) having significantly higher dN/dS ratio compared to the neutral-like ('unclassified') category (Wilcoxon rank sum, p=0.005) and EEF 2 and 5 having significantly lower dN/dS ratios (Wilcoxon rank sum, p=0.014 and p=0.0017, respectively). The ohnolog pairs showing species-specific expression divergence (EEF 3 and 4) did not have a significantly different dN/dS ratio compared to the neutral-like category (Wilcoxon rank test, p-values=0.36 and 0.26, respectively). Further, we used the high-quality genome of Atlantic salmon to annotate and compare known transcription factor motifs divergence in the promoters (-1000, +200 bp from transcription start site) of ohnologs. Under the assumption that expression divergence is, at least partly, driven by changes in transcription factor binding motifs, we tested if ohnolog regulatory divergence (EEF1 and EEF3 for salmon) was associated with divergence of promoter motifs. Indeed, the results add validation to the different EEF classifications (Supplementary Figure S9), with EEF1 and EEF3 having significantly less similar promoter motif content compared to EEFs ohnolog with conserved tissue expression regulation (EEF2, 4, and 5) (Wilcoxon test all contrasts between EEF1/3 vs EEF2/4/5, p<0.04-0.002).

To evaluate if the ohnologs in different EEF classes were associated with distinct biological processes, we performed GO term enrichment tests on genes in EEF 1-5. EEF5 ohnologs under strict selective constraints are highly enriched in brain-specific expression and enriched for GO functions related to behaviour and neural functions. In contrast, EEF1, which represents ohnologs that underwent divergence in gene regulation following WGD, are associated with functions related to lipid metabolism, development, and immune system (GO test results in Supplementary file 2).

Highly connected genes in protein-protein interaction networks are often placed under strong constraints to maintain stoichiometry [32, 33]. To test if the strong constraints on the EEF5 class are associated with having higher protein-protein interactions, we extracted all the zebrafish genes from the genes trees corresponding to the ohnologs in the EEFs and queried them against the STRING database [34] (version 10.5). Only associations with a score of above 7.0, suggesting high confidence associations, were retained. We found that, as expected, EEF5 genes were indeed enriched for PPI (enrichment p-value, 1.05e-05) in comparison to the genes in the groups (EEF1, 3 ,4) with diverged expression (enrichment p-value, 0.785).

Evolution of gene expression levels following WGD

As the EEF classification is based on expression correlation across a tissue atlas with a single biological replicate from each tissue, we acquired an independent validation expression dataset with biological replicates, from liver, for grayling (n=4) and Atlantic salmon (n=9). For ohnolog pairs with at least one duplicate assigned to the liver expression cluster, we computed the fold change (FC) expression differences between the duplicates (within each species) and the significance levels for expression level differences (FDR adjusted p-values) in the liver expression data. The correlation of fold change (FC) in ohnolog expression levels (Figure 3) corroborated the conclusions from the tissue atlas analysis. In the EEF1 group, we observe a positive correlation between ohnolog fold change, as expected under conserved ancestral divergence. Lineage-specific divergence (EEF3 and 4) reflects species-specific ohnolog divergence in expression levels (Atlantic salmon specific = horizontal trend , grayling specific = vertical trend), while EEF2 and 5 classes showed no trend in duplicate expression level divergence implying conservation of expression levels among the ohnologs. The statistical significance of expression level differences among ohnologs further supported the EEF classification (Supplementary Table S5). EEF1 showed high proportions of tetrads with significant fold change differences between salmon duplicates (73%) and grayling

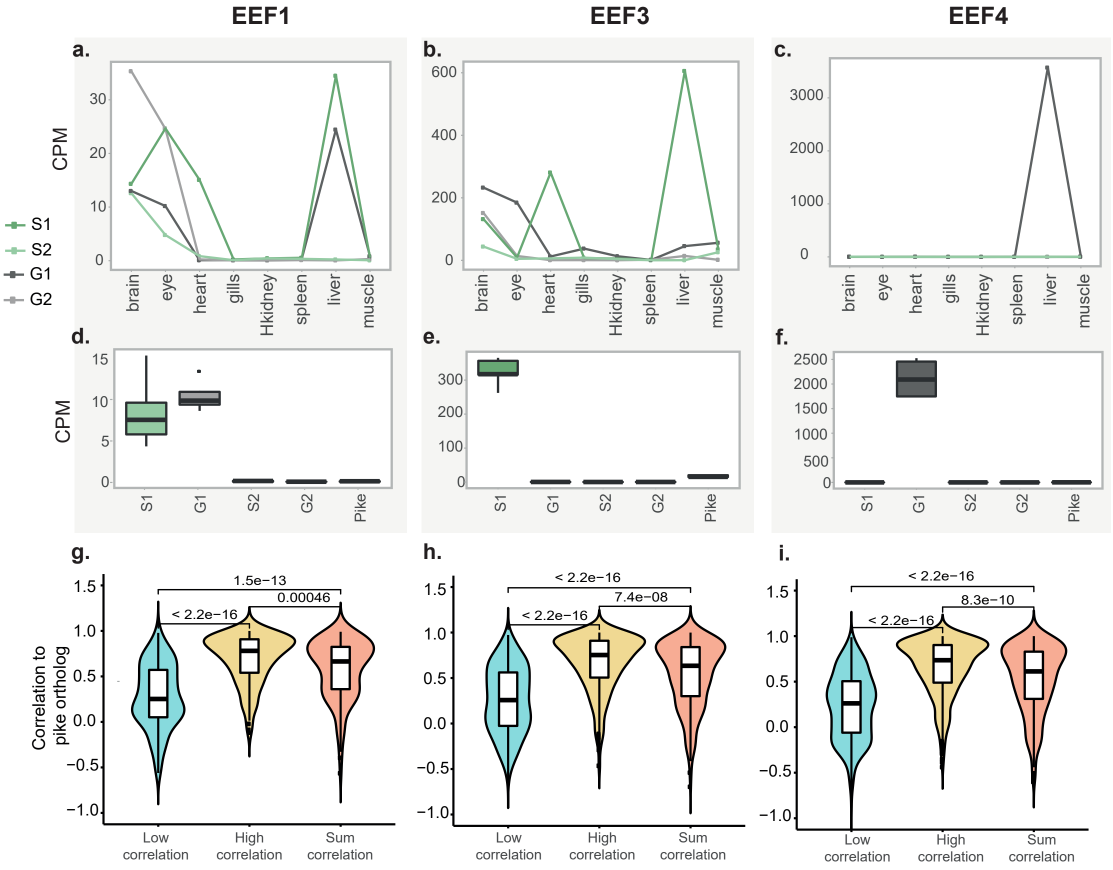

Figure 4 Expression evolution in EEF1, 3 and 4.
Expression levels, in terms of CPM (counts per million), from the tissue atlas data (a-c) and the corresponding data from the liver expression data are plotted in (boxplots in d-f) for a selected example with a liver-specific gain of expression in each of EEF1, 3 and 4. The examples indicated in (a-f) include ohnologs of ephrin type-B receptor 2-like (EEF1), contactin-1a-like gene (EEF3) and a E3 ubiquitin-protein ligase-like gene (EEF4). The ohnologs in Atlantic salmon and grayling are represented as S1, S2 and G1 and G2 respectively. Violin plots (g-i) show the distributions of expression correlations between each of the duplicates, and the sum of the duplicates, to the corresponding pike ortholog for EEF1,3 and 4 respectively. Duplicates belonging to each of the EEFs are divided based on correlation to the pike ortholog into High and Low groups. For EEF1 and 3, the plot represents the distributions of correlations of salmon duplicates, while for EEF4 grayling duplicates are plotted. P-values from pairwise comparisons using Wilcoxon test are indicated.

duplicates (75%), and was the EEF with the highest proportion of tetrads with both species showing differential expression (60%). One such example of an EEF1 tetrad with significant different ohnolog expression levels for both species is the ephrin type-B receptor 2-like duplicate pair (Figure 4a and 4d). Ephrin receptors are classical receptor thyrosine kinases involved in signalling, and play diverse roles in developmental processes, immune function, and organ homeostasis such as insulin regulation [35]. Indeed, the expression of the liver-active copy of this ohnolog pair is induced by addition of insulin in to Atlantic salmon vitro liver slice cultures (personal communication, Thomas Nelson Harvey). Hence, it is plausible that the tissue regulation divergence of the ephrin type-B receptor 2-like is linked to evolution of metabolism function. For the lineage-specific divergence, 70% of EEF3 and 75% EEF4 also showed significant difference in ohnolog expression levels. Two examples of such genes are ohnologs of a contactin-1a-like gene (EEF3, Figure 4b and 4e) and a E3 ubiquitin-protein ligase-like gene (EEF4, Figure 4c and 4f), likely involved in immune function.

Genome-wide patterns of tissue-specific expression divergence among WGD ohnologs in teleosts highlight asymmetric divergence as the most common evolutionary fate, with one ohnolog copy retaining more regulatory similarity with unduplicated orthologs [36]. Yet, a very small proportion (<1%) of ohnologs display expression evolution characteristics that resemble

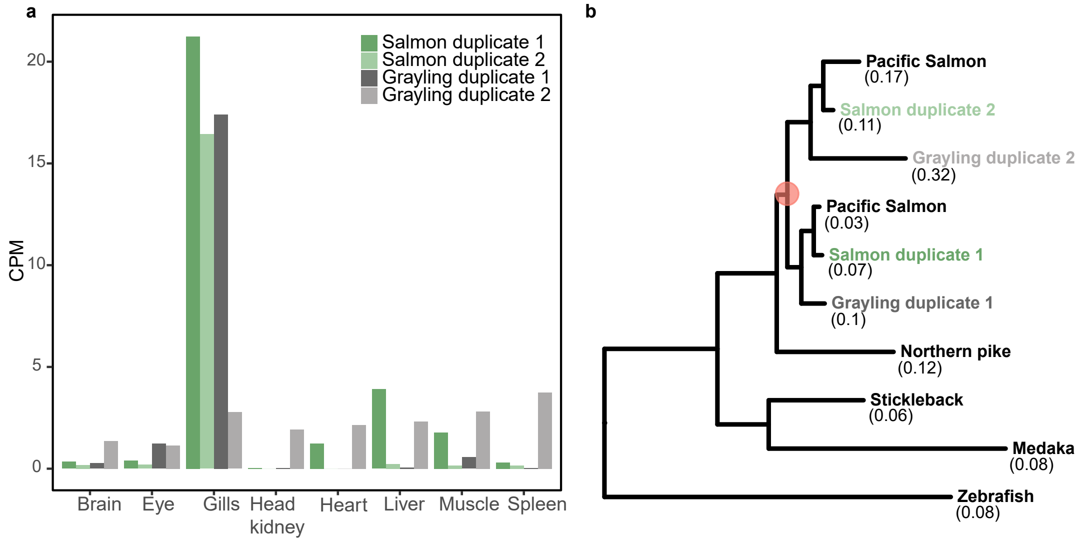

Figure 5 Divergent selection on cystic fibrosis transmembrane conductance regulator.
a) Expression values, in terms of counts per million (CPM), of the cystic fibrosis transmembrane conductance regulator (CFTR) ohnologs in Atlantic salmon and grayling across eight tissues. b) CFTR gene tree. The orange circle represents the Ss4R duplication. Branch-specific dN/dS of the tip nodes are indicated within parentheses.

sub-functionalization of tissue regulation [14]. Since such a partitioning of tissue expression contribution would necessarily lead to selection for maintaining both copies, the EEF1 class could indeed be associated with this atypical mode of expression divergence. Under a model of sub-functionalization, the sum of expression levels of both ohnologs should correlate better to the assumed ancestral expression regulation than any of the individual ohnologs [36]. To test this, we included tissue expression data from pike orthologs in the analyses as a proxy for the ancestral regulatory state. As expected, the predominant correlation pattern reflected regulatory neo-functionalization (Figure 4g,h and i). This further supports the EEF1 class as representing cases of ancestral adaptive divergence of ohnolog expression regulation since the neo-functionalized gene regulation has been conserved over 50 million years in both species. Adaptive expression divergence could likely be linked to a gain of tissue expression levels. Unfortunately, tissue gene expression correlations are difficult to interpret directly with regard to the direction of expression level shifts. Utilizing the additional liver expression dataset, we could not only independently validate the expression patterns in the EEF categories by verifying the correspondence with liver replicates, but also highlight some examples of putative liver-specific expression gains in the ohnologs in EEF1, 3 and 4 (examples in Figure 4 a-f).

Loss of purifying selection on chloride ion transporter regulation in non-anadromous grayling

The most apparent difference in biology between grayling and Atlantic salmon is the anadromous life history in Atlantic salmon, i.e. the ability to migrate between freshwater and saltwater, a trait that grayling has not evolved. Saltwater acclimation involves changes in switching from ion absorption to ion secretion to maintain osmotic homeostasis. To assess whether key genes associated with the ability to adapt to seawater are under divergent selection for expression regulation in Atlantic salmon and grayling, we probed into EEF 3 and 4 for overrepresented GO terms related to ion-homeostasis (i.e. potassium, sodium or chloride regulation/transport). Interestingly, in the group exhibiting grayling-specific regulatory divergence (EEF4), we found that 'regulation of chloride transport' was overrepresented. One of the genes associated with this GO term was the classical anadromy-associated, salinity-induced, cystic fibrosis transmembrane conductance regulator (CFTR). The CFTR gene transports chloride ions over cell membranes in the gill and is involved in saltwater adaptations in Atlantic salmon [37]. Looking into the tissue expression profiles of this tetrad (Figure 5a) it was evident that the divergence of tissue regulation in grayling was associated with a loss of gill tissue expression specificity compared to Atlantic salmon. To determine if the grayling CFTR duplicate with diverged expression also had signatures of coding sequence divergence, we computed branch-specific dN/dS. Notably, the grayling CFTR displaying diverged expression regulation also displays a two-fold increase in dN/dS compared to its Ss4R duplicate with conserved expression regulation, reflecting relaxation of purifying selection pressure on one ohnolog in the non-anadromous grayling (Figure 5b). These results could indicate that there could be a fitness advantage of maintaining two copies of gill expressed CFTR for anadromous species, but not for pure freshwater adapted species, such as grayling.

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 novel shifts in expression [16]. Our comparative approach provides new insights into the importance of selection in the contrasting processes that maintain ancestral gene regulation and drive spatial expression divergence of ohnologs following WGD.

The least common, yet possibly most intriguing, expression evolution fate are EEF1 duplicates that reflect regulatory divergence in a shared salmonid ancestor, followed by strong purifying selection in both grayling and Atlantic salmon. These ohnologs represent key candidates for salmonid-specific adaptive evolution of novel gene regulation enabled by the WGD. Salmonids are suggested to have evolved from a pike-like ancestor; a relatively stationary ambush predator [38]. 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 (e.g [39, 40]). In line with this, we see an overrepresentation of immune-related genes in the EEF1 class. Furthermore, pikes are generally piscivorous throughout their lifespan, while salmonids depend more on aquatic and terrestrial invertebrate prey with significantly lower input of lipids (especially in early life) [41]. Interestingly, the EEF1 duplicates are also enriched for liver-expressed genes involved in lipid-homeostasis metabolism and energy storage (glycogen) related functions (GO test results in Supplementary file 2). Taken together, our results suggest a role of Ss4R ohnologs in adaptive evolution of novel gene expression regulation related to new pathogenic pressures in a new type of habitat, and also optimization of lipid-homeostasis and glycogen metabolism-related functions in response to evolution of a more active pelagic/riverine life with limited lipid resources.

The most commonly observed non-neutral expression evolution fate of ohnologs following Ss4R is the remarkable conservation of tissue-specific expression, predominantly in brain, across the 60 million years of independently evolving salmonid lineages (Table 2, EEF5). A strong expression conservation pattern in brain related genes has been described across vertebrates [42, 43, 44, 45]. Brain-specific genes are typically under strong purifying selection pressure owing to their specialized functions in specific cell types and complex networks of signaling cascades involving high-dimensional protein-protein interactions. Involvement in protein-protein interaction networks is expected to result in constraints to maintain stoichiometry. Our results also demonstrate that EEF5 genes are enriched in protein-protein interactions, and thereby corroborate the previous observation of biased retention of brain expressed WGD-derived duplicate genes.

Our comparative analysis of Ss4R duplicates in Atlantic salmon and grayling suggests a difference in the rate of rediploidization between the two species. We find a set of LORe regions, corresponding to whole chromosome arms in Atlantic salmon [15, 14], represented by single copy genes in grayling as a result of assembly collapse. This strongly suggests that sequences are in fact present as near-identical duplicated regions in the grayling genome. The larger chromosome arm-sized regions still being virtually indistinguishable at the sequence level (10% in total, i.e. blue ribbons in Figure 2b) 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 [46, 14]. The difference we observe in the rediploidization history is thus likely linked to the distinctly different chromosome evolution in Atlantic salmon and grayling (Supplementary Figure S1) [47].

LORe regions also showed a strong enrichment of specific-specific conservation of tissue- specific expression pattern in EEF2 (Table 2 and Supplementary Table S5), as expected under lineage-specific rediploidization and subsequent regulatory divergence. However, we also find a significant proportion (~5%) of EEF2 in AORe regions of the genome. This observation is more difficult to explain, but it is likely a real biological observation as the coding- and promoter sequence evolution analyses (Supplementary figure S9) support the validity of the EEF2 category in AORe regions. Possible explanations for this observation could be a result of non-homologous gene conversion (e.g. [48]) or alternatively, rare local lineage-specific non-homologous recombination events outside the LORe regions. Evolution of anadromy, the ability to migrate between fresh- and seawater, is a fundamental difference in life history strategies between Atlantic salmon and European grayling. Ohnologs of CFTR, a key gene regulating chloride ion export in gills [49] have been shown to be involved in saltwater adaptation in Atlantic salmon [37]. We found that one of the grayling ohnologs of CFTR has lost gill specificity and has evolved under relaxed purifying selection (Figure 5a and b). Our results suggest that that maintaining two functional CFTR genes could be adaptive trait in anadromous salmonids in that it improves their ability to remove excess chloride ions and maintain ion homeostasis in the sea. Conversely, in non-anadromous species, there is no selective pressure to maintain both CFTR copies, and this has resulted in the return to a single functional CFTR ohnolog copy in grayling.

Conclusion

In summary, we present the draft genome assembly of European grayling using an efficient and cost-effective short read sequencing strategy. We show that this draft assembly is very valuable for comparative studies in salmonids. Our comparative genome and transcriptome analysis between Atlantic salmon and grayling provides novel insights into evolutionary fates of ohnologs subsequent to WGD and into associations between signatures of selection pressures on gene duplicate regulation and the evolution of key traits, including anadromy. 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.

Methods

Sampling and sequencing

A male grayling specimen was sampled outside of its spawning season (October 2012) from the River Glomma at Evenstad, Norway. The fish was humanely sacrificed and various tissue samples were immediately extracted and conserved for later DNA and RNA analysis. Fin clips were stored on 96% ethanol for DNA sequencing. Tissues from muscle, gonad, liver, head kidney, spleen, brain, eye, gill and 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 (150 bp read length) and mate pair libraries of insert size 3kb and 6 kb (100bp read length) were sequenced using the Illumina HiSeq2000 platform (Table S1). 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 was carried out using Illumina TruSeq RNA Preparation kit on Illumina HiSeq2000 (Table S2). All the library preparation and sequencing was performed at the McGill University and 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) [50]. A *de novo assembly* was generated with Allpaths-LG (release R48777) [26] using the 180bp paired-end library and the mate pair (3kb and 6kb) libraries. Assembly polishing was carried out using pilon (version 1.9) [27]. 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 ID: NC_012928), followed by a variant calling step using Genome Analysis Toolkit (GATK) (version 3.4-46) [51]. 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 [14] using megablast (E-value threshold 1e-250). Stringent filtering of the aligned scaffolds (representing 1.3 Gbp of the 1.4 Gbp assembly) identified 13 likely chimeric scaffolds mapping to two or more salmon chromosomes (Supplementary File 1), which were then selectively 'broken' between, apparently, incorrectly linked contigs.

Transcriptome assembly

The RNAseq 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) [52] 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 [53], 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) [54]. Cufflinks (version 2.1.1) [54, 55] and TransDecoder [56] 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 (version2.31.8), 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) [57] and SNAP (version 20131129) [58] trained on core eukaryotic gene dataset (CEGMA). The first round of MAKER was then run using the thus generated *ab initio* models, with the UniProt database as the protein evidence, the *de novo* identified repeat library and EST evidences from the transcriptomes assembled using *de novo* and the reference guided approaches, along with the transcript sequences from the recent Atlantic salmon annotation [14]. The second pass involved additional data from training AUGUSTUS [59] 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) [60]. Using the MAKER standard filtering approach, the resulting set of genes were first filtered using the threshold of AED (Annotation Edit Distance), retaining gene models with AED score less than 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 [61] and BUSCO [62].

Analysis of orthologous groups

We used orthofinder (version 0.2.8, e-value threshold at 1e-05) [63] to identified 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*, *Oryzias latipes*), and two mammalian outgroups (*Homo sapiens*, *Mus musculus*). Rainbow trout protein annotations were taken from `https://www.genoscope.cns.fr/trout/`. Atlantic salmon, *Esox lucius* data were downloaded from NCBI ftp server (`ftp://ftp.ncbi.nih.gov/genomes/`, release 100). The transcriptome data for coho salmon was obtained from NCBI (GDQG00000000.1) and translated using TransDecoder. All other annotations were downloaded from ENSEMBL.

Each set of orthogroup proteins were then aligned using MAFFT(v7.130) [64] using default settings and the resulting alignments were then used to infer maximum likelihood gene trees using FastTree (v2.1.8) [65] (Figure 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 [66] with two monophyletic salmonid tips as well as non-salmonid outgroups resulting in a final set of 20,342 gene trees. In total, 31,291 grayling genes were assigned to a clan (Figure 1 and Supplementary Figure S2). We then identified homoelogy in the Atlantic salmon genome by integrating all-vs-all protein BLAST alignments with a priori information of Ss4R synteny as described in Lien et al. 2016 [14]. Using the homeology information, we inferred a set of high confidence ohnologs originating from Ss4R. The clans were grouped based on the gene tree topology into duplicates representing LORe and those with ancestrally diverged duplicates. The LORe regions were further categorized into two (duplicated or collapsed) based on the number of corresponding *T.thymallus* orthologs. This data was plotted on Atlantic salmon chromosomes using circos plot generated using OmicCircos (`https://bioconductor.org/packages/release/bioc/html/OmicCircos.html`).

Expression divergence and conservation

The grayling RNA-seq reads from each of the eight tissues (liver, muscle, spleen, heart, head kidney, eye, brain, 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 [67]. The CPM value (counts per million), here used as a proxy for expression, was then calculated using *edgeR* [68]. Similar CPM datasets were obtained from Atlantic salmon RNA-seq data reported in Lien et al [14].

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, 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, referred to as ohnolog-tetrads. 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 EEF categories (see Table 2). 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 [69] for all the genes in ohnolog-tetrads, where a value approaching 1 indicates higher tissue specificity while 0 indicates ubiquitous expression.

Expression comparison in liver

Utilizing independant liver tissue samples, we compared conservation and divergence in liver tissue gene expression among Ss4R duplicates with their ohnolog-tetrad EEF category. The liver samples from four grayling individuals were sampled in the river Gudbrandsdalslågen. 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 100bp single-end read libraries were generated for two individuals and sequenced using using the Illumina HiSeq4000 platform. For the other two individuals, 150bp paired-end read libraries were generated and sequenced using the Illumina HiSeq2500 platform. RNA-seq data for an additional 8 Atlantic salmon liver tissue samples was obtained from a feeding experiment [70]. Pre-smolt salmon were raised on fish oil based diets under freshwater conditions.

The RNA-seq read data was quality processed using CutAdapt [50] before alignment to grayling or Atlantic salmon (ICSASG_v2, [14]) genomes respectively using STAR [54]. RSEM [71] expected counts were generated for gene features. EdgeR [68] 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 grayling and Atlantic salmon Ss4R duplicates in an 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 [72]. The *seqinr* R package (`http://seqinr.r-forge.r-project.org/`) was used to calculate pairwise dN and dS values for all sequences in each alignment using the *"kaks"* function. For in-depth analyses of branch specific sequence evolution of the CFTR genes, we used the codeml model in PAML (version 4.7a) [73]. To assess if sequences in the CFTR gene tree evolved under similar selection pressure we contrasted a fixed dN/dS 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 dN/dS values were extracted from the ML results for the free ratios model.

The two Pacific salmon genes in the CFTR tree (Figure 5) correspond to a gene from Rainbow trout and another from Coho salmon. A blat 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 is 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 factors binding sites

A total of 13544 metazoan transcription factor protein sequences together with their DNA motifs in position specific scoring Matrix (PSSM) 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 (-1000bp/+200bp from TSS) were extracted. A Markov model of 1-order was created from the entire set of FASTA file of the DNA sequences of the upstream promoters region using the fasta-get-markov application obtained from MEME Suite [74]. The background model of Markov 1-order was used to estimate the p-values of match scores obtained from frequency matrix conversion to a log-odds score matrix.

We performed a genome-wide transcription factors binding sites prediction in Atlantic salmon genome using the PSSM collection. Finding Individual Motif Occurrences (FIMO) [75] tool from MEME Suite was used to scan the sequences (p-value = 0.0001 and FDR = 0.2). Similarity between Atlantic salmon ohnologs was done using Jaccard coefficient. The promoter *Jaccard coefficient* is defined as;

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

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

Gene Ontology (GO) analysis

The gene ontology term (GO) enrichment analysis was performed using *elim* algorithm of topGO R package (`http://www.bioconductor.org/packages/2.12/bioc/html/topGO.html`), with a significance threshold of 0.05 against the reference set of all Ss4R duplicates. GO terms were assigned to salmon genes using Blast2GO[76].

Data availability

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`. Liver expression data will be made publicly available upon acceptance.

Supplementary Files

1. *SupplementaryFile1_AssemblyValidation.xlsx*
A list of scaffolds 13 scaffolds that were ”broken” based on comparison with Atlantic salmon chromosomes.
2. *SupplementaryFile2_GOtests_expression_divergence.xlsx*
GO enrichment analysis table for each of the EEFs

Acknowledgements

This research 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 KSJ. SRS 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 for helpful discussions. We greatly appreciate Daniel J. Macqueen, Marine S. Brieuc and Monica H. Solbakken for critical reading of the manuscript.

Author’s contributions

KSJ, LAV, SJ and SL conceived and planned the project and generation of the data. SRS and SV performed all the analyses with help from AJN and OKT. Differential expression analysis on the liver dataset was performed by GBG and TRH, and the promoter motif analysis was prepared by TDM. SRS, SV and AJN drafted the manuscript. All authors read and commented on the manuscript.

Competing interests

The authors declare that they have no competing interests.

References

1. Van de Peer, Y., Maere, S., Meyer, A.: The evolutionary significance of ancient genome duplications. Nat. Rev. Genet. **10**(10), 725–732 (2009)
2. Ohno, S.: Evolution by Gene Duplication, (1970)
3. Lynch, M., Conery, J.S.: The evolutionary fate and consequences of duplicate genes. Science **290**(5494), 1151–1155 (2000)
4. Zhang, J.: Evolution by gene duplication: an update. Trends Ecol. Evol. **18**(6), 292–298 (2003)
5. Conant, G.C., Wolfe, K.H.: Turning a hobby into a job: how duplicated genes find new functions. Nat. Rev. Genet. **9**(12), 938–950 (2008)
6. Osborn, T.C., Pires, J.C., Birchler, J.A., Auger, D.L., Chen, Z.J., Lee, H.-S., Comai, L., Madlung, A., Doerge, R.W., Colot, V., Martienssen, R.A.: Understanding mechanisms of novel gene expression in polyploids. Trends Genet. **19**(3), 141–147 (2003)
7. Carroll, S.B.: Endless forms: the evolution of gene regulation and morphological diversity. Cell **101**(6), 577–580 (2000)
8. Wray, G.A.: The evolution of transcriptional regulation in eukaryotes. Mol. Biol. Evol. **20**(9), 1377–1419 (2003)
9. Sémon, M., Wolfe, K.H.: Preferential subfunctionalization of slow-evolving genes after allopolyploidization in xenopus laevis. Proc. Natl. Acad. Sci. U. S. A. **105**(24), 8333–8338 (2008)
10. Kassahn, K.S., Dang, V.T., Wilkins, S.J., Perkins, A.C., Ragan, M.A.: Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates. Genome Res. **19**(8), 1404–1418 (2009)
11. 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., Quillet, E., Guyomard, R., Galiana, D., Bobe, J., Volff, J.-N., Genêt, C., Wincker, P., Jaillon, O., Roest Crollius, H., Guiguen, Y.: The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nat. Commun. **5**, 3657 (2014)
12. Li, J.-T., Hou, G.-Y., Kong, X.-F., Li, C.-Y., Zeng, J.-M., Li, H.-D., Xiao, G.-B., Li, X.-M., Sun, X.-W.: The fate of recent duplicated genes following a fourth-round whole genome duplication in a tetraploid fish, common carp (cyprinus carpio). Sci. Rep. **5**, 8199 (2015)
13. Acharya, D., Ghosh, T.C.: Global analysis of human duplicated genes reveals the relative importance of

whole-genome duplicates originated in the early vertebrate evolution. BMC Genomics **17**, 71 (2016)

14. 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., Vigeland, M.D., Caler, L., Grimholt, U., Jentoft, S., Våge, D.I., de Jong, P., Moen, T., Baranski, M., Palti, Y., Smith, D.R., Yorke, J.A., Nederbragt, A.J., Tooming-Klunderud, A., Jakobsen, K.S., Jiang, X., Fan, D., Hu, Y., Liberles, D.A., Vidal, R., Iturra, P., Jones, S.J.M., Jonassen, I., Maass, A., Omholt, S.W., Davidson, W.S.: The atlantic salmon genome provides insights into rediploidization. Nature **533**(7602), 200–205 (2016)
15. 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., Macqueen, D.J.: Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification. Genome Biol. **18**(1), 111 (2017)
16. Hermansen, R.A., Hvidsten, T.R., Sandve, S.R., Liberles, D.A.: Extracting functional trends from whole genome duplication events using comparative genomics. Biol. Proced. Online **18**(1) (2016)
17. Alexandrou, M.A., Swartz, B.A., Matzke, N.J., Oakley, T.H.: Genome duplication and multiple evolutionary origins of complex migratory behavior in salmonidae. Mol. Phylogenet. Evol. **69**(3), 514–523 (2013)
18. Macqueen, D.J., Johnston, I.A.: 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 (2014)
19. Limborg, M.T., Seeb, L.W., Seeb, J.E.: Sorting duplicated loci disentangles complexities of polyploid genomes masked by genotyping by sequencing. Mol. Ecol. **25**(10), 2117–2129 (2016)
20. Wolfe, K.H.: Yesterday's polyploids and the mystery of diploidization. Nat. Rev. Genet. **2**(5), 333–341 (2001)
21. Hendry, A.P., Stearns, S.C.: Evolution Illuminated: Salmon and Their Relatives. Oxford University Press on Demand, ??? (2004)
22. Nygren, A., Nilsson, B., Jahnke, M.: Cytological studies in thymallus thymallus and coregonus albula. Hereditas **67**(2), 269–274 (1971)
23. Phillips, R., Ráb, P.: Chromosome evolution in the salmonidae (pisces): an update. Biol. Rev. Camb. Philos. Soc. **76**(1), 1–25 (2001)
24. Hartley, S.E.: THE CHROMOSOMES OF SALMONID FISHES. Biol. Rev. Camb. Philos. Soc. **62**(3), 197–214 (1987)
25. Ocalewicz, K., Furgala-Selezniow, G., Szmyt, M., Lisboa, R., Kucinski, M., Lejk, A.M., Jankun, M.: Pericentromeric location of the telomeric DNA sequences on the european grayling chromosomes. Genetica **141**(10-12), 409–416 (2013)
26. Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., Berlin, A.M., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E.S., Jaffe, D.B.: High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. U. S. A. **108**(4), 1513–1518 (2011)
27. Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., Earl, A.M.: Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One **9**(11), 112963 (2014)
28. UniProt Consortium: UniProt: a hub for protein information. Nucleic Acids Res. **43**(Database issue), 204–12 (2015)
29. Li, H.: Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM (2013). 1303.3997
30. Garrison, E., Marth, G.: Haplotype-based variant detection from short-read sequencing (2012). 1207.3907
31. Gu, X., Su, Z.: Tissue-driven hypothesis of genomic evolution and sequence-expression correlations. Proc. Natl. Acad. Sci. U. S. A. **104**(8), 2779–2784 (2007)
32. Freeling, M., Thomas, B.C.: Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res. **16**(7), 805–814 (2006)
33. Sémon, M., Wolfe, K.H.: Consequences of genome duplication. Curr. Opin. Genet. Dev. **17**(6), 505–512 (2007)
34. Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., Jensen, L.J., von Mering, C.: The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res. **45**(D1), 362–368 (2017)
35. Pasquale, E.B.: Eph-ephrin bidirectional signaling in physiology and disease. Cell **133**(1), 38–52 (2008)
36. Sandve, S.R., Rohlfs, R.V., Hvidsten, T.R.: Subfunctionalization versus neofunctionalization after whole-genome duplication (2017)
37. Nilsen, T.O., Ebbesson, L.O.E., Madsen, S.S., McCormick, S.D., Andersson, E., Th. Bjornsson, B., Prunet, P., Stefansson, S.O.: Differential expression of gill na ,k -ATPase - and -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 (2007)
38. Craig, J.F.: A short review of pike ecology. Hydrobiologia **601**(1), 5–16 (2008)
39. Haase, D., Roth, O., Kalbe, M., Schmiedeskamp, G., Scharsack, J.P., Rosenstiel, P., Reusch, T.B.H.: Absence of major histocompatibility complex class II mediated immunity in pipefish, syngnathus typhle: evidence from deep transcriptome sequencing. Biol. Lett. **9**(2), 20130044 (2013)
40. Solbakken, M.H., Voje, K.L., Jakobsen, K.S., Jentoft, S.: Linking species habitat and past palaeoclimatic events to evolution of the teleost innate immune system. Proc. Biol. Sci. **284**(1853) (2017)
41. Carmona-Antoñanzas, G., Tocher, D.R., Taggart, J.B., Leaver, M.J.: An evolutionary perspective on elovl5 fatty acid elongase: comparison of northern pike and duplicated paralogs from atlantic salmon. BMC Evol. Biol. **13**, 85 (2013)
42. Zheng-Bradley, X., Rung, J., Parkinson, H., Brazma, A.: Large scale comparison of global gene expression patterns in human and mouse. Genome Biol. **11**(12), 124 (2010)
43. Chan, E.T., Quon, G.T., Chua, G., Babak, T., Trochesset, M., Zirngibl, R.A., Aubin, J., Ratcliffe, M.J.H., Wilde, A., Brudno, M., Morris, Q.D., Hughes, T.R.: Conservation of core gene expression in vertebrate tissues. J. Biol. **8**(3), 33 (2009)
44. Khaitovich, P., Hellmann, I., Enard, W., Nowick, K., Leinweber, M., Franz, H., Weiss, G., Lachmann, M., Pääbo, S.: Parallel patterns of evolution in the genomes and transcriptomes of humans and

chimpanzees. Science **309**(5742), 1850–1854 (2005)

45. Roux, J., Liu, J., Robinson-Rechavi, M.: 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 (2017)
46. Comai, L.: The advantages and disadvantages of being polyploid. Nat. Rev. Genet. **6**(11), 836–846 (2005)
47. Qumsiyeh, M.B.: Evolution of number and morphology of mammalian chromosomes. J. Hered. **85**(6), 455–465 (1994)
48. Hastings, P.J.: Mechanisms of ectopic gene conversion. Genes **1**(3), 427–439 (2010)
49. Marshall, W.S., Singer, T.D.: Cystic fibrosis transmembrane conductance regulator in teleost fish. Biochimica et Biophysica Acta (BBA) - Biomembranes **1566**(1-2), 16–27 (2002)
50. Martin, M.: Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal **17**(1), 10 (2011)
51. Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K.V., Altshuler, D., Gabriel, S., DePristo, M.A.: From FastQ data to High-Confidence variant calls: The genome analysis toolkit best practices pipeline. In: Current Protocols in Bioinformatics, pp. 11–101111033 (2013)
52. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A.: Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. **29**(7), 644–652 (2011)
53. Faust, G.G., Hall, I.M.: YAHA: fast and flexible long-read alignment with optimal breakpoint detection. Bioinformatics **28**(19), 2417–2424 (2012)
54. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R.: STAR: ultrafast universal RNA-seq aligner. Bioinformatics **29**(1), 15–21 (2013)
55. Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., Pachter, L.: Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. **28**(5), 511–515 (2010)
56. Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., MacManes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., LeDuc, R.D., Friedman, N., Regev, A.: De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. Nat. Protoc. **8**(8), 1494–1512 (2013)
57. Lomsadze, A.: Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. **33**(20), 6494–6506 (2005)
58. Korf, I.: Gene finding in novel genomes. BMC Bioinformatics **5**(1), 59 (2004)
59. Stanke, M., Diekhans, M., Baertsch, R., Haussler, D.: Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics **24**(5), 637–644 (2008)
60. Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., Lopez, R.: InterProScan: protein domains identifier. Nucleic Acids Res. **33**(Web Server issue), 116–20 (2005)
61. Parra, G., Bradnam, K., Korf, I.: CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics **23**(9), 1061–1067 (2007)
62. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M.: BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics **31**(19), 3210–3212 (2015)
63. Emms, D.M., Kelly, S.: OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. **16**, 157 (2015)
64. Katoh, K., Misawa, K., Kuma, K.-I., Miyata, T.: MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. Nucleic Acids Res. **30**(14), 3059–3066 (2002)
65. Price, M.N., Dehal, P.S., Arkin, A.P.: FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One **5**(3), 9490 (2010)
66. Wilkinson, M., McInerney, J.O., Hirt, R.P., Foster, P.G., Embley, T.M.: Of clades and clans: terms for phylogenetic relationships in unrooted trees. Trends Ecol. Evol. **22**(3), 114–115 (2007)
67. Anders, S., Pyl, P.T., Huber, W.: HTSeq–a python framework to work with high-throughput sequencing data. Bioinformatics **31**(2), 166–169 (2015)
68. Robinson, M.D., McCarthy, D.J., Smyth, G.K.: edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**(1), 139–140 (2010)
69. Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A., Horn-Saban, S., Safran, M., Domany, E., Lancet, D., Shmueli, O.: Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics **21**(5), 650–659 (2005)
70. Gillard, G., Harvey, T.N., Gjuvsland, A., Jin, Y., Thomassen, M., Lien, S., Leaver, M., Torgersen, J.S., Hvidsten, T.R., Vik, J.O., Sandve, S.R.: Life stage associated remodeling of lipid metabolism regulation in the duplicated Atlantic salmon genome (2017)
71. Li, B., Dewey, C.N.: RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics **12**(1), 323 (2011)
72. Suyama, M., Torrents, D., Bork, P.: PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. **34**(Web Server issue), 609–12 (2006)
73. Yang, Z.: PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13**(5), 555–556 (1997)
74. Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., Noble, W.S.: MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. **37**(Web Server issue), 202–8 (2009)
75. Grant, C.E., Bailey, T.L., Noble, W.S.: FIMO: scanning for occurrences of a given motif. Bioinformatics **27**(7), 1017–1018 (2011)
76. Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., Robles, M.: Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics **21**(18), 3674–3676 (2005)

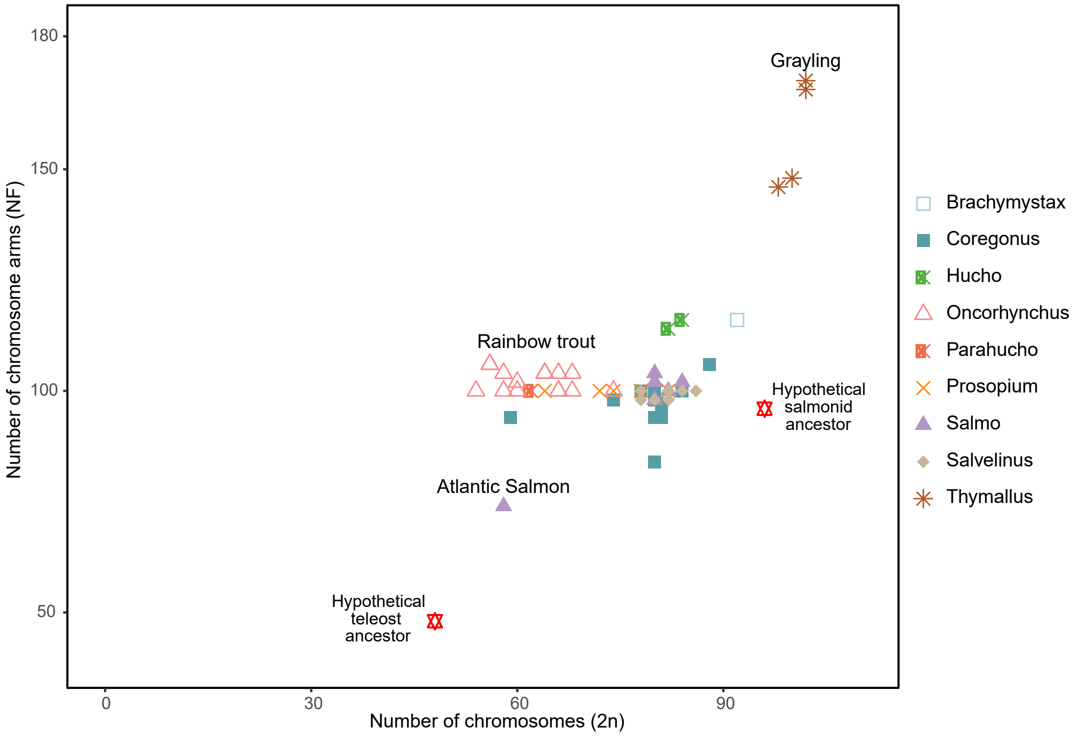

Figure S1 Chromosome evolution in salmonids
Chromosome number (2N) plotted against Number of chromosome arms (NF). Data from (Hartley et al. 1987). Based on the karyotype of most extant teleosts (48-50 acrocentric chromosomes), it is hypothesized that the salmonid ancestor had a karyotype of around 96-100 uni-armed chromosomes (NF 100). While most of the salmonids have a karyotype consisting of chromosome number of 52 to 102 and NF of 72-170, Atlantic salmon and grayling seem to be the exceptions on the opposite extremes. It has also been seen that the bi-armed metacentric chromosomes in grayling are much smaller than those in other salmonids. Thus, it has been hypothesized that, while most salmonids have reduced the chromosome number and retained NF close to the ancestral karyotype through translocations and fusions, the grayling karyotype has evolved through inversions (Phillips and Ráb 2007; Ocalewicz et al. 2013).

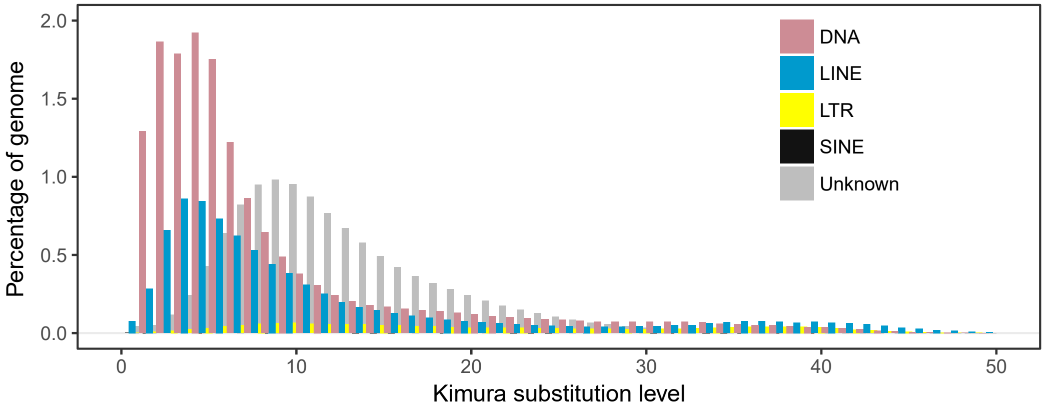

Figure S2 Repeat landscape of grayling genome based on Kimura distance.
X-axis represents divergence from repeat consensus sequence and y-axis represents the proportion of the transposable element family in the genome (where LTR stands for long terminal repeats, LINE represents long interspersed nuclear elements and SINE stands for short interspersed nuclear elements).

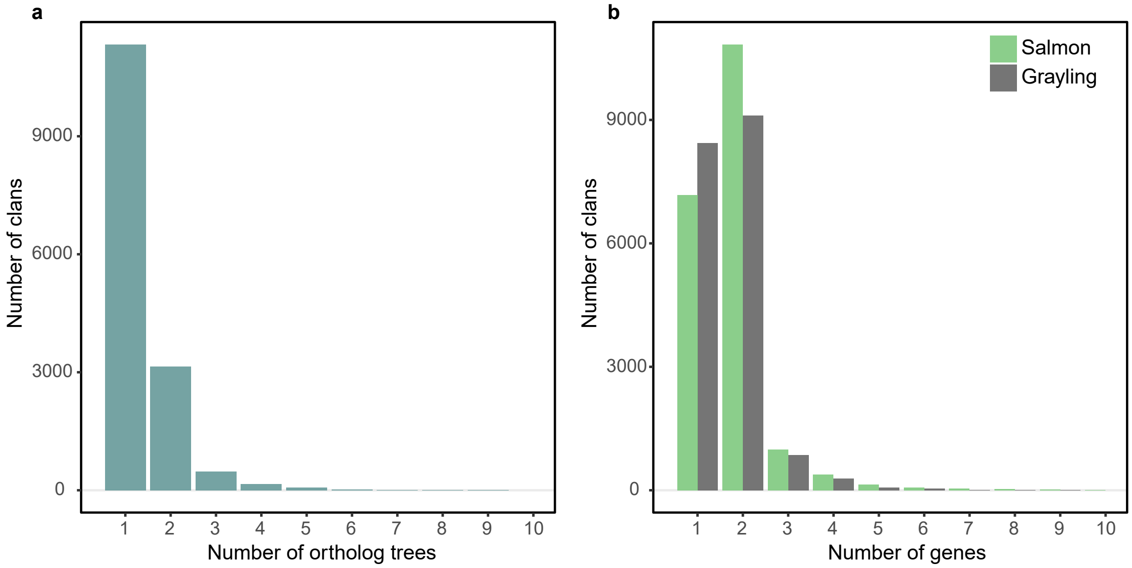

Figure S3 a) Distribution of clans per ortholog tree. b) Number of Atlantic salmon and grayling genes per clan.

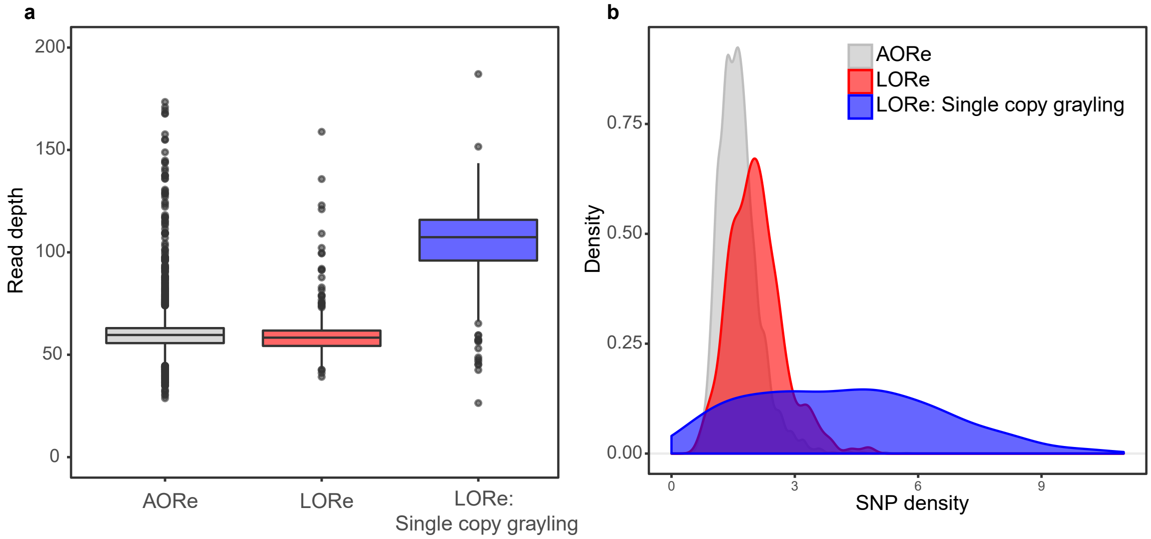

Figure S4 Distribution of (a) mapped read depth and (b) SNP density per Kb across all Ss4R duplicates in grayling grouped by the ohnolog resolution models.

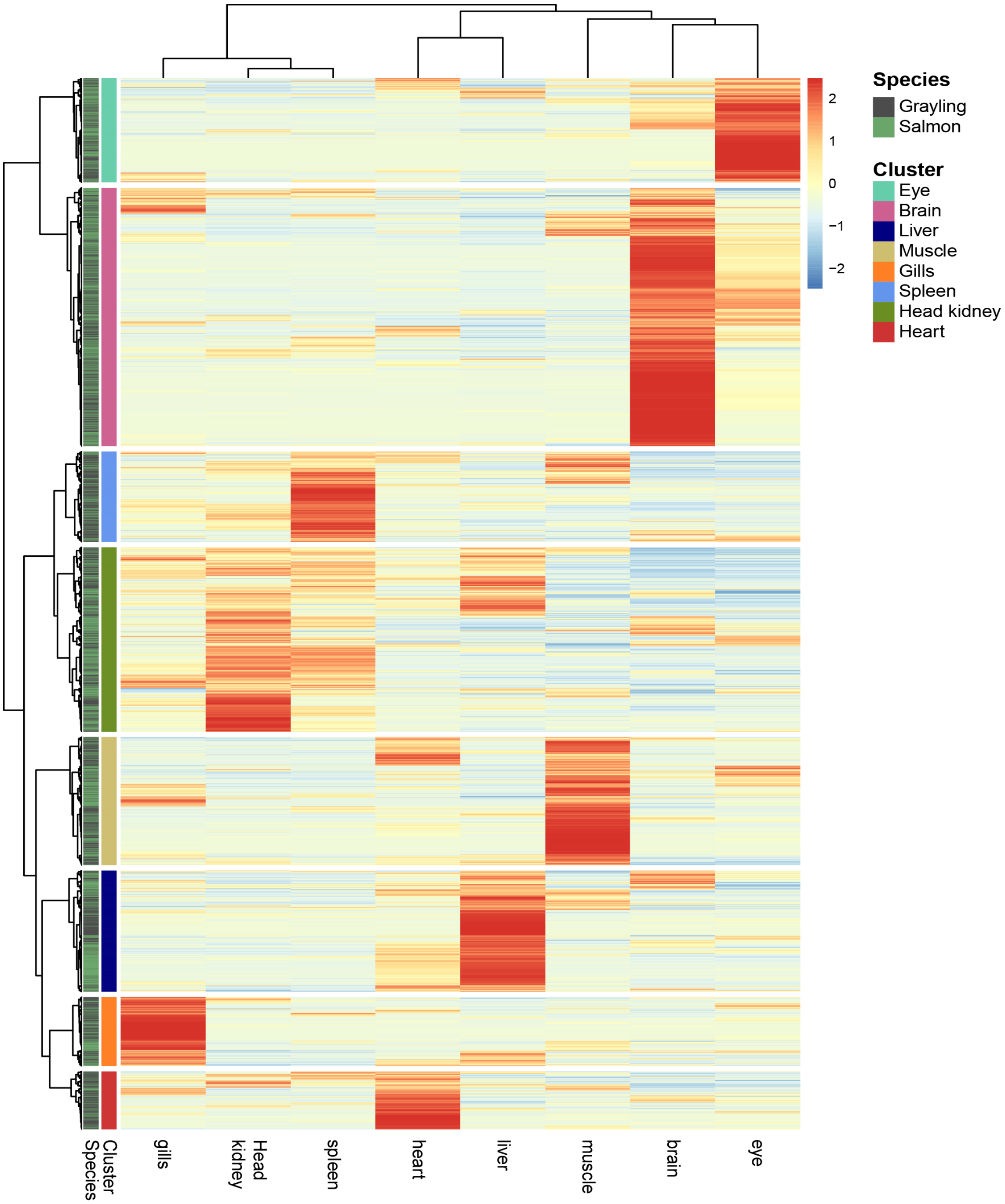

Figure S5 Heatmap of tissue expression clusters from Atlantic salmon and grayling. Tissue expression profile of ohnologs from Atlantic salmon and grayling using hierarchical clustering. The color scale of the heatmap corresponds to the relative abundance of the transcript across all the tissues within the two species. The first vertical bar ('Cluster') represents the 8 distinct 'tissue-specific' clusters. The 'Species' bar represents the respective species corresponding to the gene.

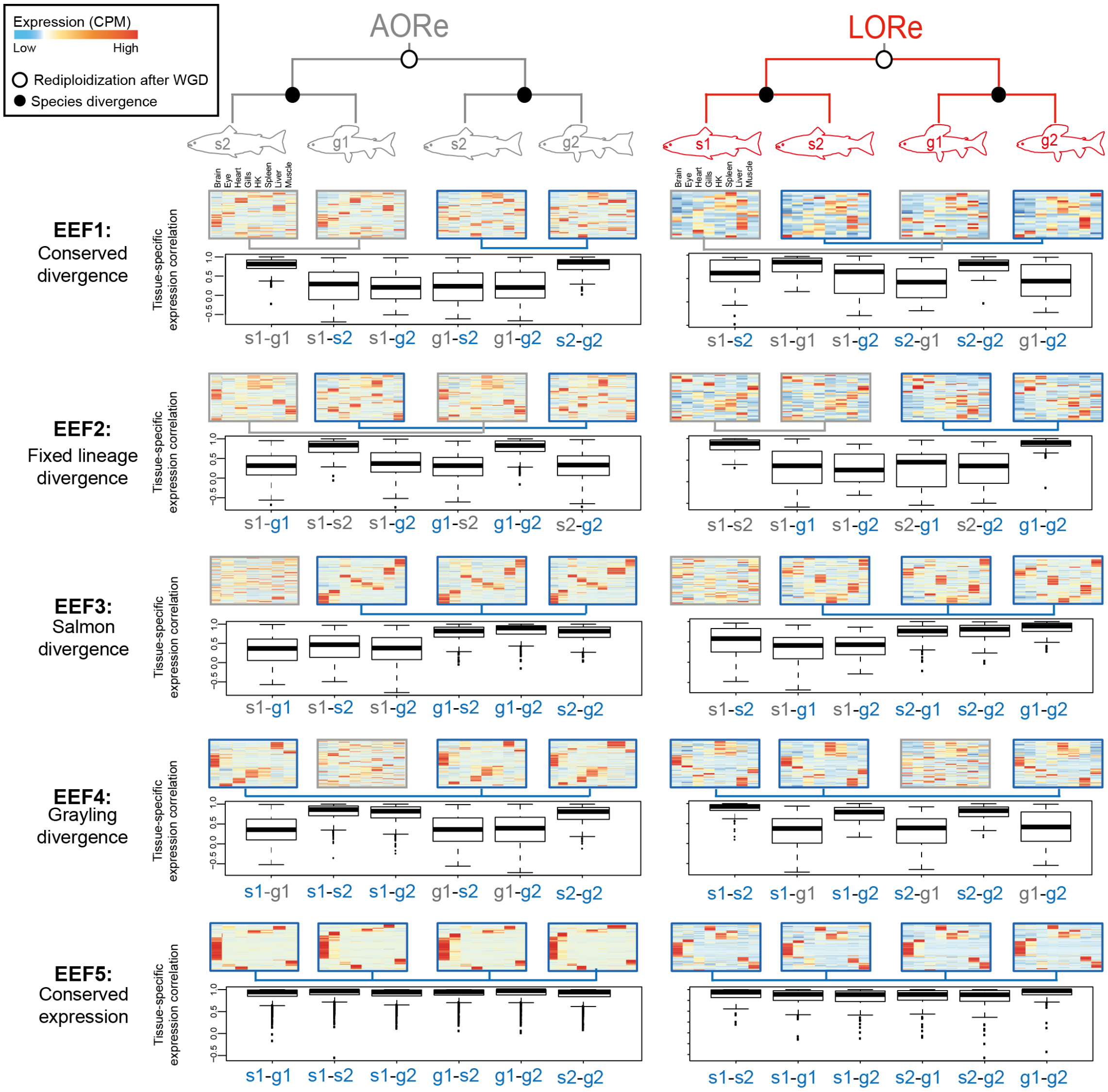

Figure S6 Expression pattern evolution of ohnologs in LORe and AORe regions.
An extension of Figure 3 with heatmaps showing 5 expression evolution fates (EEFs, see Table 2) reflecting differential selection on tissue expression regulation after Ss4R WGD over genes with LORe and AORe histories. The color scale of the heatmaps correspond to the relative abundance of the transcript across all the tissues within the two species, in terms of counts per million (CPM). Each row across the four heatmaps represents one ortholog group of an ohnolog-tetrad. Connecting lines below heatmaps indicate duplicates belonging to same tissue clusters (conserved). Below the heatmaps are the boxplots representing expression correlation between and within duplicates in Atlantic salmon and grayling. The ohnologs in Atlantic salmon and grayling are represented as s1, s2 and g1 and g2 respectively.

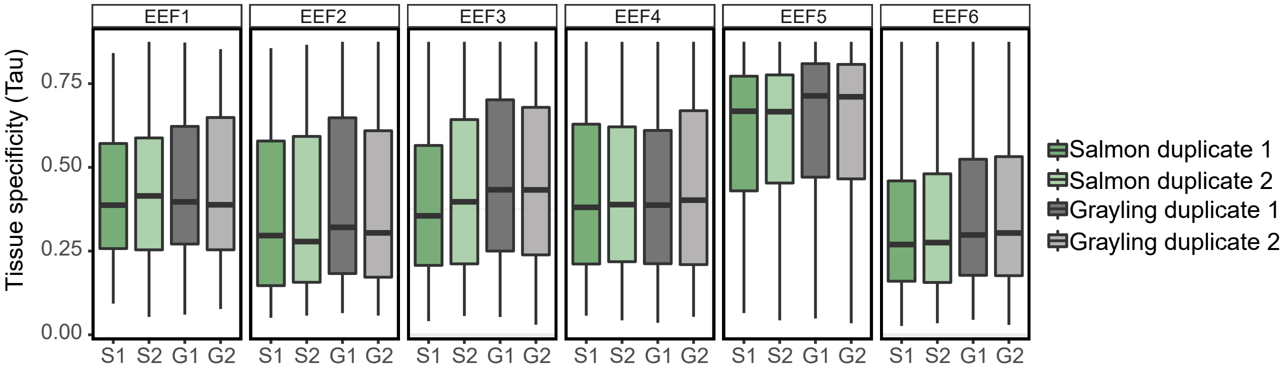

Figure S7 Tissue-specificity of ohnologs. Overall tissue specificity (Tau) distribution for each of the EEFs. The ohnologs in Atlantic salmon and grayling are represented as S1, S2 and G1 and G2 respectively.

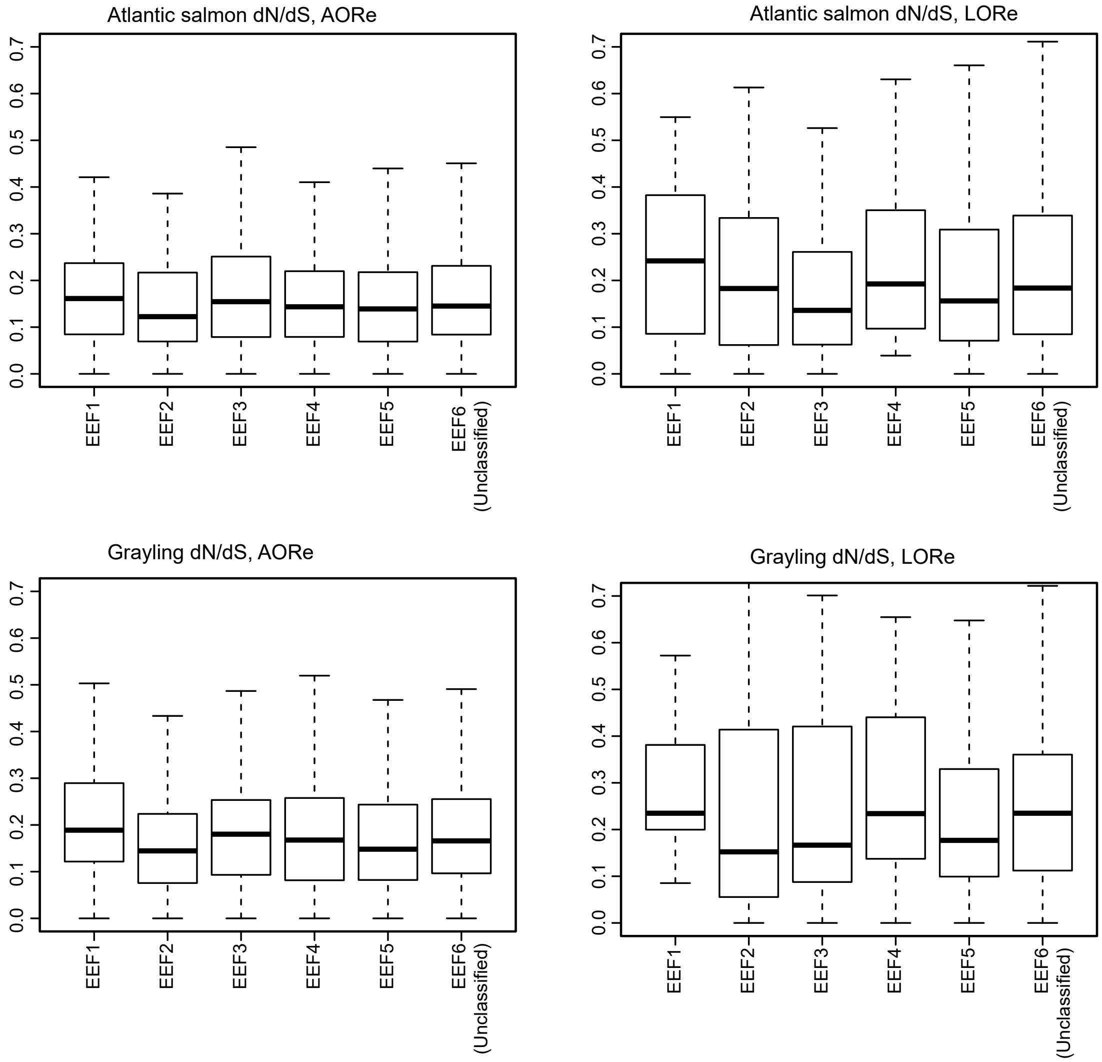

Figure S8 Coding sequence evolution. Distribution of dN/dS, representing coding sequence evolution, across different EEF categories across the LORe and AORe regions in Atlantic salmon and grayling.

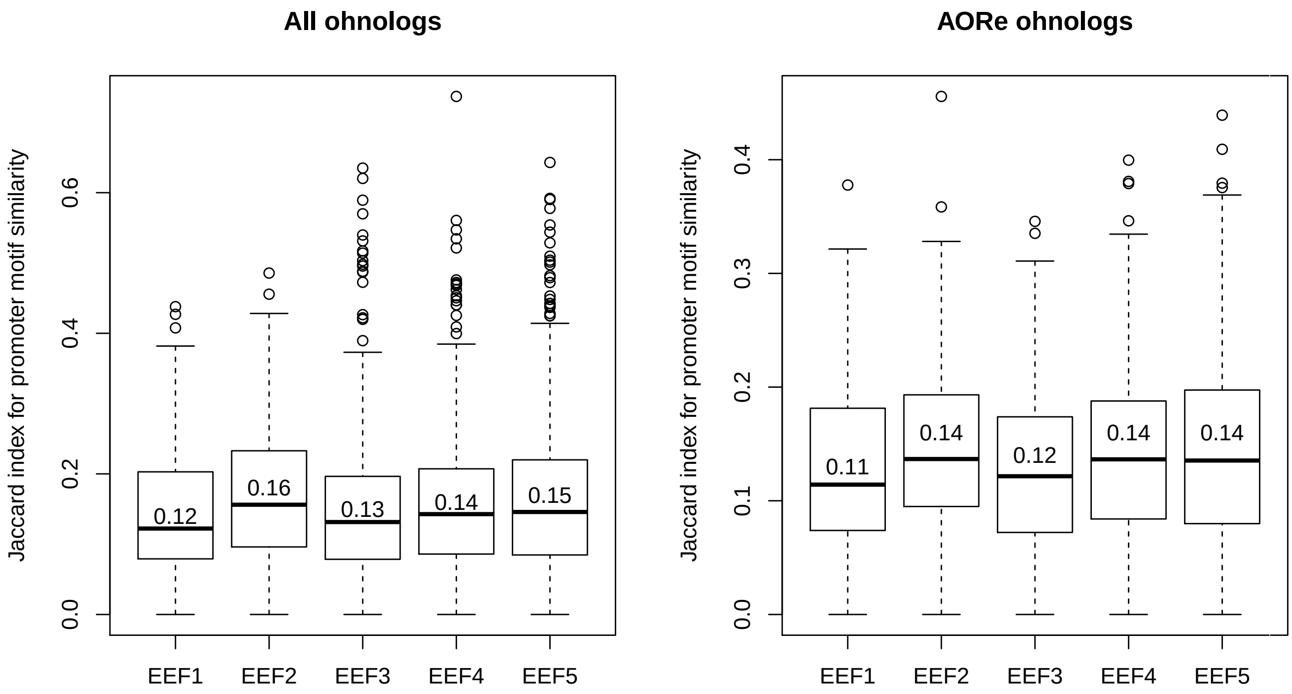

Figure S9 Salmon ohnolog promoter motif conservation. Boxplots of the similarity of promoter motif presence between ohnolog pairs in salmon in, AORe regions, is shown as Jaccard Index for the EEF1-5. Median Jaccard Indexes are indicated within boxes. The classes EEF1 and EEF3 , where tissue regulation are diverged among salmon ohnologs, have significantly lower rank sum means in a Wilcoxon test compared to the ohnologs in EEF2, EEF4, and EEF5 (p=0.02-0.002) which have conserved expression in salmon.

Table S1 Sequencing libraries and data produced.

Insert size	**Read length (bp)**	**Number of bases**	**Coverage***
180bp	150	103,520,976,600	57.51
3kb	100	85,304,163,600	47.39
3kb	100	42,464,706,800	23.59
6kb	100	90,372,684,400	50.21

* Based on a genome size estimate of 1.8Gbp

Table S2 Summary of RNAseq data generated

Tissue	**Number of bases**
Liver	17,086,217,700
Muscle	26,352,566,700
Spleen	27,942,424,800
Heart	28,336,371,600
Headkidney	23,154,448,800
Gonad	19,270,169,100
Eye	19,541,207,700
Brain	21,784,344,900
Gills	25,275,737,700

Table S3 Repeats and transposable elements

	Number of elements	Length occupied (bp)	Percentage of sequences
SINES	69,830	8,566,799	0.58 %
ALU	0	0	0.0 %
MIRs	216	8,430	0.0 %
LINES	316,144	117,430,749	8.0 %
LINE1	11,346	4,292,742	0.29 %
LINE2	120,455	40,284,357	2.74 %
L3/CR1	2,850	463,243	0.03 %
LTR elements	86,307	22,365,017	1.52 %
ERVL	91	13,152	0.0 %
ERVL-MaLRs	8	488	0.0 %
ERV_classI	9,443	2,223,741	0.15 %
ERV_classII	3,252	195,686	0.01 %
DNA elements	830,457	235,731,278	16.05 %
hAT-Charlie	11,902	3,316,142	0.23 %
TcMar-Tigger	141	41,487	0.0 %
Unclassified	777,695	167,400,996	11.40 %
Total interspersed repeats		551,494,839	37.55 %
Small RNA	1,867	150,973	0.01 %
Satellites	18,039	2,929,358	0.20 %
Simple repeats	599,983	43,271,286	2.95 %
Low complexity	64,487	4,726,681	0.32 %

Table S4 Distribution of tissue-dominated expression clusters in tetrads of different regulatory evolution categories.
Red cells represents genes in tissue expression clusters that were disproportionately represented compared to ‘all’ tetrads.

Expression evolution	**Brain**	**Eye**	**Gills**	**Heart**	**Headkidney**	**Liver**	**Muscle**	**Spleen**
EEF1	🟥 124 (17%)	50 (7%)	🟥 118 (16%)	36 (5%)	124 (17%)	🟥 144 (20%)	68 (9%)	56 (8%)
EEF2	🟥 162 (16%)	112 (11%)	96 (10%)	62 (6%)	148 (15%)	120 (12%)	144 (15%)	🟥 140 (14%)
EEF3	🟥 363 (20%)	150 (8%)	234 (13%)	109 (6%)	205 (12%)	🟥 321 (18%)	🟥 294 (17%)	104 (6%)
EEF4	578 (24%)	173 (7%)	276 (12%)	149 (6%)	🟥 453 (19%)	363 (15%)	🟥 228 (10%)	164 (7%)
EEF5	🟥 1948 (49%)	328 (8%)	404 (10%)	🟥 52 (1%)	🟥 336 (8%)	🟥 368 (9%)	508 (13%)	🟥 56 (1%)
All	4113 (26%)	1357 (9%)	1848 (12%)	996 (6%)	2134 (13%)	2231 (14%)	2057 (13%)	1220 (8%)

Red = Fisher test Bonferreoni corrected p-value <= 0.05

Table S5 Differential expression analysis using the liver expression data. Number of ohnolog tetrads with at least one ohnolog assigned to liver expression cluster, and the number and percentage of those tetrads with significant (FDR adjusted p-value < 0.001) fold change (FC) in liver expression for duplicates of salmon, grayling, and for both species.

Expression evolution	**Ohnolog tetrads with liver clustering**	**Significant FC in salmon**	**Significant FC in grayling**	**Significant FC in both species**
EEF1	63	46 (73%)	47 (75%)	38 (60%)
EEF2	57	25 (44%)	29 (51%)	16 (28%)
EEF3	142	99 (70%)	70 (49%)	51 (36%)
EEF4	152	90 (59%)	114 (75%)	69 (45%)
EEF5	85	49 (58%)	51 (60%)	34 (40%)