The grayling genome reveals selection on gene expression regulation after whole genome duplication

Srinidhi Varadharajan*[1] , Simen R. Sandve*[2‡] , Ole K. Tørresen[1] , Sigbjørn Lien[2] , Leif Asbjørn Vøllestad[1] , Sissel Jentoft[1] , Alexander J. Nederbragt[1,3] 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]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, yet little is known about how selection operates on the resulting gene duplicates. Here, we present a draft genome assembly of a salmonid species, European grayling (*Thymallus thymallus*) and use comparative genomics and transcriptomics to understand evolutionary consequences of WGD in the genome of salmonid ancestor ∼80 million years ago (Ss4R). We find evidence for lineage-specific rates in rediploidization and that ∼60% of the Ss4R ohnologs have experienced different types of non-neutral evolution of tissue-specific gene expression regulation. Distinct selective pressures were associated with tissue type, biological function and selection pressure on protein coding sequence. Finally, our results indicate the role of adaptive divergence of Ss4R duplicates in the evolution of salmonid metabolism and identifies loss of purifying selection on one Ss4R ohnolog encoding a key chloride pump linked to the evolution of anadromy.

Introduction

Whole genome duplication (WGD) has played a vital role in the evolution of vertebrate genome complexity. Two rounds of genome duplication events occurred in the ancestor of all vertebrates (1R and 2R events), a third WGD at the base of the teleosts (3R) 225-333 million years ago (MYA) [1]. Several teleost lineages experienced additional WGD events including the salmonid ancestor 88-103 MYA (Ss4R) [2, 3]. 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 [4]. Hence, sequence divergence or structural rearrangements are indispensable for blocking multivalent formation, suppressing recombination and driving the process of returning to a functional diploid state, a process called rediploidization. As the mutational process is stochastic, the resolution of ohnologs (gene duplicates resulting from WGD) is achieved independently for different duplicated chromosomes and hence occurs at different rates in various genomic regions.

The functional redundancy arising from gene duplication is believed to be a source for the evolution of novel traits and adaptations [2]. Duplicate genes that escape loss or pseudogenization are known to acquire novel regulation and expression divergence [5, 6, 7]. 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 [8]. As changes in gene regulation are known to be important in the evolution of phenotypic diversity and complex trait variation [9, 10], these post-WGD shifts in expression regulation provide a large substrate for adaptive evolution. Several studies have investigated the genome-wide consequences of WGD on gene expression evolution in vertebrates(e.g. [11, 12, 13, 14, 15, 16, 17] and have revealed that a large proportion of gene duplicates have evolved substantial regulatory divergence, and that in most cases one copy retains ancestral-like regulation (consistent with regulatory neo-functionalization). 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 [18], which allows us to distinguish neutral

divergence in biological function from that maintained by purifying selection [19]. Further, confidently identifying gene duplicates retained from paleopolyploidy events like 2R and 3R dating back to >300-500 million years (MY) is challenging.

Salmonids have emerged as a model for studying functional consequences of autopolyploidization in vertebrates, owing to their relatively young WGD event (<100MYA) and ongoing rediploidization [20, 16]. Recent studies on genome evolution subsequent to Ss4R have shown that the rediploidization process has been temporally overlapping with species radiation, resulting in lineage-specific ohnolog resolution (LORe) that may fuel differentiation of genome structure and function [21, 17]. In fact, only 75% of the duplicated ancestral salmonid genome had rediploidized at the time of the basal split in the Salmonidae family ∼60 MYA . Consequently, ∼25% of the Ss4R duplicates have undergone rediploidization independently 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 [22]. The species in the subfamily Salmoninae have fewer and highly derived chromosomes resulting from large-scale chromosomal rearrangements and chromosomal fusions, display extreme phenotypic plasticity, and have evolved the capability of migrating between fresh and saltwater habitats (anadromy) [3]. In contrast, the Thymallinae species (graylings) have a more ancestral genome structure with few or no chromosome fusions [23, 24, 25, 26] (Supplementary Figure S1). Further, grayling species are generally less plastic and have not evolved anadromy. This unique combination of both shared and lineage-specific rediploidization histories and striking difference in genome structure and adaptations provides an intriguing study system for addressing key questions about the evolutionary consequences of WGD.

In order to gain deeper insights into how selection has shaped the evolution of gene duplicates post WGD, we have sequenced, assembled and annotated the draft 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 gain insights into the consequences of lineage-specific rediploidization and genome-wide selective constraints on gene expression regulation. Our analyses of expression patterns across the two duplicated salmonid genomes (grayling and Atlantic salmon) demonstrate that a large fraction of the duplicates originating from Ss4R have experienced purifying selection to maintain ancestral tissue-specific expression regulation. Moreover, widely 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 the 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 [27], followed by assembly correction using Pilon [28], 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 1.9 [25]. To annotate gene structures, we used RNA-seq data from nine tissues extracted from the sequenced individual. 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 data, the *de novo* identified repeats along with the UniProt proteins [29] and Atlantic salmon coding sequences [16] 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 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 25% of the genome of the most recent common salmonid ancestor was still tetraploid when the grayling and Atlantic salmon lineages diverged [16, 17]. To test this hypothesis, we used a phylogenomic approach to characterize rediploidization following Ss4R in grayling.

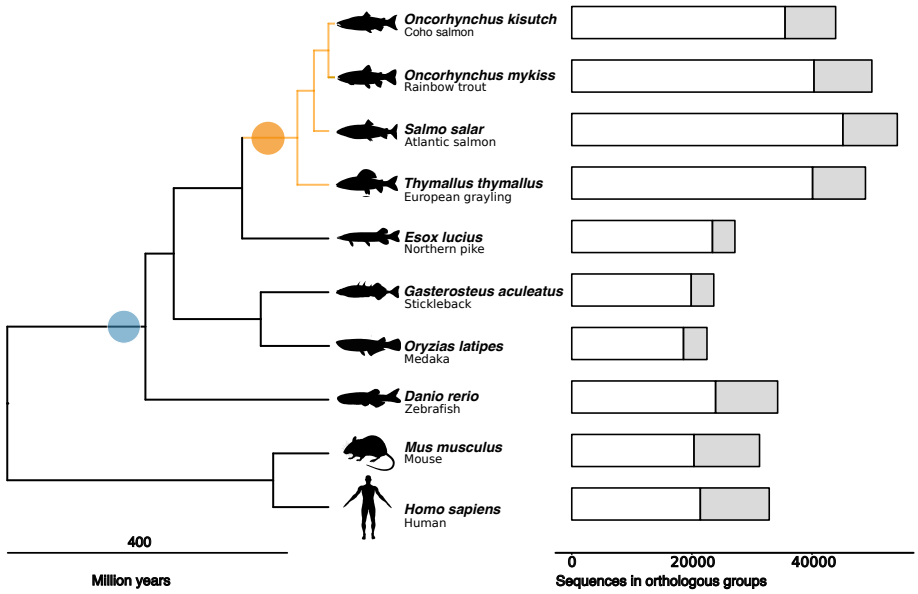

Figure 1 Species and genes in orthologous groups.
Left: phylogenetic relationship of species used for constructing orthologous groups and gene trees. The blue colored circle indicates the 3R-WGD event while the Ss4R event is indicated with an orange colored circle. Right: number of genes assigned to orthologous groups in each of the species used in the analysis. Total size of the bars indicates genes in orthologous groups, with the white portion indicating genes included in the gene trees.

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

We inferred 23,782 orthologous gene groups among 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 subgroups (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 [16]. A synteny-aware blast approach [16] was first used to identify Ss4R duplicate pairs/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 and thus following the topology expected under ancestral ohnolog resolution (henceforth referred to as AORe) (Figure 2a). In total, 3,362 gene trees correspond to LORe regions (2,403 with a single copy in grayling) and 5,113 correspond to an AORe-like topology. This data was cross-checked

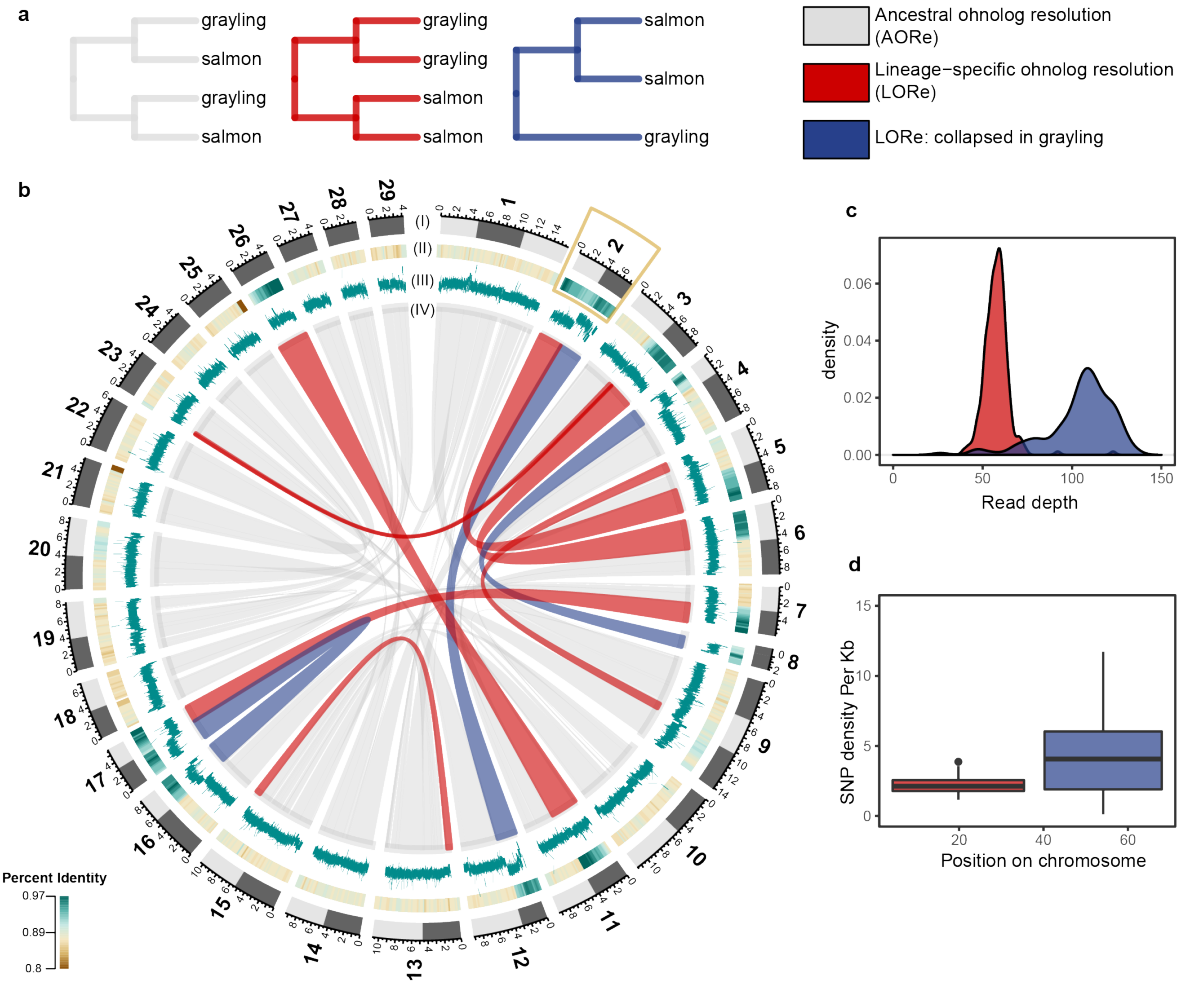

Figure 2 Rediploidization in grayling genome.
a) Gene tree topologies corresponding to corresponding to the different models of ohnolog resolution. b) Circos plot (generated using *OmicCircos* package in R): Outer track (I) represents the 29 chromosomes of Atlantic salmon with chromosome arms indicated using light and dark grey. (II) Percent identity between duplicated genomic regions in Atlantic salmon with darker green representing higher percent identity. (III) Average number of reads mapped to grayling genes in the corresponding regions. (IV) The grey ribbon links represent the ancestrally diverged gene duplicate pairs (AORe), while the red ribbon links represent the LORe duplicate pairs and the blue ribbon links 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 in yellow in (b)).

with the LORe coordinates suggested by Robertson et al [17] and cases that did not conform were omitted from further analyses. This final set consisted of 5,340 gene trees containing Ss4R duplicates from both species (4603 AORe, 737 LORe), in addition to 482 ortholog sets containing Ss4R duplicates in Atlantic salmon but not 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 2). In Atlantic salmon, several homeologous 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 [16, 17](indicated in Figure 2b as red and blue ribbons). Interestingly, the homeologous LORe regions 2q-12qa, 4p-8q and 16qb-17qa had only one grayling ortholog corresponding to two copies in Atlantic salmon, suggesting either loss of large duplicated blocks or sequence assembly collapse. To probe further into these regions, we mapped the grayling Illumina paired end reads that were used for the assembly back to the grayling genome sequence using BWA-

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*	190 (5.5%)	19 (3.7%)
EEF2	Fixed-specific divergence: *Tissue regulation among duplicates are conserved within species but different between species*	195 (5.7%)	53 (10.5%)
EEF3	Salmon-specific divergence: *One Atlantic salmon duplicate has diverged in expression regulation*	396 (11.5%)	70 (13.8%)
EEF4	Grayling-specific divergence: *One grayling duplicate has diverged in expression regulation*	527 (15.3%)	87 (17.2%)
EEF5	Conserved: *All genes in the ohnolog-tetrad have conserved tissue regulation*	828 (24%)	127 (25%)
EEF6	Unclassified: *Tetrads with neutral-like expression evolution*	1308 (38%)	151 (29.8%)
Total		3444	507

MEM [30] 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 genes 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 [31] (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). The observed assembly collapse in some Ss4R regions in grayling could be related to a generally slower rediploidization rate compared to the Atlantic salmon lineage. To test for the difference in sequence divergence in the AORe and LORe, we computed the synonymous substitution rates (dS) between duplicate pairs in Atlantic salmon and grayling. Indeed, the difference between Atlantic salmon-pair-dS and grayling-pair-dS was significantly larger in LORe compared to AORe regions (Wilcoxon test, p=4.62e-11, 95% Confidence Interval: -Inf, -0.006), supporting a slower rediploidization rate in grayling LORe regions than in Salmoninae (Supplementary Figure S5).

Selection on gene expression regulation following Ss4R 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 first applying hierarchical clustering (see Methods) of tissue expression across duplicate pairs in Atlantic salmon and grayling. Each set of Ss4R duplicate pairs, hereafter referred to as ohnolog-tetrads, was assigned to one out of eight tissue-dominant expression clusters (Supplementary Figure S6). Next, ohnolog-tetrads were classified into groups of five distinct EEF categories each representing differences in past selection pressure on the tissue-regulation of ohnolog pairs. (see Table 2, Figure 3). The conserved divergence (EEF1) category represents expresison 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 duplicate. Lastly, EEF5 are 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 be a result of low tissue specificity (Table 2). After applying a gene tree topology-based filtering criteria (see Methods), 3,951 ohnolog-tetrads that conformed to expectations of LORe (507) or AORe (3444) gene tree topologies were used in further analyses.

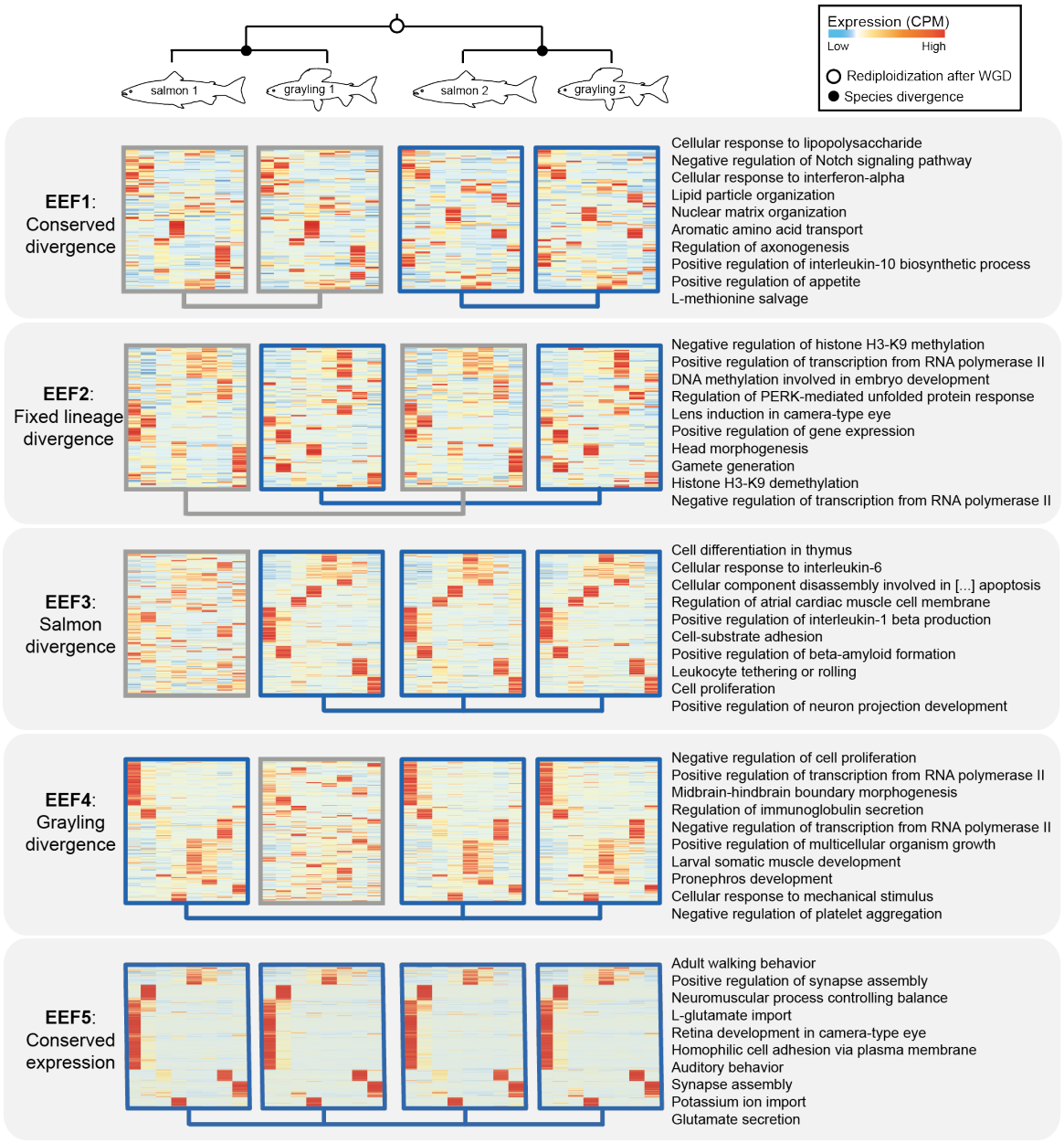

Figure 3 Selection on tissue expression regulation after whole genome duplication.
Heatmaps showing clustering of expression values across the five non-neutral Expression Evolution Fates (EEFs) reflecting differential selection on tissue expression regulation after Ss4R WGD. The phylogenetic tree (top) represents a typical AORe (ancestrally diverged) topology corresponding to the ohnolog-tetrads represented in the figure (similar patterns were observed in LORe, see Supplementary Figure S7). Each row across the four heatmaps represents one gene of a ohnolog-tetrad, with darker red corresponding to higher expression level, represented in terms of counts per million (CPM). Connecting lines below heatmaps indicate duplicates belonging to same tissue clusters (conserved expression pattern). The top 10 overrepresented Gene Ontology (GO) terms in each of the EEFs are indicated next to the heatmaps.

Of the 6 classes of EEFs, unclassified (EEF6, 30-38%) and conserved tissue regulation (EEF5, ∼25%) 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 relative size of the EEF categories were similar in rank in LORe and AORe regions (Table 2), the EEF category sizes were significantly different (Fisher's exact test, two sided, p-value < 0.0005, Supplementary

Table S4). This difference was caused by enrichment of conserved-diverged expression evolution (EEF1) in AORe tetrads and enrichment for lineage-specific expression divergence (EEF2) among LORe tetrads.

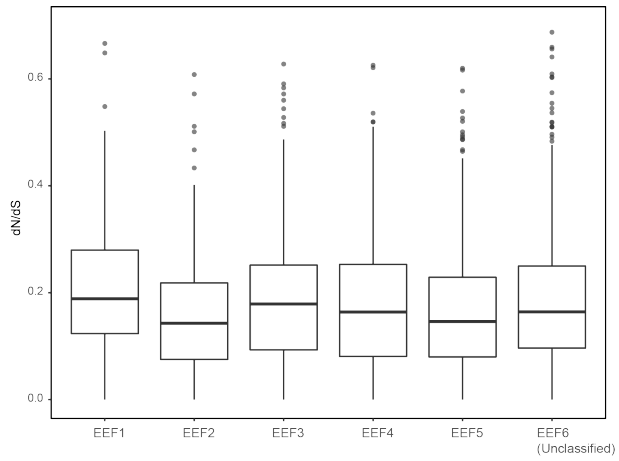

Figure 4 a. Distribution of dN/dS, representing coding sequence evolution, across the EEFs in grayling AORe regions. Similar patterns were observed in Atlantic salmon (see Supplementary Figure S9).

As different tissues are involved in different biological functions, we expect that regulatory evolution of gene expression is shaped by tissue-specific selective pressures [32]. 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 but one EEF-class (EEF3), between 2-5 tissue-expression clusters were significantly over- or underrepresented (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 bias towards brain-specific expression (Table S4). This finding was supported by high tissue-specificity (Tau score) of genes in ohnolog-tetrads associated with EEF5 (Supplementary Figure S8).

To evaluate if the ohnologs in different EEF classes were associated with distinct biological processes, we applied GO term enrichment tests on genes in EEF 1-5. Only 27 (among 721) overrepresented GO terms (p-adjusted < 0.05) were shared among >=2 of the five groups of expression evolution fates. Further inspection of top 10 GO terms in each EEF classes (Figure 3) shows links between EEF-categories and involvements in biological functions. EEF5 ohnologs under strict evolutionary 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.

Further, to test whether distinct evolutionary trajectories at the regulatory level (EEFs 1-5) were coupled to distinct patterns of protein-coding sequence evolution, we estimated dN/dS ratios for each duplicate pair within each species and compared the dN/dS distribution in each EEF class with that of the neutral-like ('unclassified') regulatory evolution (Figure 4 and Supplementary Figure S9). Low dN/dS (<<1) indicates strong purifying selection pressure. As with gene expression evolution, EEF 1-5 show clear 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=0.02) and EEF 2 and 5 having significantly lower dN/dS ratios (Wilcoxon rank sum, p=0.00016 and p=8.405e-05, 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=0.27 and 0.58, respectively).

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 EEF4, where a single grayling gene displayed diverged tissue-specific expression regulation, 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), which transports chloride ions over cell membranes in the gill. 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

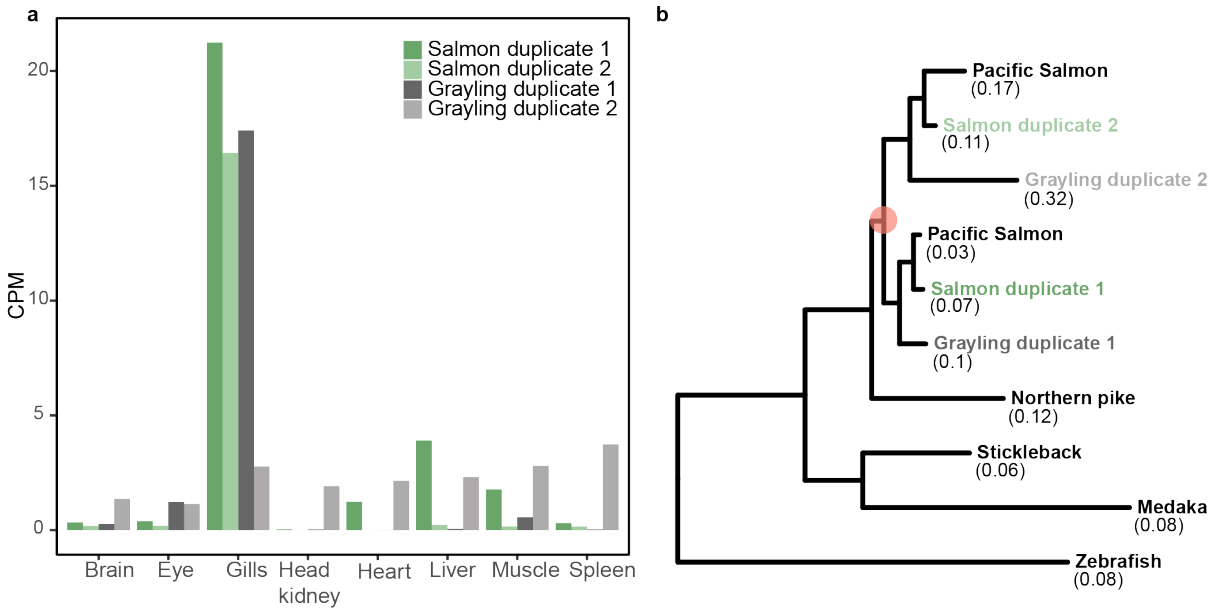

Figure 5 Divergent selection on cystic fibrosis transmembrane conductance regulator.
a) Expression values (Counts per million, CPM) of the cystic fibrosis transmembrane conductance regulator ohnologs in Atlantic salmon and grayling across eight tissues. b) Cystic fibrosis transmembrane conductance regulator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

with the conserved expression regulation, reflecting relaxation of purifying selection pressure in the non-anadromous grayling (Figure 5a and 5b).In Atlantic salmon, both CFTR Ss4R copies have been found to be involved in saltwater adaptations [33]. The combined activity of both CFTR copies might provide a fitness advantage in Atlantic salmon while the divergence of the second copy could simply indicate a different functional fate in freshwater 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 [18]. Our comparative approach provides valuable insights into the importance of selection in the contrasting processes of maintaining ancestral gene regulation and driving spatial expression divergence of ohnologs following WGD. 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). Our results corroborate the observation of biased retention of WGD derived duplicate genes related to nervous system and a strong expression conservation pattern in brain that has been described across vertebrates [34, 35, 36, 37]. 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.

The least common expression evolution fate (EEF1, ∼5%) are duplicates that reflect adaptive regulatory divergence followed by strong purifying selection in both grayling and Atlantic salmon. Although rare, duplicates of class EEF1 are particularly interesting as they represent key candidates for salmonid specific adaptive evolution of novel gene regulation enabled by the WGD. Salmonids are believed 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 (Figure 3). 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 (Figure 3 and 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.

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 [17, 16], 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. In combination with an overall lower neutral sequence divergence observed among Ss4R duplicates in grayling, this finding further supports a lower rate of rediploidization in the grayling genome as compared to that in the Atlantic salmon lineage. 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 [42]. The difference we observe in the rediploidization history of the two salmonids is thus likely linked to the distinctly different chromosome evolution in Atlantic salmon and grayling (Supplementary Figure S1) [43].

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. Among the ohnologs with grayling-specific divergence (EEF4), we found overrepresentation of genes associated with "ion homeostasis" functions. One of these ohnolog pairs comprises two CFTR genes, encoding a membrane chloride channel that exports chloride ions out of cells [44]. The grayling CFTR ohnolog with diverged tissue regulation has lost gill-tissue specificity and shows relaxed purifying selection pressure at the protein coding sequence level as well (>2-fold increase in dN/dS, Figure 5). The Ss4R CFTR ohnologs in Atlantic salmon are both primarily expressed in gills (Figure 5), and both are upregulated upon exposure to seawater [45]. It is therefore plausible that maintaining two functional CFTR genes is an 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.

In summary, we present the draft genome of European grayling using an efficient and cost effective short read sequencing strategy. The 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) [46]. A *de novo* assembly was generated with Allpaths-LG (release R48777) [27] using the 180bp paired-end library and the mate pair (3kb and 6kb) libraries. Assembly polishing was carried out using pilon (version 1.9) [28]. 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) [47]. 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 [16] 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) [48] 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 [49], 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) [50]. Cufflinks (version 2.1.1) [50, 51] and TransDecoder [52] 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) [53] and SNAP (version 20131129) [54] 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 [16]. The second pass involved additional data from training AUGUSTUS [55] 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) [56]. 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 [57] and BUSCO [58].

Analysis of orthologous groups

We used orthofinder (version 0.2.8, e-value threshold at 1e-05) [59] 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) [60] using default settings and the resulting alignments were then used to infer maximum likelihood gene trees using FastTree (v2.1.8) [61] (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 [62] 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 homoeology 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 [16]. 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 [63]. The CPM value (counts per million), here used as a proxy for expression, was then calculated using *edgeR* [64]. Similar CPM datasets were obtained from Atlantic salmon RNA-seq data reported in Lien et al [16]. 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 genes with non-zero CPM value were considered. This filtering resulted in a set of 5,026 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. To quantify the breadth of expression (i.e., the number of tissues a gene is expressed in), we calculated the tissue specificity index Tau [65] for all the genes in ohnolog-tetrads, where a Tau value approaching 1 indicates higher tissue specificity while 0 indicates ubiquitous expression.

Sequence evolution

To estimate coding sequence evolution rates, we converted amino acid alignments to codon alignments using pal2nal [66]. 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) [67]. 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.

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.

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`.

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. We thank Kim M. Bærum for sampling of grayling and Marianne H. S. Hansen for excellent technical assistance. Sample preparation, library contruction and sequecning 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 greatly appreciate Daniel J. Macqueen, Torgeir R. Hvidsten and Marine S. Brieuc for critical reading of the manuscript.

Author's contributions

LAV, KSJ, SJ and SL planned the project and generation of the data. SRS and SV performed all the analyses with help from AJN, OKT and SL. 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. 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)
4. Wolfe, K.H.: Yesterday's polyploids and the mystery of diploidization. Nat. Rev. Genet. **2**(5), 333–341 (2001)
5. Lynch, M., Conery, J.S.: The evolutionary fate and consequences of duplicate genes. Science **290**(5494), 1151–1155 (2000)
6. Zhang, J.: Evolution by gene duplication: an update. Trends Ecol. Evol. **18**(6), 292–298 (2003)
7. 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)
8. 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)
9. Carroll, S.B.: Endless forms: the evolution of gene regulation and morphological diversity. Cell **101**(6), 577–580 (2000)
10. Wray, G.A.: The evolution of transcriptional regulation in eukaryotes. Mol. Biol. Evol. **20**(9), 1377–1419 (2003)
11. 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)
12. 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)
13. 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)
14. 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)
15. 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)
16. 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)
17. 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)
18. 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)
19. Sandve, S.R., Rholfs, R., Hvidsten, T.R.: Subfunctionalization versus neofunctionalization after whole-genome duplication. Nat. Genet. Accepted
20. Allendorf, F.W., Thorgaard, G.H.: Tetraploidy and the evolution of salmonid fishes. In: Evolutionary Genetics of Fishes, pp. 1–53 (1984)
21. 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)
22. Hendry, A.P., Stearns, S.C.: Evolution Illuminated: Salmon and Their Relatives. Oxford University Press.(2004)
23. Nygren, A., Nilsson, B., Jahnke, M.: Cytological studies in *Thymallus thymallus* and *Coregonus albula*.

Hereditas **67**(2), 269–274 (1971)
24. Phillips, R., Ráb, P.: Chromosome evolution in the salmonidae (pisces): an update. Biol. Rev. Camb. Philos. Soc. **76**(1), 1–25 (2001)
25. Hartley, S.E.: THE CHROMOSOMES OF SALMONID FISHES. Biol. Rev. Camb. Philos. Soc. **62**(3), 197–214 (1987)
26. 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)
27. 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)
28. 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)
29. UniProt Consortium: UniProt: a hub for protein information. Nucleic Acids Res. **43**(Database issue), 204–12 (2015)
30. Li, H.: Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM (2013). 1303.3997
31. Garrison, E., Marth, G.: Haplotype-based variant detection from short-read sequencing (2012). 1207.3907
32. 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)
33. 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)
34. 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)
35. 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)
36. 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)
37. 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 (2016)
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. Comai, L.: The advantages and disadvantages of being polyploid. Nat. Rev. Genet. **6**(11), 836–846 (2005)
43. Qumsiyeh, M.B.: Evolution of number and morphology of mammalian chromosomes. J. Hered. **85**(6), 455–465 (1994)
44. 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)
45. Singer, T.D., Clements, K.M., Semple, J.W., Schulte, P.M., Bystriansky, J.S., Finstad, B., Fleming, I.A., Scott McKinley, R.: Seawater tolerance and gene expression in two strains of Atlantic salmon smolts. Can. J. Fish. Aquat. Sci. **59**(1), 125–135 (2002)
46. Martin, M.: Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal **17**(1), 10 (2011)
47. 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)
48. 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)
49. Faust, G.G., Hall, I.M.: YAHA: fast and flexible long-read alignment with optimal breakpoint detection. Bioinformatics **28**(19), 2417–2424 (2012)
50. 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)
51. 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)
52. 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)
53. Lomsadze, A.: Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. **33**(20), 6494–6506 (2005)
54. Korf, I.: Gene finding in novel genomes (2004)
55. 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)
56. 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)
57. Parra, G., Bradnam, K., Korf, I.: CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics **23**(9), 1061–1067 (2007)
58. 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)
59. Emms, D.M., Kelly, S.: OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. **16**, 157 (2015)
60. 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)
61. Price, M.N., Dehal, P.S., Arkin, A.P.: FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One **5**(3), 9490 (2010)
62. 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)
63. Anders, S., Pyl, P.T., Huber, W.: HTSeq–a python framework to work with high-throughput sequencing data. Bioinformatics **31**(2), 166–169 (2015)
64. 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)
65. 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)
66. 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)
67. Yang, Z.: PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13**(5), 555–556 (1997)

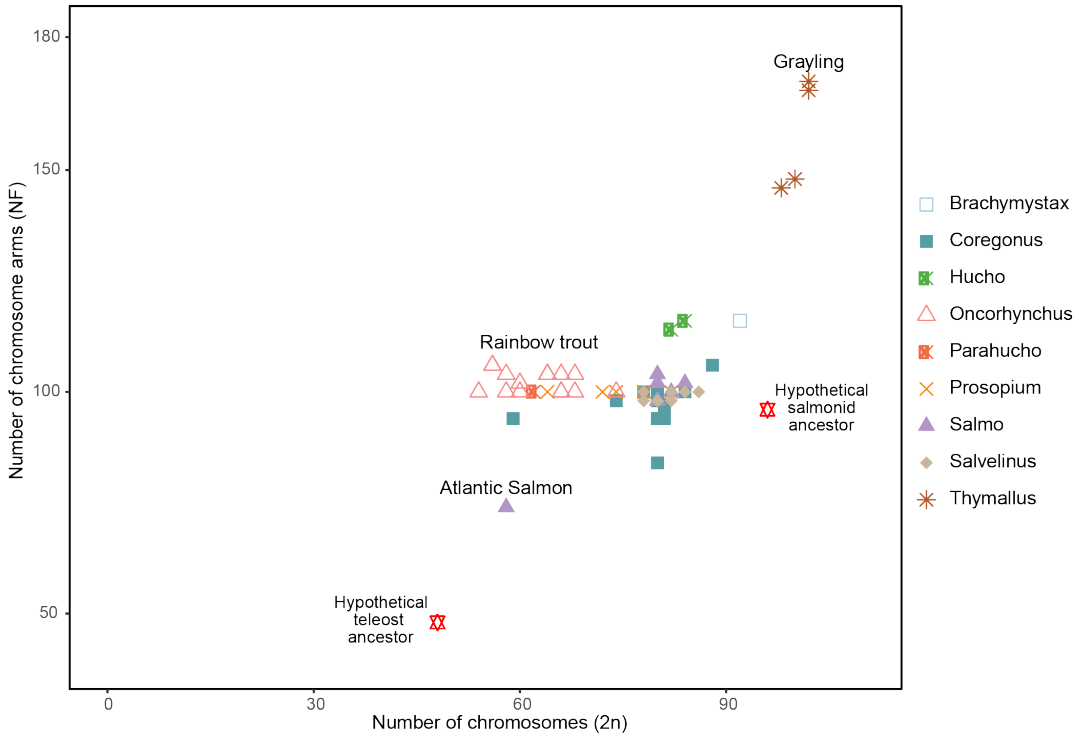

Figure S1 Chromosome evolution in salmonids
Chromosome number (2N) plotted against Number of chromosome arms (NF). Data from (Hartley et al. 1987). Based on the karyotype of most extant teleosts (48-50 acrocentric chromosomes), it is hypothesized that the salmonid ancestor had a karyotype of around 96-100 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).

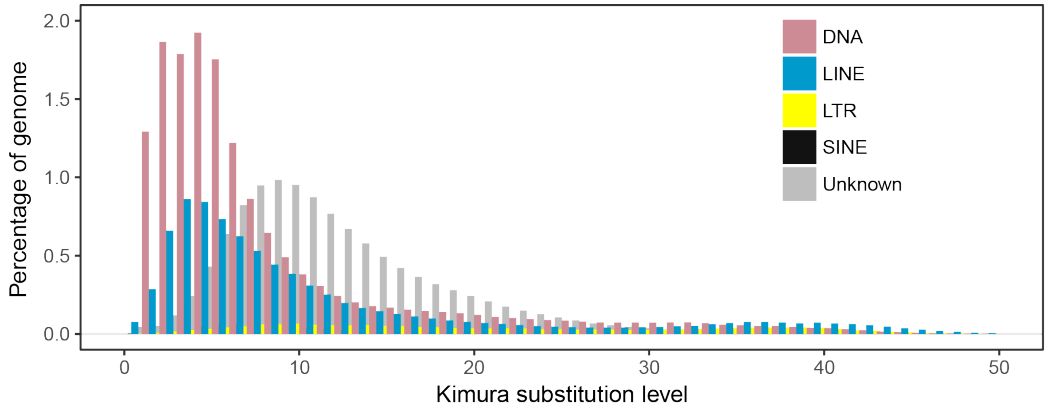

Figure S2 Repeat landscape of grayling genome based on Kimura distance.
X-axis represents divergence from repeat consensus sequence and y-axis represents the proportion of the transposable element family in the genome (where LTR stands for long terminal repeats, LINE represents long interspersed nuclear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

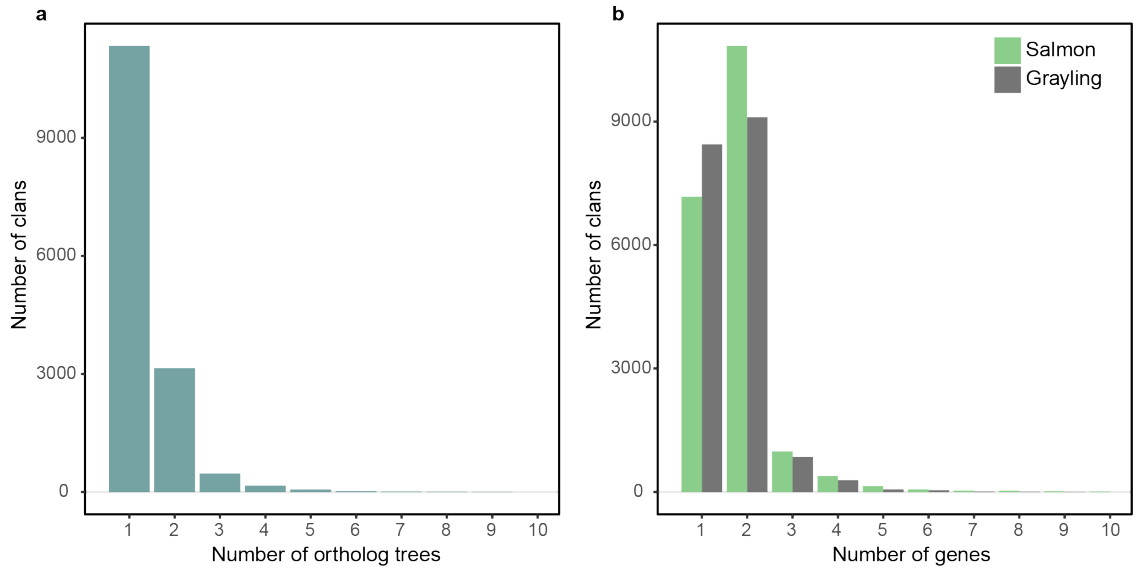

Figure S3 a) Distribution of clans per ortholog tree. b) Number of Atlantic salmon and grayling genes per clan.

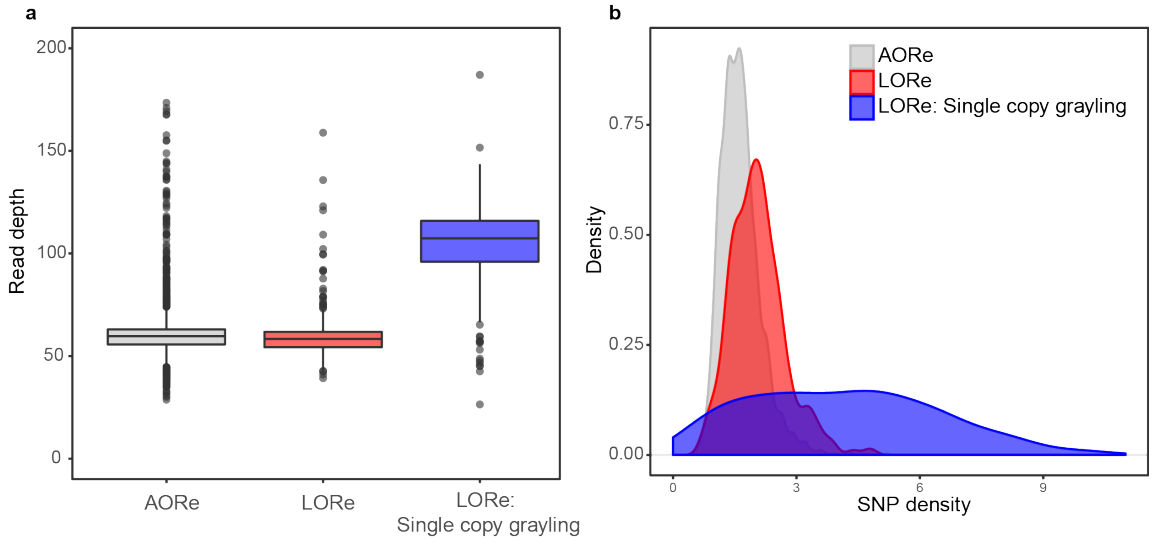

Figure S4 Distribution of (a) mapped read depth and (b) SNP density per Kb across all Ss4R duplicates in grayling grouped by the ohnolog resolution models.

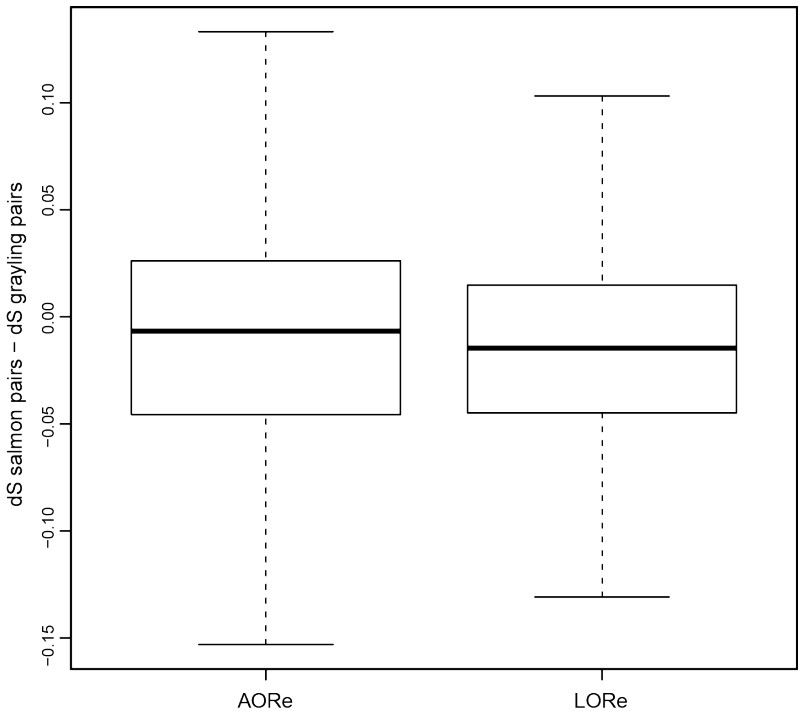

Figure S5 Difference in synonymous substitution rates (dS) between Atlantic salmon and grayling in AORe and LORe regions.

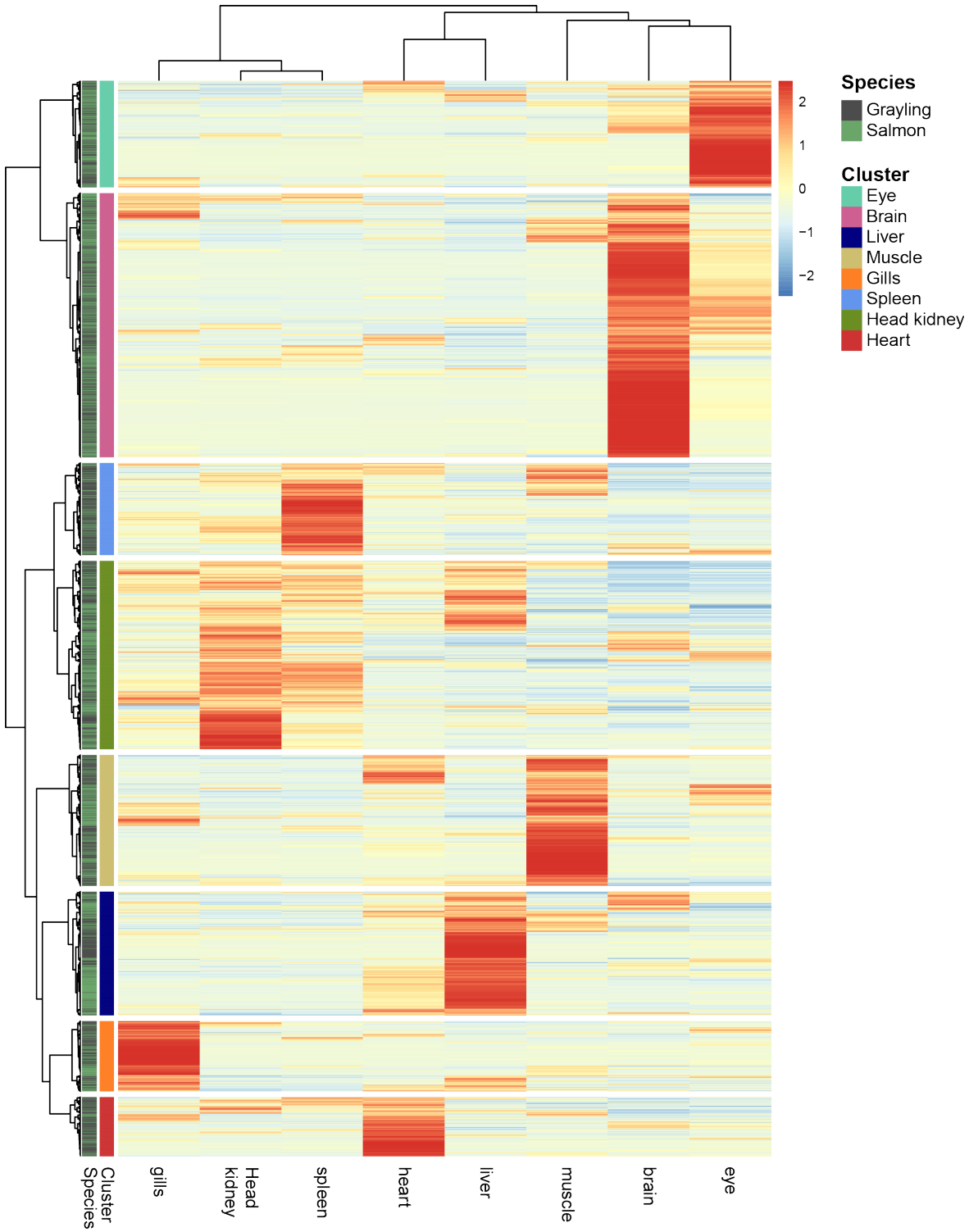

Figure S6 Heatmap of tissue expression clusters from Atlantic salmon and grayling.
Tissue expression profile of ohnologs from Atlantic salmon and grayling using hierarchical clustering. The color scale of the heatmap corresponds to the relative abundance of the transcript across all the tissues within 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.

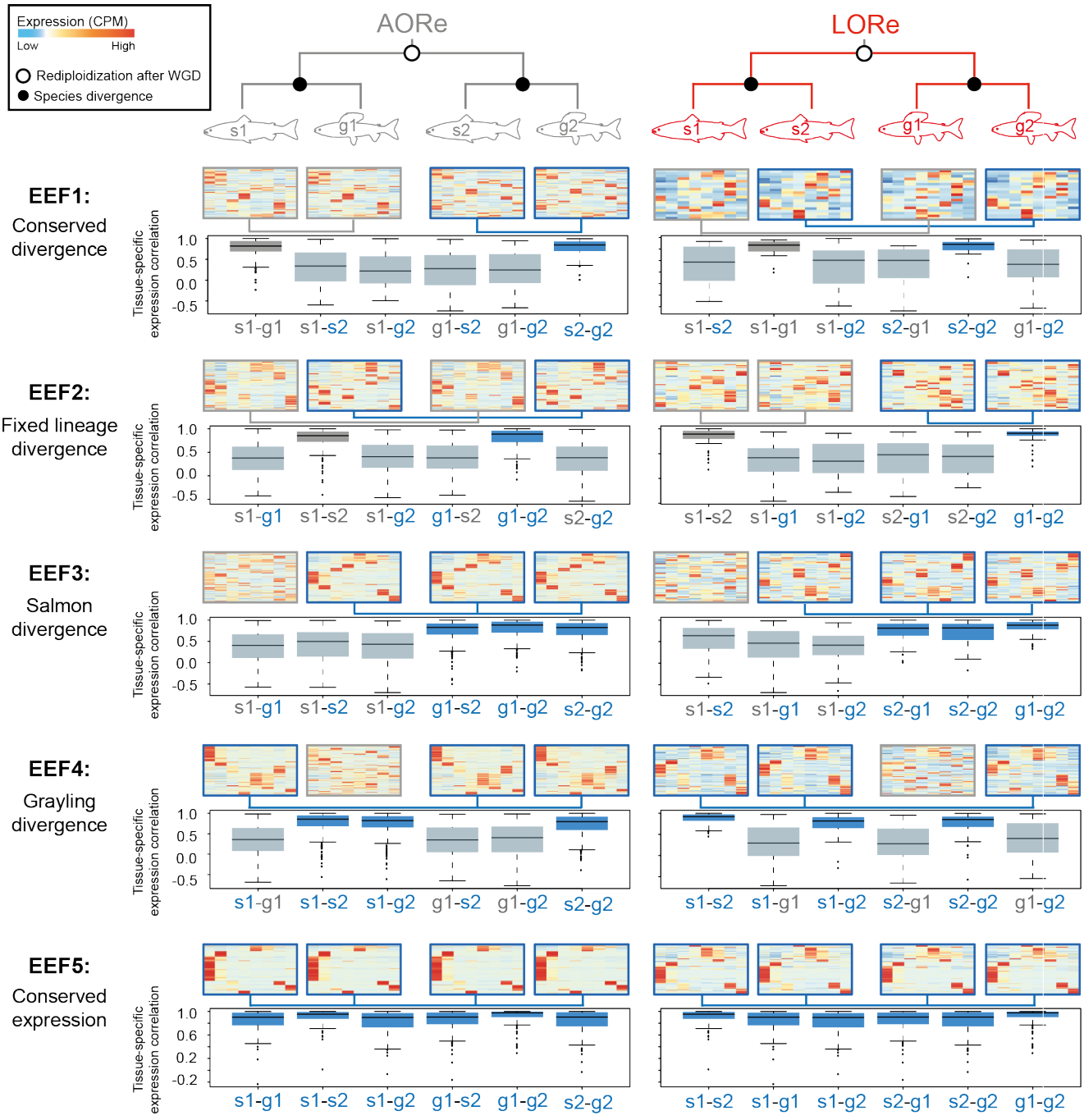

Figure S7
An extension of Figure 3 with heatmaps showing 5 expression evolution fates (EEFs, see Table 2) reflecting differential selection on tissue expression regulation after Ss4R WGD over genes with LORe and AORe histories. The color scale of the heatmaps correspond to the relative abundance of the 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.

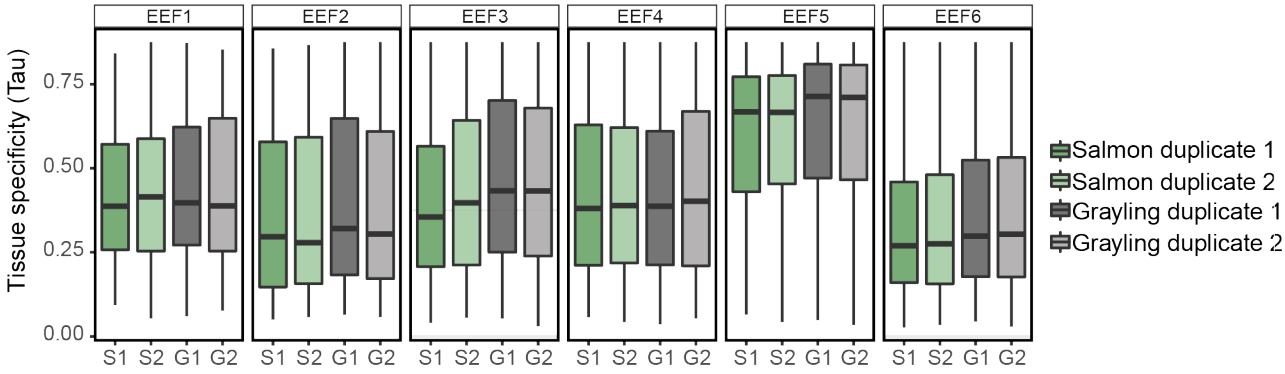

Figure S8 Overall tissue specificity (Tau) distribution for each of the EEFs. The ohnologs in Atlantic salmon and grayling are represented as s1, s2 and g1 and g2 respectively.

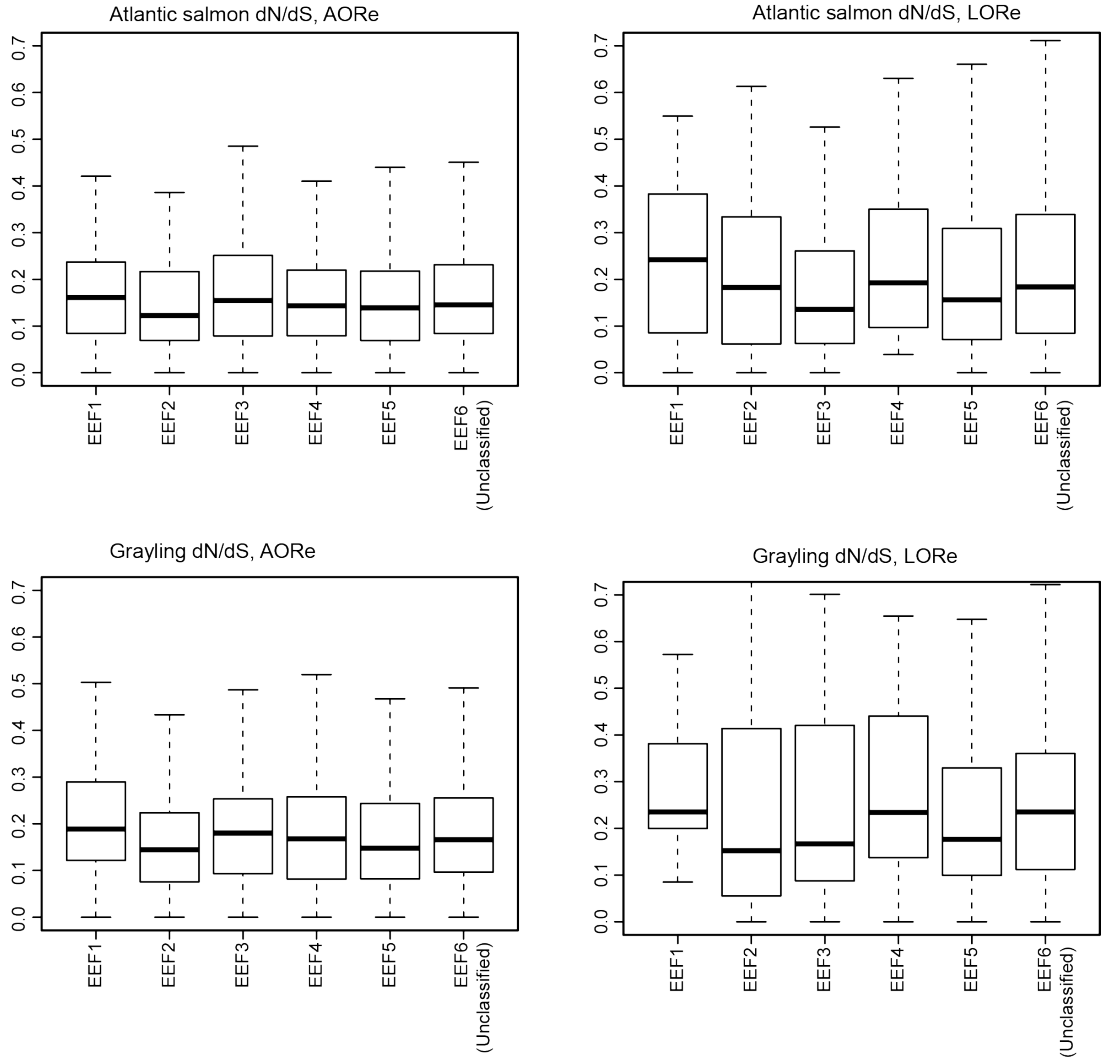

Figure S9 Distribution of dN/dS, representing coding sequence evolution, across different EEF categories across the LORe and AORe regions in Atlantic salmon and grayling.

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	130 (20%)	60 (9%)	72 (11%)	38 (6%)	122 (19%)	112 (17%)	48 (7%)	62 (10%)
EEF2	180 (18%)	136 (14%)	56 (6%)	42 (4%)	172 (17%)	86 (9%)	130 (13%)	190 (19%)
EEF3	413 (22%)	171 (9%)	149 (8%)	103 (6%)	355 (19%)	266 (14%)	251 (13%)	156 (8%)
EEF4	631 (26%)	193 (8%)	131 (5%)	128 (5%)	584 (24%)	327 (13%)	245 (10%)	217 (9%)
EEF5	1764 (46%)	396 (10%)	240 (6%)	44 (1%)	556 (15%)	328 (9%)	436 (11%)	56 (1%)
All	4024 (25%)	1559 (10%)	1124 (7%)	932 (6%)	2834 (18%)	1949 (12%)	1923 (12%)	1459 (9%)

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