



THE UNIVERSITY *of* EDINBURGH

## Edinburgh Research Explorer

# Genome-wide reconstruction of rediploidization following autopolyploidization across one hundred million years of salmonid evolution

### Citation for published version:

Gundappa, MK, To, TH, Grønvold, L, Martin, SAM, Lien, S, Geist, J, Hazlerigg, D, Sandve, SR & Macqueen, D 2021, 'Genome-wide reconstruction of rediploidization following autopolyploidization across one hundred million years of salmonid evolution', *Molecular Biology and Evolution*.  
<https://doi.org/10.1093/molbev/msab310>

### Digital Object Identifier (DOI):

[10.1093/molbev/msab310](https://doi.org/10.1093/molbev/msab310)

### Link:

[Link to publication record in Edinburgh Research Explorer](#)

### Document Version:

Publisher's PDF, also known as Version of record

### Published In:

Molecular Biology and Evolution

### General rights


Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

### Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact [openaccess@ed.ac.uk](mailto:openaccess@ed.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.



# Genome-Wide Reconstruction of Rediploidization Following Autopolyploidization across One Hundred Million Years of Salmonid Evolution

Manu Kumar Gundappa,<sup>1,2</sup> Thu-Hien To,<sup>†,3</sup> Lars Grønvold,<sup>†,3</sup> Samuel A.M. Martin,<sup>2</sup> Sigbjørn Lien,<sup>3</sup> Juergen Geist ,<sup>4</sup> David Hazlerigg,<sup>5</sup> Simen R. Sandve,<sup>3</sup> and Daniel J. Macqueen<sup>\*,1</sup>

<sup>1</sup>The Roslin Institute and Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Midlothian, United Kingdom

<sup>2</sup>School of Biological Sciences, University of Aberdeen, Aberdeen, United Kingdom

<sup>3</sup>Department of Animal and Aquacultural Sciences, Faculty of Biosciences, Centre for Integrative Genetics (CIGENE), Norwegian University of Life Sciences, Ås, Norway

<sup>4</sup>Aquatic Systems Biology Unit, TUM School of Life Sciences, Technical University of Munich, Freising, Germany

<sup>5</sup>Department of Arctic and Marine Biology, Faculty of BioSciences Fisheries & Economy, University of Tromsø, Norway

<sup>†</sup>These authors contributed equally to this work.

\*Corresponding author: E-mail [daniel.macqueen@roslin.ed.ac.uk](mailto:daniel.macqueen@roslin.ed.ac.uk).

Associate editor: Mary O'Connell

## Abstract

The long-term evolutionary impacts of whole-genome duplication (WGD) are strongly influenced by the ensuing rediploidization process. Following autopolyploidization, rediploidization involves a transition from tetraploid to diploid meiotic pairing, allowing duplicated genes (ohnologs) to diverge genetically and functionally. Our understanding of autopolyploid rediploidization has been informed by a WGD event ancestral to salmonid fishes, where large genomic regions are characterized by temporally delayed rediploidization, allowing lineage-specific ohnolog sequence divergence in the major salmonid clades. Here, we investigate the long-term outcomes of autopolyploid rediploidization at genome-wide resolution, exploiting a recent “explosion” of salmonid genome assemblies, including a new genome sequence for the huchen (*Hucho hucho*). We developed a genome alignment approach to capture duplicated regions across multiple species, allowing us to create 121,864 phylogenetic trees describing genome-wide ohnolog divergence across salmonid evolution. Using molecular clock analysis, we show that 61% of the ancestral salmonid genome experienced an initial “wave” of rediploidization in the late Cretaceous (85–106 Ma). This was followed by a period of relative genomic stasis lasting 17–39 My, where much of the genome remained tetraploid. A second rediploidization wave began in the early Eocene and proceeded alongside species diversification, generating predictable patterns of lineage-specific ohnolog divergence, scaling in complexity with the number of speciation events. Using gene set enrichment, gene expression, and codon-based selection analyses, we provide insights into potential functional outcomes of delayed rediploidization. This study enhances our understanding of delayed autopolyploid rediploidization and has broad implications for future studies of WGD events.

**Key words:** whole-genome duplication, rediploidization, ohnolog, phylogenomics, genome evolution.

## Introduction

Whole-genome duplication (WGD) leading to polyploidy has occurred extensively during eukaryotic evolution (Soltis et al. 2015; Van de Peer et al. 2017). This includes complex WGD histories in plant evolution (Qiao et al. 2019), lineage-defining WGD events ancestral to vertebrates (Simakov et al. 2020) and teleosts (Jaillon et al. 2004), and additional WGDs in several fish families, including salmonids (Lien et al. 2016), cyprinids (Li and Guo 2020), and sturgeons (Du et al. 2020). WGD is widely thought to promote evolutionary diversification through mechanisms that remain incompletely understood (Van de Peer et al. 2017).

Rediploidization follows all WGD events and creates novel genetic diversity (Wolfe 2001). After WGD within the same species (autopolyploidization), rediploidization involves a transition from multivalent (tetraploid inheritance) to bivalent chromosome pairing (diploid inheritance) during meiosis (Furlong and Holland 2002; Lien et al. 2016). Consequently recombination among four alleles ceases, promoting sequence divergence between duplicated genes (ohnologs) on distinct chromosomes (Furlong and Holland 2002). This, in turn, creates novel pathways of functional evolution compared with before WGD (Ohno 1970; Conant and Wolfe 2008; Innan and Kondrashov 2010). Rediploidization depends

on mutations that promote preferential bivalent pairing during meiosis, such as structural rearrangements (e.g., inversions) and transposable element (TE) insertions (Ohno 1970; Weiss and Maluszynska 2000; Lien et al. 2016). The same rediploidization process will be absent in allopolyploids (WGD following hybridization of different species) showing immediate preferential bivalent pairing of the subgenomes descended from each parent species (Cifuentes et al. 2010; Mason and Wendel 2020), but in theory can occur whenever sequence similarity is sufficient for multivalent pairings to arise, for example, in segmental allopolyploids (Martin and Holland 2014; Robertson et al. 2017).

A past body of work in salmonid fishes revealed that rediploidization occurred at distinct times in evolution for different genomic regions following an ancestral autopolyploidization (hereafter: “Ss4R”) dated at 88–103 Ma (Macqueen and Johnston 2014). The Ss4R is the fourth WGD in salmonid evolutionary history (Berthelot et al. 2014; Lien et al. 2016) following earlier events at the base of vertebrate (Simakov et al. 2020) and teleost evolution (Jaillon et al. 2004). The variable timing of rediploidization for duplicated regions retained from Ss4R is reflected as a “snapshot” within all salmonid genome by distinct levels of sequence divergence among large syntenic blocks of ohnologs (Lien et al. 2016). Although rediploidization occurred in the ancestor to all living salmonids in large genomic regions, including several entire chromosomes, speciation occurred before rediploidization had completed in several large genomic segments (Robertson et al. 2017). Consequently, some duplicated chromosome arms share very high sequence similarity (>95%) in all salmonid species (Lien et al. 2016; De-Kayne and Feulner 2018; Campbell et al. 2019; Blumstein et al. 2020) and experienced ohnolog divergence independently in the three salmonid subfamilies, which diverged ~50 Ma (Robertson et al. 2017). This process was coined “lineage-specific ohnolog resolution” (“LORe”) and is thought to be possible whenever the evolutionary transition from multivalent to bivalent pairing (i.e., from tetraploid alleles to ohnolog pairs) occurs after speciation events separating lineages descended from the same WGD event (Martin and Holland 2014; Robertson et al. 2017).

Delayed rediploidization and LORe has implications for phylogenetic inference, as it removes any possibility of 1:1 ortholog relationships among ohnologs retained in affected sister clades, which is the classic expectation following ancestral gene duplication and WGD events (Martin and Holland 2014; Robertson et al. 2017). LORe can readily be mistaken for lineage-specific (e.g., tandem) gene duplication during phylogenetic analyses, if global expectations of WGD (i.e., collinearity among ohnolog pairs on distinct chromosomes) are overlooked (Robertson et al. 2017). When considering the potential evolutionary impacts of WGD events, LORe was hypothesized to promote lineage-specific adaptation (Robertson et al. 2017) and offers a plausible framework to explain frequent observations of time-lags between WGD event and subsequent species or phenotypic diversification regimes (Schranz et al. 2012; Clark and Donoghue 2017; Carretero-Paulet and Van de Peer 2020). Following the discovery of delayed rediploidization and LORe in salmonids,

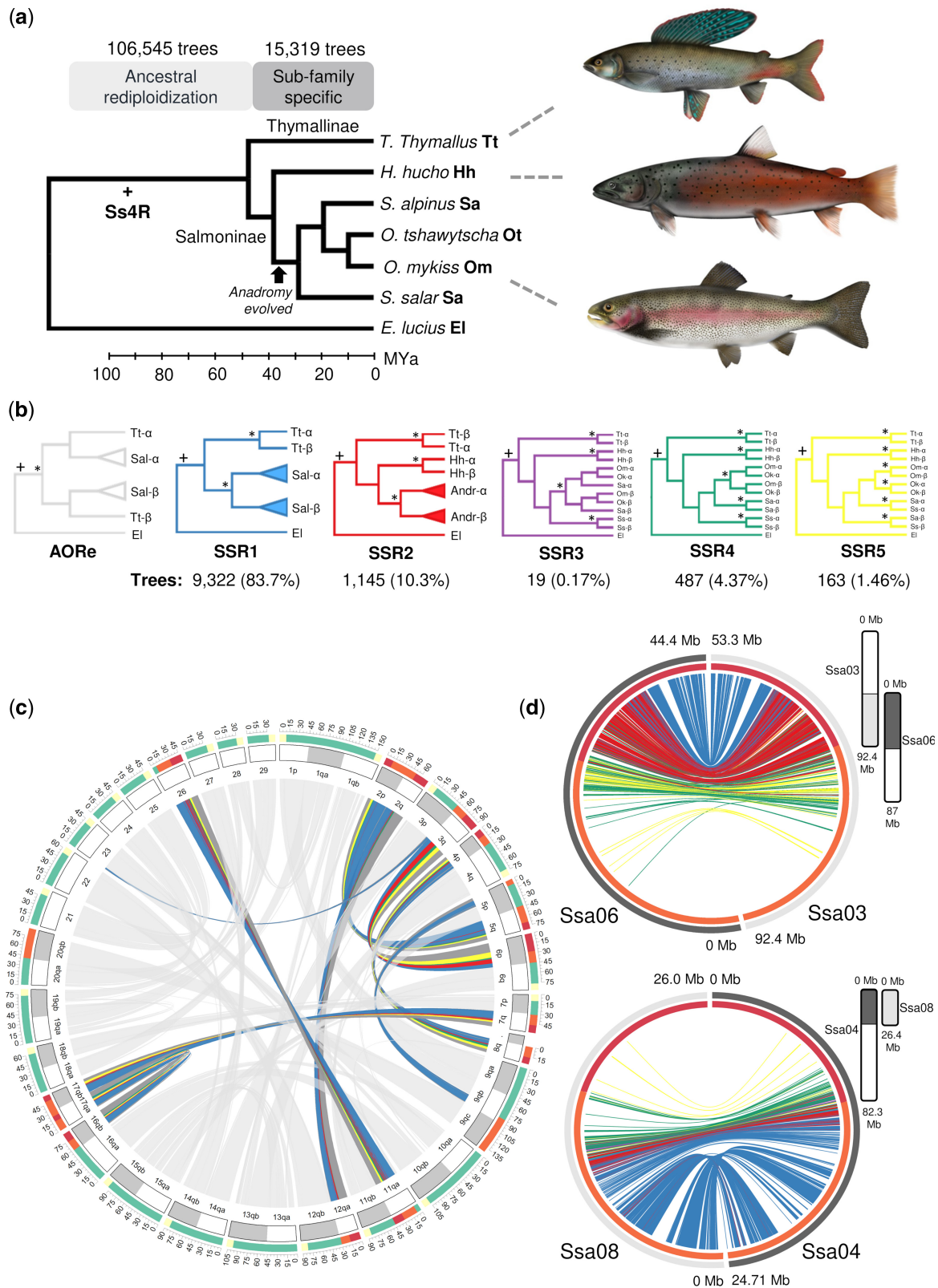
many authors have realized the possible importance of these processes for WGD events in divergent taxa (Macqueen and Johnston 2014; Martin and Holland 2014; Clark and Donoghue 2017; Van de Peer et al. 2017; Rozenfeld et al. 2019; Carretero-Paulet and Van de Peer 2020). Although salmonids represent an outstanding study system, our understanding of rediploidization outcomes in this group of fishes has remained fragmented due to a lack of genome-wide sequence information and/or limited phylogenetic resolution in past reconstructions.

The overarching aim of this study was to reconstruct the post-Ss4R rediploidization process and its outcomes with vastly increased genomic and phylogenetic resolution compared with past work. We sequenced a genome for a species holding a particularly informative phylogenetic position within the salmonid family, and developed a whole-genome alignment approach to capture ohnolog regions across genome assemblies recently generated for multiple salmonid species. This unique data set allowed us to reconstruct genome-wide rediploidization dynamics using phylogenetic methods, capturing two major waves of rediploidization in addition to complex lineage-specific ohnolog divergence histories that scale in complexity with speciation history. Exploiting this new high-resolution “map” of rediploidization, we enhance our understanding of the influence of rediploidization dynamics on a range of gene functional properties. Finally, we discuss the broader implications of our findings for ongoing research into the evolutionary outcomes of WGD events.

## Results

### Reference Genome Assembly for the Huchen (Danube Salmon)

To enhance scope to reconstruct rediploidization dynamics across salmonid evolution, we generated a high-quality genome sequence for the huchen (*Hucho hucho*), also known as the Danube salmon (supplementary fig 1, tables 1–3 and supplementary methods 1, Supplementary Material online). This species holds a key phylogenetic position for salmonid comparative genomics. It is part of a species-poor clade within subfamily Salmoninae that is sister to a species-rich clade including Atlantic salmon (*Salmo salar*) and Pacific salmon (*Oncorhynchus*) species (fig. 1a). The common ancestor of the latter clade is thought to have evolved the capability to migrate into seawater during the life-cycle (termed anadromy), which represents a dominant life-history strategy in extant member species (Alexandrou et al. 2013). In contrast, all species within the huchen’s clade complete the full life-cycle in freshwater, which represents the inferred ancestral state for salmonids (as observed in grayling species; subfamily Thymallinae) (Alexandrou et al. 2013). Past work has shown that ~25% of the Salmoninae genome experienced rediploidization after the split from Thymallinae (Robertson et al. 2017). Adding the huchen to this study captures the most basal speciation event in Salmoninae, allowing us distinguish regions in the genome that underwent rediploidization in the common Salmoninae ancestor, from regions that experienced lineage-specific rediploidization after the split of



**FIG. 1.** Genome-wide phylogenetic reconstruction of rediploidization history following the Ss4R autoploidy. (a) Phylogeny and divergence times for species used in genome-wide alignment. Also highlighted are the number of captured ohnolog trees in ancestral rediploidization (AORE) and subfamily specific rediploidization (SSR) regions of the genome and the timing of the Ss4R event (Macqueen and Johnston 2014). (b) The number of phylogenetic trees matching predicted lineage-specific rediploidization scenarios in SSR regions. (c) Circos plot mapping inferred rediploidization histories (i.e., SSR categories) from ohnolog trees along the Atlantic salmon genome; colors match to the SSR topologies shown in (b). Chromosome arm names follow established nomenclature for Atlantic salmon (Lien et al. 2016). (d) Circos plots mapping SSR topologies for two example chromosome arms. Additional data provided in [supplementary figs. 2–10, Supplementary Material online](#).

huchen from the ancestrally anadromous Salmoninae lineage (fig. 1a). As the huchen is an endangered species, a reference genome can also be used to support genetic research aimed at wild stock conservation and restoration (Geist et al. 2009; Kucinski et al. 2015).

Our huchen assembly was generated using Illumina technology from a haploid individual (supplementary fig. 1, Supplementary Material online) and had 2.49 Gb total sequence length, contig/scaffold N50 of 37.6/287.3 kb and 90.2% BUSCO (Simão et al. 2015) completeness (supplementary methods 1, Supplementary Material online). An annotated version with 50,114 coding gene models is available on the Ensembl genome browser ([https://www.ensembl.org/Hucho\\_hucho](https://www.ensembl.org/Hucho_hucho)).

### Multispecies Genome Alignment Including Ohnologs

With the goal to reconstruct genome-wide rediploidization dynamics in salmonids, we developed a genome alignment approach to capture Ss4R ohnolog regions across multiple species (see Materials and Methods). These alignments included species from Salmoninae and Thymallinae as well as northern pike *Esox lucius*, a representative of Esociformes—a sister lineage to salmonids that diverged before Ss4R (Macqueen and Johnston 2014; Lien et al. 2016; Robertson et al. 2017) (fig. 1a). We generated multispecies alignments for a priori defined syntenic ohnolog blocks retained from Ss4R (Lien et al. 2016) in two genome portions where rediploidization was either ancestral to all salmonids (ancestral ohnolog resolution “AORe” regions; Robertson et al. 2017), or occurred after the split of Salmoninae and Thymallinae (subfamily-specific rediploidization “SSR” regions) (supplementary data 1 and 2, Supplementary Material online).

Using this approach, 3,709,704 and 511,436 raw alignments were generated densely covering ohnolog blocks in AORe and SSR regions, respectively (supplementary data 3 and 4, Supplementary Material online). We next applied a step to filter the alignments according to the number of sequences represented to ensure we retained informative multispecies representation of ohnolog regions (see Materials and Methods), and further applied Gblocks (Castresana 2000) to remove low confidence positions in each alignment. This led to 106,545 (sum length: 92.2 Mb) and 15,319 (sum length: 32.3 Mb) high-quality alignments that broadly represent the length of every defined AORe and SSR regions, respectively (supplementary data 5 and 6, Supplementary Material online), hence achieving our goal of providing a genome-wide representation of duplicated regions retained from Ss4R. Each alignment was used to generate the same number of maximum-likelihood phylogenetic trees (fig. 1a). The alignments and trees used in subsequent analyses are provided and described in supplementary data 7 and 8, Supplementary Material online, respectively.

### High-Resolution Reconstruction of Lineage-Specific Rediploidization Histories

Using our genome-wide data set of phylogenetic trees, we classified rediploidization histories based on the onset of ohnolog divergence (Robertson et al. 2017) in the SSR regions

of the genome. Five distinct tree topologies (hereafter: SSR1, 2, 3, 4, and 5) capture the spectrum of predicted rediploidization histories according to the species tree (fig. 1b). For instance, SSR1 indicates independent rediploidization (i.e., ohnolog divergence) in Thymallinae and the common ancestor of Salmoninae members. At the other end of the spectrum, SSR5 indicates independent rediploidization in every species included (fig. 1c).

A total of 11,136 (72.7%) of the available 15,319 trees matched to one of the predicted SSR histories (done as described in Materials and Methods section: “Reconstructing rediploidization history and LORe in SSR regions”). Among these, most trees (83.7% of 11,136) supported rediploidization in the Salmoninae ancestor (SSR1 topology; fig. 1b). A total of 10.3% of the trees matched to expectations of two independent rediploidization histories in Salmoninae, once during *Hucho* evolution and again in the ancestor to *Salmo*, *Salvelinus*, and *Oncorhynchus* (SSR2 topology; fig. 1b). A smaller number of trees matched to predictions of additional rediploidization events nested within Salmoninae (SSR3–5 topologies; fig. 1b).

By positionally mapping the phylogenetic topologies along the Atlantic salmon genome (Lien et al. 2016), it was evident that SSR classifications are not randomly distributed, with large genomic regions dominated by common phylogenetic signals (fig. 1c and d and supplementary figs. 2–10, Supplementary Material online). Different ohnolog blocks (chromosome arm nomenclature used standard for *S. salar*; e.g., Lien et al. 2016) have distinct rediploidization histories, with most dominated by the SSR1 category. Some regions, including Ssa03-06 and Ssa04-08 (fig. 1d and supplementary figs. 2 and 3, Supplementary Material online) in addition to Ssa02-12 (supplementary fig. 4, Supplementary Material online) harbor large genomic regions dominated by the SSR2 category. The mapping of SSR topologies was closely associated with the level of sequence divergence between ohnolog regions (fig. 1c and d). Duplicated regions sharing >97% identity either represent SSR4/SSR5 topologies or more commonly, missing data (fig. 1d); these regions often harbored insufficient data to pass our alignment filtering criteria (see Materials and Methods). In particular, these alignments often contained a single sequence in multiple species, suggesting collapse during genome assembly due to the high similarity of ohnolog sequences (e.g., Varadharajan et al. 2018).

Although the genomic location of different SSR tree topologies was strongly clustered, some regions contained a mixture of different, closely related phylogenetic topologies; a pattern prevalent in regions dominated by SSR2–4 trees (e.g., fig. 1c and d). Although this could reflect weak phylogenetic signal, the average filtered alignment length was >2 kb, and our quality control efforts removed weakly supported topologies (see Materials and Methods). We initially hypothesized that this observation reflected errors in the reference genome, which has lowest accuracy in late rediploidization regions showing high ohnolog similarity (Bertolotti et al. 2020). To test this, we positioned all trees mapping to Ssa03-06 (1,898 trees) along the homologous regions of a more recent long-read-based genome for brown trout *S.*

*trutta*, with markedly higher contiguity. An identical pattern was observed (supplementary figs. 11 and 12, Supplementary Material online), suggesting assembly error was not an important factor, or that both assemblies suffered the same issue. Another possibility we considered was that the mixing of phylogenetic signals resulted from using genome alignment data, rather than gene trees. To test this, we mapped 236 high-quality gene trees including Ss4R ohnologs across the same salmonid species (generated by Bertolotti et al. 2020) to Ssa03-06, and observed the same pattern (supplementary fig. 13, Supplementary Material online). Our current interpretation is that the mixing of related SSR topologies is explained by small-scale intrachromosomal rearrangements that reordered the position of genomic regions sharing common rediploidization histories.

To summarize, these findings reveal complex histories of lineage-specific ohnolog divergence resulting from delayed rediploidization, which scale in number and complexity with the number and timing of speciation events during salmonid evolution.

### Spectrum of Rediploidization Ages across the Genome

The complex lineage-specific rediploidization histories inferred in SSR regions led us to ask if rediploidization timing varied across AORE regions. The challenge to answering this question is that all trees in AORE regions are characterized by the same topology, representing a single onset of ohnolog divergence in the salmonid common ancestor (Macqueen and Johnston 2014; Robertson et al. 2017) (fig. 1b). In other words, unlike SSR regions, tree topologies provide no information on rediploidization age.

As an alternative approach to empirically estimate rediploidization age, we applied the Bayesian relaxed clock approach MCMCtree (Yang 2007) using the genome-wide alignment data and published temporal constraints on species divergence (Campbell et al. 2013; Macqueen and Johnston 2014; Lecaudey et al. 2018) to estimate the temporal onset of ohnolog divergence (after Macqueen and Johnston 2014). For AORE regions, we concatenated sequence alignments across 23 defined ohnolog regions in the Atlantic salmon genome (Lien et al. 2016) assuming each represented a single shared rediploidization history (supplementary data 10, Supplementary Material online). This assumption was based on the fact that these duplicated regions have maintained extensive collinearity because Ss4R, so evidently have not undergone major rearrangements (Lien et al. 2016). As major rearrangements are thought to block homeologous pairing during meiosis leading to rediploidization (Lien et al. 2016), it is reasonable to assume an absence of unique rediploidization histories within each duplicated AORE block. The estimated rediploidization ages (Bayesian means) ranged from 68 to 106 Ma (fig. 2a and supplementary data 11, Supplementary Material online). Although representing a spectrum of rediploidization ages, the 95% credibility intervals overlapped for 21 out of the 23 regions (fig. 2a). This makes it impossible, even with the maximal available sequence data, to distinguish scenarios where rediploidization occurred concomitantly from scenarios where rediploidization was staggered across tens of

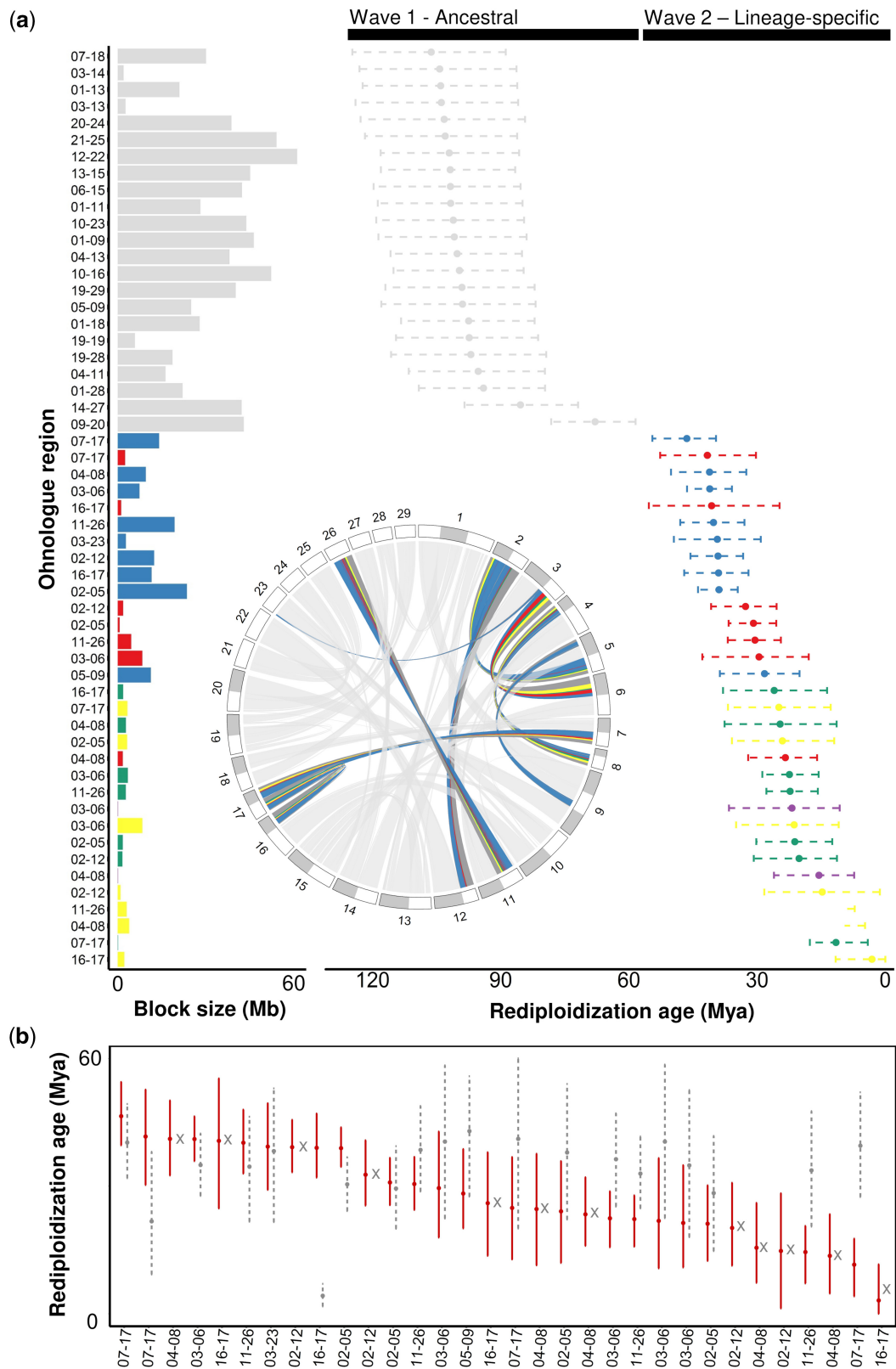
millions of years. Nonetheless, the duplicated region Ssa09-20 underwent rediploidization later in time, as its upper 95% credibility interval does not overlap with the lower 95% credibility interval for all but one of the remaining AORE regions (fig. 2a). This analysis suggests that AORE regions representing 61.4% (1.37 out of 2.24 Gb) of the chromosome-anchored Atlantic salmon genome underwent rediploidization no later than ~80 Ma according to the lower 95% credibility intervals (fig. 2a and supplementary data 11, Supplementary Material online). Finally, the most ancient inferred rediploidization ages indicated an older absolute date for the timing of Ss4R than current estimates (Berthelot et al. 2014; Macqueen and Johnston, 2014; Lien et al. 2016), with Ssa07-18 showing a mean rediploidization age of 106 Ma (fig. 2a and supplementary data 11, Supplementary Material online).

The same approach allowed us to estimate rediploidization ages for SSR regions, where it is possible to infer the timing of lineage-specific ohnolog divergence in different species. This was done for Atlantic salmon and European grayling, which split ~50–60 Ma (Campbell et al. 2013; Macqueen and Johnston 2014; Lecaudey et al. 2018), concatenating alignments for genomic regions showing strong support for different SSR topologies (supplementary data 10, Supplementary Material online; visualized in fig. 1c). We observed cases where ohnolog regions have similar rediploidization age estimates in both lineages, despite their independent histories of divergence, with overlapping 95% credibility intervals, as well as regions with very different rediploidization ages. For example, a large SSR1 region within Ssa05-09 has the oldest rediploidization age estimate (Bayesian mean: ~43 Ma) in European grayling; ~15 Ma older than in Salmoninae (Bayesian mean: ~28 Ma) (fig. 2b). Two chromosome arms, Ssa04-08 and Ssa02-12, containing regions (spanning SSR1-5 in Salmoninae) with estimated rediploidization ages from 14–42 to 15–39 Ma, respectively, are represented by a single sequence in the European grayling genome, indicative of assembly collapse and possible maintenance of tetraploidy (Varadharajan et al. 2018). Ssa16-17 had estimated rediploidization ages of 3–41 Ma in Salmoninae (spanning SSR1-5), but again showed assembly collapse in the European grayling genome (fig. 2b).

To summarize, this analysis reveals distinct waves of ancestral and lineage-specific rediploidization following Ss4R, separated by a period of comparative stasis (where just one genomic region underwent rediploidization) spanning 17–39 My, in addition to the existence of several homologous genomic regions in different salmonid subfamilies with markedly different rediploidization ages.

### Does Rediploidization Age Influence the Retention of Gene Functions?

We next asked if the functions of gene duplicates retained or lost after WGD is influenced by rediploidization age. Past work in Atlantic salmon (Robertson et al. 2017) and rainbow trout (Campbell et al. 2019) identified striking differences in functional enrichment among Ss4R ohnologs from genomic regions that experienced ancestral (i.e., AORE) or delayed rediploidization (i.e., SSR). We advanced these efforts using a



**FIG. 2.** Absolute rediploidization age estimation for ohnolog blocks retained from the Ss4R autoployploidization. (a) Onset of ohnolog divergence estimated by MCMCtree using concatenated genome alignments ([supplementary fig. 10, Supplementary Material online](#)). For each ohnolog block, the plotted circle is the posterior mean, and the dotted line is the 95% credibility interval. (b) Comparison of estimated rediploidization age in SSR regions for Atlantic salmon (red lines) and European grayling (gray lines). Gray crosses indicate regions of presumed assembly collapse in the European grayling genome due to highly delayed rediploidization or maintenance of tetraploidy.

higher resolution map of rediploidization history (fig. 1c) and an expanded set of Ss4R ohnolog pairs and singleton genes (where one ohnolog in a pair was lost during evolution) (Bertolotti et al. 2020). We extracted eight nonoverlapping gene sets from regions in the Atlantic salmon genome with distinct rediploidization ages, representing AORE (14,325 ohnologs; 5,887 singletons), SSR1 (3,140 ohnologs; 539 singletons), SSR2 (650 ohnologs; 78 singletons), and SSR3, 4, and 5 combined (hereafter: “SSR345”) (426 ohnologs; 162 singletons) (data in [supplementary data 12 and 13, Supplementary Material online](#)). We then tested for enrichment ( $P < 0.001$ ) in each set using GOslim terms (fig. 3 and [supplementary data 12, Supplementary Material online](#)). When using standard GO terms many functions are represented by a small number of genes, which may cause substantial biases comparing genes extracted from nonoverlapping genomic regions, as done previously (Robertson et al. 2017; Campbell et al. 2019). GOslim provides a much coarser description of gene functions, typically inclusive of hundreds to thousands of genes per term, which should circumvent this problem by ensuring each term is represented extensively even when the genes are drawn from nonoverlapping genomic regions.

For singletons, only the AORE set showed significant enrichment of GOslim terms, representing a small number of metabolic process and molecular functions that did not overlap with any ohnolog gene set (fig. 3a and b and [supplementary table 12, Supplementary Material online](#)). AORE ohnologs showed the largest number of overrepresented terms, most of which were not shared with other gene sets, including ohnologs from SSR categories (fig. 3a and b and [supplementary table 12, Supplementary Material online](#)). This likely reflects statistical power, as many more ohnologs are available in AORE than SSR regions. However, it may also capture the substantially larger evolutionary time for selection to act on duplicated genes in AORE regions (fig. 2a). In this respect, GOslim terms unique to AORE ohnologs included *mitotic cell cycle*, *enzyme regulator activity*, *kinase activity*, and *transcription factor binding*, which were enriched terms for ohnologs previously shown to have evolved dosage balance following early vertebrate WGD events (Makino and McLysaght 2010).

Few GOslim terms were shared between ohnologs from AORE and the SSR regions. However, most terms enriched for ohnologs from the three SSR categories were separately shared with AORE ohnologs (14 out of 16 for SSR1; 6 out of 8 for SSR2 and 5 out of 8 for SSR345), despite not being shared across SSR categories (fig. 3a and b). This may again reflect limited power due to the comparatively smaller number of genes available in SSR regions. However, GOslim terms showing enrichment for ohnologs in SSR regions that are not shared with AORE regions are more difficult to explain on the same grounds, and may be linked to selection on gene functions linked to rediploidization. For SSR1 ohnologs, this included the terms *extracellular matrix organization* and *cytoplasmic chromosome* and for SSR2 ohnologs, *plasma membrane organization* (fig. 3b). Strongly enriched terms for SSR345 ohnologs were *secondary metabolic process* in addition to *peroxisome* and *mitochondrion* (fig. 3b and [supplementary table 12, Supplementary Material online](#)).

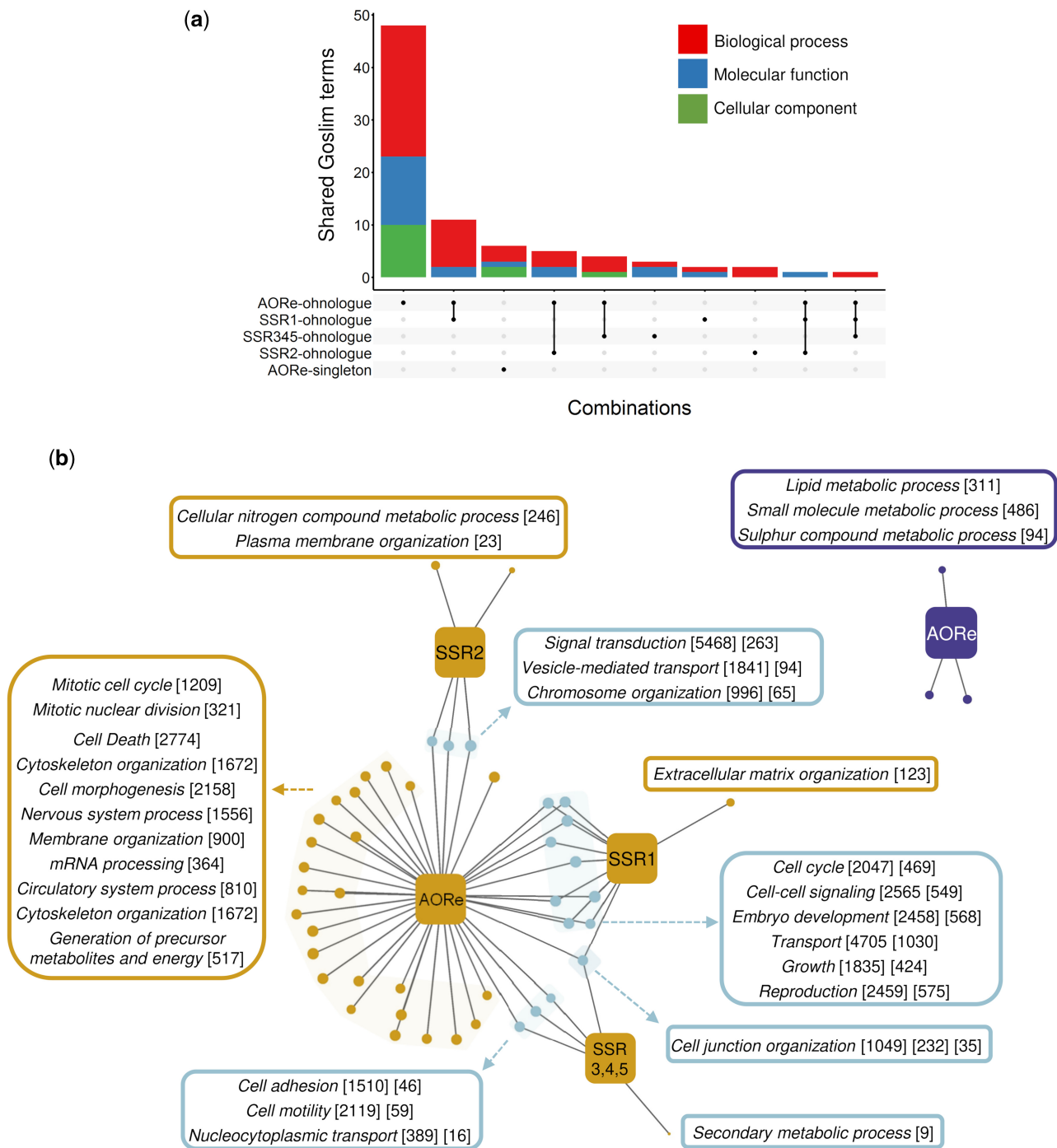
In summary, these results capture differences in functional enrichment between ohnologs and singletons, consistent with past investigations (Blomme et al. 2006; Makino and McLysaght 2010; Smet et al. 2013; Inoue et al. 2015; Han et al. 2016; Parey et al. 2020). In contrast to past work (Robertson et al. 2017; Campbell et al. 2019), our findings indicate common functional enrichment biases between Ss4R ohnologs with different rediploidization ages, highlighting the pitfalls of using gene set enrichment to compare nonoverlapping genomic regions. Finally, this work provides evidence for enrichment of a small number of unique functions for ohnologs from regions with different rediploidization ages.

### Rediploidization Age and Gene Expression Evolution

Both neutral evolution and selection have resulted in pervasive remodeling of gene expression after Ss4R (Lien et al. 2016; Sandve et al. 2018; Varadharajan et al. 2018; Gillard et al. 2021). However, our understanding of how rediploidization dynamics impact genome regulatory evolution, or vice versa, is limited. To explore whether rediploidization history influences gene regulatory evolution, we compared expression levels across a large set of ohnolog and singleton genes sampled from genomic regions with different rediploidization ages (SSR3-5 excluded due to small sample size) (fig. 4 and [supplementary data 14–17, Supplementary Material online](#)). For this we used RNA-Seq expression (transcript per million [TPM] data) from a panel of tissues shared by Atlantic salmon and northern pike. The TPM values for each gene were added together across all tissues to capture cumulative expression level. The salmon ohnolog expression data are represented in two ways: first as the pair-sum, where the TPMs were added together in each ohnolog pair (fig. 4b), and then individually, treating each copy as separate genes (fig. 4c and [supplementary data 14, Supplementary Material online](#)). The rationale for reporting both comparisons is to provide insights into the evolution of Ss4R ohnolog pair dosage, assuming selection acts on their total rather than individual expression level.

By examining pike orthologs, we established background expectations for gene expression in a nonduplicated genome (fig. 4a and [supplementary data 14 and 15, Supplementary Material online](#)). Across all rediploidization categories, pike orthologs of salmon ohnologs showed higher expression than pike orthologs of salmon singletons (fig. 4a). Assuming pike as a proxy for the ancestral state, Ss4R ohnologs are thus biased toward more highly expressed genes. The respective ratio of ohnolog-to-singleton median expression of pike orthologs was 1.24, 1.30, and 1.20 in AORE, SSR1, and SSR2 regions (fig. 4a). Considering the equivalent data for salmon ohnolog pair-sum expression levels, a much higher ratio (2.01, 2.44, and 2.38 in AORE, SSR1, and SSR2, respectively) was observed (fig. 4b). Although ohnolog pair-sum expression was higher than singletons in all rediploidization categories ([supplementary data 15, Supplementary Material online](#)), these ratios are less than twice the expression level for pike orthologs to AORE genes, and close to twice the level for pike orthologs of SSR1/2 genes. This indicates the evolution of lower total transcript dose for ohnolog pairs from AORE

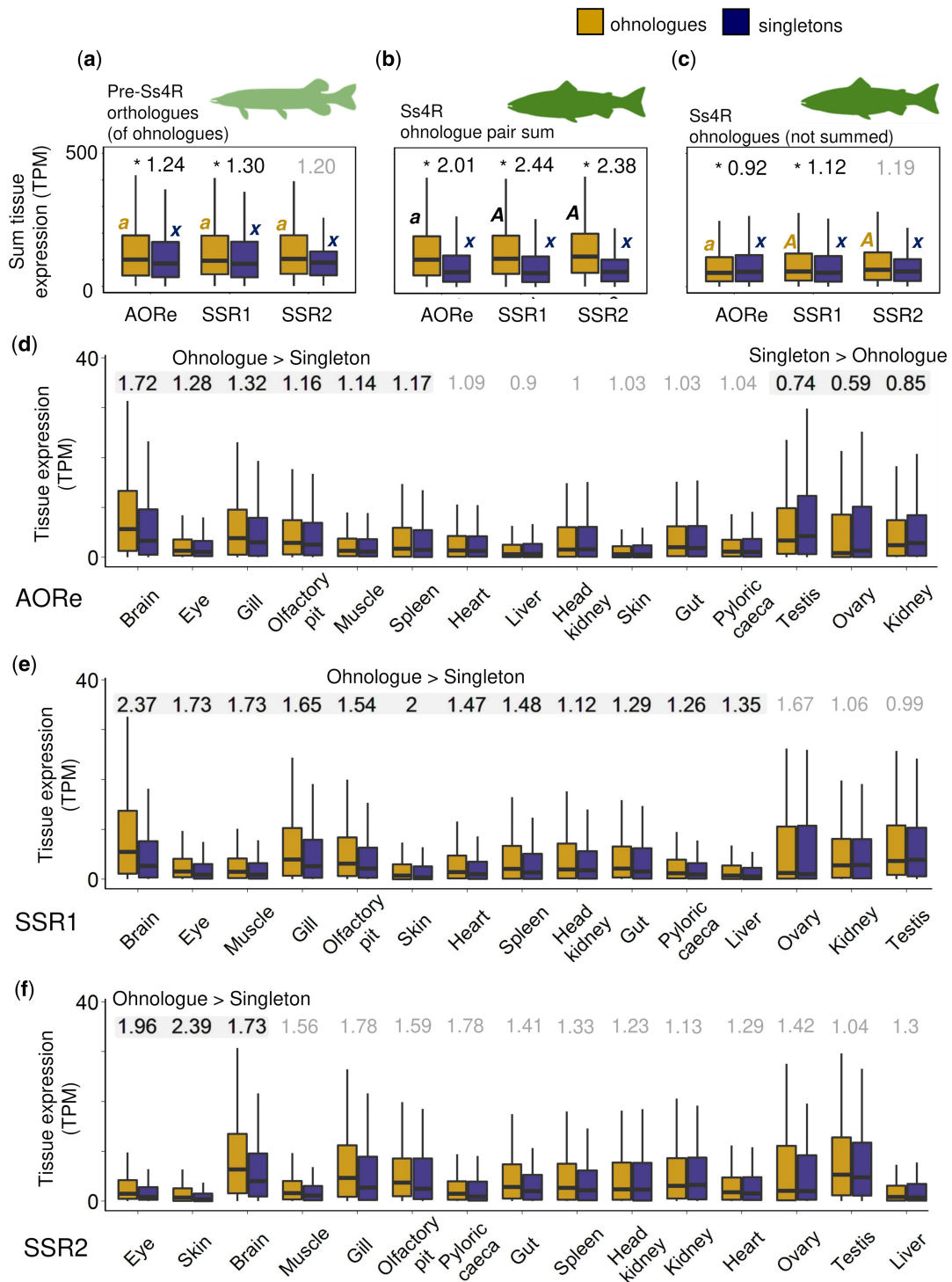




**Fig. 3.** Gene set enrichment analyses contrasting regions with distinct rediploidization ages in the Atlantic salmon genome. (a) Upset plot highlighting shared enriched Goslim terms among Ss4R ohnologs and singleton genes extracted from AORE and SSR regions. (b) Network visualizing shared enriched Goslim biological processes among the same categories. Each node represents a unique category (yellow nodes: Ss4R ohnologs; purple nodes: Ss4R singleton genes) in the genome and the lines extending from nodes connect to either category-specific (ending in yellow/purple circles for ohnologs/singletons) or category-shared Goslim terms (ending in turquoise circles). Examples of category-specific and category-shared Goslim terms are provided. Full data in [supplementary data 12, Supplementary Material online](#).

regions compared with SSR regions. Consistently, when we summarize TPM data from salmon ohnologs as independent genes, the ohnolog-to-singleton median expression ratio is closer to parity, and below that observed for pike orthologs (0.92, 1.12, and 1.19 in AORE, SSR1, and SSR2 regions, respectively) (fig. 4c). Interestingly, the expression of all salmon ohnologs is lower than singletons in AORE regions, whereas

ohnologs from SSR regions show higher expression than from AORE regions (fig. 4c and [supplementary data 15, Supplementary Material online](#)). Together, this analysis captures a dominant evolutionary trend of reduction in expression level for Ss4R ohnologs, consistent with a recent study (Gillard et al. 2021), with the most substantial reduction occurring in regions that rediploidized early.



**Fig. 4.** Rediploidization dynamics and gene expression evolution. (a–c) Boxplots of TPM values added together for a panel of tissues for genes from regions of the Atlantic salmon genome with unique rediploidization ages (figs. 1 and 2). (a) Northern pike data (panel of 13 tissues shared with Atlantic salmon), restricted to single-copy genes orthologous to salmon ohnologs and singletons (Bertolotti et al. 2020). (b) Atlantic salmon data for the same 13 tissues, adding together TPM values for each Ss4R ohnolog pair. (c) Atlantic salmon data for the same 13 tissues, with all ohnologs treated as a set of independent genes. Asterisks indicate a significant difference between ohnologs and singleton gene sets according to a Wilcoxon signed rank test ( $P \leq 0.01$ ). The ratio of median ohnolog versus singleton TPM is shown in black font when significant (gray font when not significant). Different letter cases (“a” vs. “A” for ohnologs—“x” vs. “X” for singletons) indicate a significant difference according to Wilcoxon signed rank test ( $P \leq 0.02$ ). (d–f) Boxplots of TPM values for 15 Atlantic salmon tissues (including the same 13 included in the global comparison) with other details as described in the legend for parts (a–c), except that we only compared differences between ohnologs and singletons per tissue using a Wilcoxon signed rank test ( $P \leq 0.01$  after Bonferroni–Holm correction). d = AORe, e = SSR1, f = SSR2. Tissues showing a significant difference are highlighted by light gray shading. The left-to-right position of tissues is ordered by  $P$  value (lowest to highest for ohnolog TPM > singleton TPM). Summary statistics and full information on statistical tests provided in [supplementary tables 14–17, Supplementary Material online](#).

Previous analyses have demonstrated that the regulatory evolution of gene duplicates is dependent on ancestral tissue expression context, with brain-biased ohnologs evolving under highest selective constraints (Lien et al. 2016; Varadharajan et al. 2018). We therefore also dissected the association between rediploidization history and gene expression levels for individual tissues, supporting clear tissue-specific differences (fig. 4d–f and supplementary data 16 and 17, Supplementary Material online). In AORE regions, despite the overall higher expression of singletons versus ohnologs (fig. 4c), the expression level of ohnologs in brain, eye, gill, olfactory pit, muscle, and spleen was higher than for singletons (fig. 4d and supplementary data 16 and 17, Supplementary Material online). Conversely, the opposite pattern was observed for testis, ovary, and kidney, that is, lower ohnolog expression than singletons (fig. 4d). By contrast, in SSR regions, no tissue had lower ohnolog expression level compared with singletons (fig. 4e and f). Instead, several tissues showed higher ohnolog expression levels than singletons, and the ohnolog to singleton expression level ratio was higher than for AORE regions in the same tissues (fig. 4d–f), consistent with the global analysis (fig. 4c). These results reinforce that expression evolution of Ss4R ohnologs is strongly shaped by tissue-specific selective constraints, in addition to rediploidization dynamics, while reiterating the trend of higher ohnolog transcript dose in late rediploidization regions.

### Positive Selection of Ohnologs in Regions with Distinct Rediploidization Ages

We previously hypothesized that LORe promotes lineage-specific adaptation by creating a substrate of “newly diverging” ohnologs that can functionally specialize in response to lineage-specific selection pressures (Robertson et al. 2017). We have further argued that selection on ohnolog functions retained in AORE regions may be comparatively constrained by ancestral divergence and specialization inherited prior to speciation (Robertson et al. 2017). To test these ideas, we asked if the number of ohnologs targeted by positive selection was a product of rediploidization age at three distinct periods of salmonid evolution. We employed an established method (Bertolotti et al. 2020; Gillard et al. 2021) to generate codon alignments including Ss4R ohnologs for all species used in our rediploidization analyses, along with additional teleost outgroups to Ss4R. After filtering, we retained 4,145 alignments, each harboring an Ss4R ohnolog pair retained across multiple salmonid species. This was broken down as 3,351, 709, and 85 alignments from AORE, SSR1, and SSR2 regions, respectively (see supplementary data 18, Supplementary Material online for genomic locations; alignments and trees provided in supplementary data 7, Supplementary Material online).

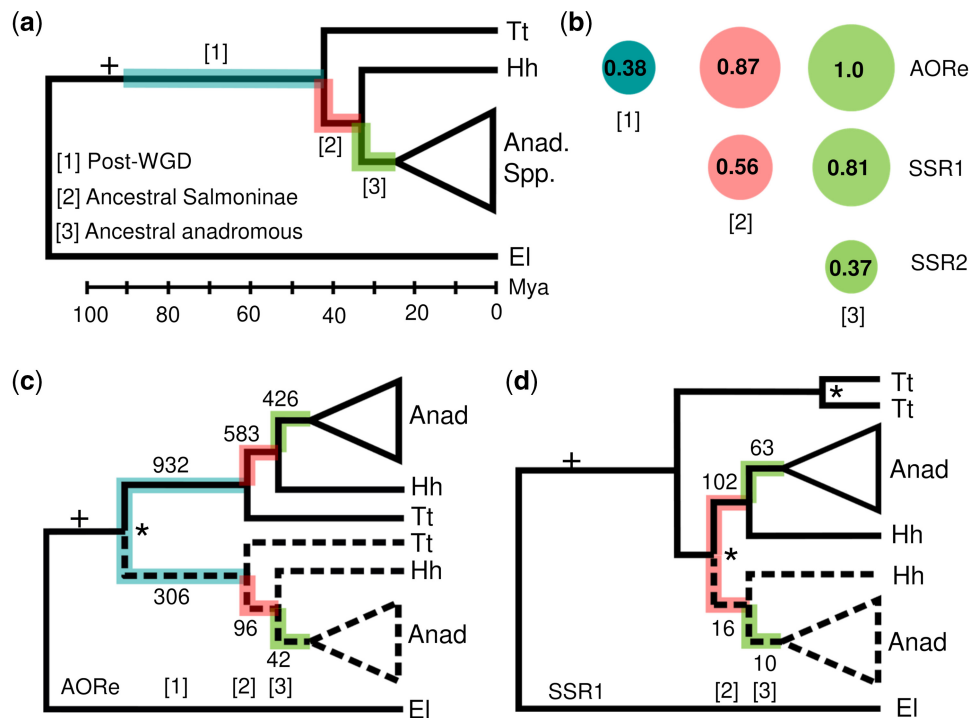
Each alignment was used in a  $d_N/d_S$  analysis employing an adaptive branch-site model (Smith et al. 2015). We documented ohnologs showing evidence for positive selection (corrected  $P < 0.05$ ) comparing each rediploidization age

category at three predefined phylogenetic branches: 1) “post-WGD,” separating the Ss4R event from the divergence of salmonid subfamilies, 2) “ancestral Salmoninae,” defining the common ancestor of Salmoninae species, and 3) “ancestral anadromous Salmoninae,” defining the common ancestor of *Salmo*, *Oncorhynchus*, and *Salvelinus* (fig. 5a and supplementary data 19 and 20, Supplementary Material online). This approach was designed to test a prediction of our hypothesis that more ohnologs will show positive selection in SSR regions than AORE regions along the tested branches.

For AORE regions, 16.4%, 12.9%, and 7.9% of the tested ohnologs showed evidence of positive selection along the post-WGD, ancestral Salmoninae, and anadromous Salmoninae branches, respectively. To facilitate data interpretation, we corrected for the effect of absolute time, that is, the post-WGD branch represents ~43 My evolution, compared with ~14.8 and 7.9 respective million years for the ancestral and anadromous Salmoninae branches (according to Macqueen and Johnston 2014). This correction indicates that 0.38%, 0.87%, and 1.0% of tested ohnologs in AORE regions experienced positive selection per million years along the three respective branches (fig. 5b). We observed a similar proportion of ohnologs under positive selection in SSR1 regions for ancestral Salmoninae branches (0.56% per million years) and ancestral anadromous Salmoninae (0.81% per million years) branches. For SSR2 regions, evidence for positive selection was obtained for just five ohnolog genes (i.e., anadromous Salmoninae branches; 0.37% per million years). As a caveat of our approach, we acknowledge that we do not know when selection occurred along the post-WGD branch, and the presented estimates through time will be misleading if the majority of positive selection occurred rapidly post-WGD, rather than steadily through time. Moreover, comparisons across branches are limited by the fact that the selected test has reduced power to detect positive selection along short compared with long branches (Smith et al. 2015).

To further interrogate how selection targeted ohnolog coding regions at different periods of salmonid evolution, we recorded cases where positive selection affected either one or both of the ohnologs in each pair comparing AORE and SSR1 regions (fig. 5c and d). In AORE regions, we observed evidence of positive selection acting on both ohnologs in a pair for ~25% of all post-WGD branches, a value higher than ancestral Salmoninae (~14%) or ancestral anadromous Salmoninae (~9%) branches (both  $P < 0.0001$ ; two-sided Fisher’s exact test). For SSR1 regions, positive selection acting on both ohnologs in each pair was inferred for a smaller number of branches than the AORE post-WGD branches in ancestral Salmoninae (~14%) and ancestral anadromous Salmoninae (~14%) branches ( $P = 0.0063$  and  $P = 0.034$ , respectively, two-sided Fisher’s exact test).

To summarize, these findings fail to support our previous hypothesis that LORe boosts adaptation through positive selection on duplicated coding regions, but the tests performed may be inherently limited by differences in power to detect positive selection between AORE and SSR regions.



**FIG. 5.** Positive selection on Ss4R ohnologs sampled from regions with distinct rediploidization ages across three periods of salmonid evolution. (a) Species tree showing lineages used in analysis (Tt = European grayling; Hh = huchen; Anad. Spp. = clade consisting of Atlantic salmon, brown trout, rainbow trout, Arctic charr, and Coho salmon) highlighting the three test branch categories. (b) Bubble plots comparing the percentage of ohnolog branches under positive selection (corrected  $P < 0.05$ ) in genomic regions with different rediploidization ages, normalized to millions of years. (c) and (d) Ohnolog trees depicting the respective topologies in AORe and SSR1 regions of the genome, highlighted with the number of branches under positive selection. The upper half (solid black line) and lower half (dotted black line) of each tree represents the respective number of branches where a single, or both, ohnologs in each pair was inferred to be under positive selection.

## Discussion

This study advances our understanding of the role played by rediploidization dynamics in long-term evolution following autopolyploidization. Our findings and the methodological advances reported have implications for future studies of WGD events. We have also generated useful resources for genomic and evolutionary investigations in salmonids, a group of fishes with extensive ecological and economic importance (Houston and Macqueen 2019; Houston et al. 2020) including a new genome for the endangered huchen (available via the Ensembl browser), genome alignments spanning multiple species, and positive selection data for a large set of ohnologs.

Delayed rediploidization and LORe are yet to be unequivocally demonstrated outside salmonids but were proposed to follow the teleost-specific WGD event (often called Ts3R) and WGD events at the base of vertebrates (Martin and Holland 2014; Robertson et al. 2017; Rozenfeld et al. 2019). There further exists a growing recognition that delayed rediploidization leading to LORe may have followed many of the large number of WGD events in plant evolution, most of which are not assigned to autopolyploid or allopolyploid origins (Clark and Donoghue 2017). Several authors have noted that delayed and nested patterns of species diversification following WGD, that is, the proposed “WGD Radiation Lag-Time Model” (Schranz et al. 2012) are consistent with LORe (Clark and Donoghue 2017; Van de Peer et al. 2017;

Carretero-Paulet and Van de Peer 2020; Li et al. 2021; Van de Peer et al. 2021). Considering the large and growing number of high-quality genome sequences in diverse eukaryotic lineages with a history of WGD, it will be feasible to adapt our phylogenomic approach to address the outcomes of rediploidization for WGD events of a comparable or younger age to Ss4R, exploiting synteny/collinearity to distinguish ohnologs from other gene duplicates. Plants offer a particularly useful lineage for such efforts, given that hundreds of WGD events have been inferred to date (Leebens-Mack et al. 2019). When studying older WGD events like Ts3R, genome alignment methods will be unsuitable due to the greater evolutionary distances involved. Nonetheless, mapping ohnolog gene tree topologies from multiple species that branched off early following an ancestral WGD, to a genome characterized by the same event, should allow for valid tests of LORe. Such an approach exploits the prior expectation that trees showing the same lineage-specific ohnolog divergence nodes will be co-located within chromosome regions sharing common rediploidization histories (fig. 1) more than expected by chance.

Most evolutionary studies of WGD events use approaches lacking scope to characterize delayed rediploidization. This results from a common assumption that genome-wide ohnolog divergence begins immediately after polyploidization, and will thus be ancestral to sister lineages sharing the same WGD event. On this basis, many authors use the distribution of

synonymous distance ( $K_S$ ) among all paralogous gene pairs within a genome to identify WGD events as peaks against the background distribution (Vanneste et al. 2013). A spectrum of ohnolog divergence ages following delayed rediploidization is not expected to generate a single peak, whereas distinct rediploidization “waves” (fig. 2a) are expected to generate multiple peaks (or tails to peaks), impacting the accuracy of such inferences. Consequently, delayed rediploidization adds to the known limitations of  $K_S$  methods (Vanneste et al. 2013; Tiley et al. 2018; Zwaenepoel and Van de Peer 2019). Gene tree—species tree reconciliation approaches are widely used to identify WGD events and likewise have known caveats (Thomas et al. 2017; Zwaenepoel and Van de Peer 2019). The complex patterns of LORe characteristic of large regions within salmonid genomes (fig. 1) strongly violate an assumption common to such methods—that ohnolog divergence starts along a specified branch in a species tree. To further compound this, our data show that ohnolog trees with the same (i.e., AORE) topology may be associated with different rediploidization ages (e.g., Ssa09-20 vs. other AORE regions; fig. 2a). This situation is invisible to gene tree—species tree reconciliation methods and has further implications for dating WGD events using molecular clocks. When faced with delayed rediploidization, genomic regions where rediploidization occurred earliest provide the best estimate for the timing of WGD—because they are “as close we can get” to the WGD event using sequence data (Macqueen and Johnston 2014). Past estimates for the timing of Ss4R either failed to exclude ohnologs showing delayed rediploidization (Berthelot et al. 2014; Lien et al. 2016), or lacked genome-wide information to identify differences in the timing of rediploidization for AORE regions (Macqueen and Johnston 2014). We thus propose that the timing of the Ss4R WGD should be treated as the genomic region with the oldest estimated rediploidization date, that is, 106 Ma (95% Bayesian credibility interval: 89–125 Ma) (fig. 2a). Although this is only marginally older than a previous reliable estimate based on 18 ohnolog pairs sampled from AORE regions (i.e., 88–103 Ma) (Macqueen and Johnston 2014), the issue would be inflated with more heterogeneity in the timing of ancestral rediploidization across genomic regions, which may be the case for other WGD events.

A past study of Ss4R considered all rainbow trout ohnologs classified here as belonging to SSR regions as tetrasomic (Campbell et al. 2019). It is well-established that tetrasomic inheritance still occurs in salmonid genomes, focused at telomeric regions and particularly impacting males (Allendorf et al. 2015). Our results warrant caution in confusing delayed rediploidization with an absence of rediploidization. Based on our data, tetrasomic inheritance is unlikely to have occurred for tens of millions of years in regions classified as tetrasomic in rainbow trout (Campbell et al. 2019). These regions have experienced ohnolog divergence descended across various Salmoninae clades predating the *Oncorhynchus* lineage, including all SSR1 and SSR2 regions (figs. 1 and 2). We would not expect truly tetrasomic regions to contain ohnologs showing sequence divergence. Instead, these regions should

harbor up to four alleles, and would collapse in haploid-representative genome assemblies.

Previous analyses revealed that both tissue expression bias (Lien et al. 2016) and rediploidization timing are important factors for ohnolog regulatory divergence in Ss4R ohnologs, with duplicates in SSR regions showing higher correlation in tissue expression than AORE regions (Robertson et al. 2017). Our results support these findings and reveal expression level as an additional factor to consider in ohnolog regulatory evolution. Higher ohnolog transcript dosage in SSR compared with AORE regions (fig. 4) may be partly explained by differences in the relative importance of drift and selection in regions with very different rediploidization ages. Ohnologs in AORE regions have been diverging as independent gene duplicates for tens of millions of years longer than SSR ohnologs, which were tetraploid for long periods of salmonid evolution (fig. 2a). As tetraploid loci have larger effective population sizes, the strength of selection on genes in AORE and SSR regions is expected to have been different for very long periods of salmonid evolution, with purifying selection limiting accumulation of deleterious regulatory mutations more effectively in tetraploid regions. On the other hand, when ohnologs undergo rediploidization, even mutations strongly downregulating expression will effectively be neutral if the total transcript “dose” does not exceed some critical lower threshold. Another explanation for higher expression of ohnologs in SSR compared with AORE regions is that selection pressure shifted after Ss4R to maintain high total transcript dose of duplicated genes in SSR regions. This predicts that ohnologs in SSR regions have functions that are particularly dosage sensitive, which could be tested in the future.

An unresolved question concerns the fundamental drivers of delayed rediploidization. It could be that rediploidization is strongly selected against in regions showing delayed rediploidization due to negative effects for certain genes or beneficial effects of maintaining genes as tetraploid. Analyses such as gene set enrichment and gene expression analyses lack resolution to resolve such possibilities, but provide clues into what is presumably a complex process. An illustrative example concerns the fact that ohnologs sampled from late rediploidization regions were highly enriched for organelle functions (*mitochondria* and *peroxisome*). This is the opposite result to past work in plants that showed organelle genes are more likely to be lost after WGD and hence were considered “duplication resistant” (Smet et al. 2013). It is thus interesting to ask, what does “duplication resistant” mean in the context of delayed rediploidization? In one respect, the long-term maintenance of tetraploidy acts to minimize protein-coding divergence (Lien et al. 2016) (fig. 2), perhaps akin to maintaining single copy genes, but on the other hand offers the chance for genes to maintain higher dosage than before WGD (fig. 4). Consequently, selection on duplicate gene retention and loss is likely multifaceted when delayed rediploidization is involved. For Ss4R, assuming selection acted to maintain tetraploidy in SSR regions across tens of millions of years (fig. 2), it is also important to ask why a second wave of

rediploidization was tolerated. As this coincides inextricably with the evolutionary origin of salmonid subfamilies, perhaps events leading to speciation coincided with reduced effective population size, lowering the efficiency of selection on deleterious impacts of rediploidization. Or perhaps novel selective pressures accompanying early species diversification (e.g., linked to the initial development of anadromy) altered selection on rediploidization and ohnolog divergence through other routes, due to effects on specific genes. Mechanistically, rediploidization is linked to a proliferation of TEs in the genome, which cause rearrangements driving the cessation of multivalent meiotic pairings, limiting ohnolog divergence (Soltis et al. 2015). As there are known bursts of TE activity throughout salmonid evolution (Lien et al. 2016), lineage-specific TE proliferation is perhaps causatively linked to delayed lineage-specific rediploidization. Unfortunately, the current generation of salmonid genomes do not allow such ideas to be tested due to their poor representation of TEs and genomic regions showing very recent rediploidization. Emerging long-read assemblies spanning all salmonid genera will allow for mechanistic insights into the relationship between TE evolution, speciation, and lineage-specific rediploidization.

In conclusion, our findings provide a useful model for delayed autopolyploid rediploidization and its macroevolutionary impacts. We advocate for in-depth investigations of rediploidization dynamics following many other eukaryotic WGD events, potentially demanding the uptake or creation of phylogenomic methods that better accommodate the expectations of delayed and lineage-specific rediploidization. Such work will be essential to define the prevalence and significance of LORe and delayed rediploidization in wider evolution.

## Materials and Methods

### Huchen Genome Assembly and RNA-Seq

Full methods are provided in [supplementary methods 1, Supplementary Material online](#). Briefly, sampling was done using genetically wild hatchery reared fish from the State Fisheries Farm Lindbergmühle, Germany. The embryo used for sequencing was generated from wild parents, with haploidy induced by UVC irradiation. Genomic DNA was extracted from a confirmed haploid ([supplementary fig. 1, Supplementary Material online](#)) and used to construct paired-end (500 bp insert) and mate-pair sequencing (~6 and ~12 kb insert libraries) (Heavens et al. 2015) libraries. Sequencing was done using an Illumina HiSeq2500 with 250 bp paired-end reads. Genome size was estimated using a k-mer approach (Vurture et al. 2017). Contig assembly, scaffolding, and gap filling were done using W2RAP-CS42\_TGACv1 (Clavijo et al. 2017), SSPACE v3.0 (Boetzer et al. 2011), and GapFiller 1.10 (Boetzer et al. 2011), respectively. Assembly completeness and quality was estimated using CEGMA v2.5 (Parra et al. 2007), BUSCO v3 (Waterhouse et al. 2018), and KAT tools v.2.3.4 (Mapleson et al. 2017). Repeat modeling and masking were performed using Repeatmodeler v1.0.9 (Smit and Hubley 2015) and

Repeatmasker v4.0.7 (Smit et al. 2015) ([supplementary table 3, Supplementary Material online](#)).

A total of 218 Gb RNA-Seq data (~720 million paired-end 150 bp reads) were generated to support annotation of the huchen genome (NCBI BioProject PRJNA480959). The samples represented 15 tissues from one individual (fork length: 30 cm, sex not identifiable), namely whole eye, whole mixed brain, swim bladder, gill filament, olfactory pit, skin, skeletal muscle, stomach, distal intestine, unidentified gonad, pyloric caeca, kidney, spleen, liver, and heart. We also generated liver and unidentified gonad data for a further three individuals (fork length: 28–31 cm; sex not identifiable). Total RNA was extracted using Trizol (Sigma) following the manufacturer's protocol. Library construction was carried out using an Illumina TruSeq RNA kit and sequencing performed on a HiSeq1500 platform by the Norwegian sequencing center. An annotated version of the huchen genome with 50,114 protein coding gene predictions is available on the Ensembl genome browser ([https://www.ensembl.org/Hucho\\_hucho](https://www.ensembl.org/Hucho_hucho)).

### Whole-Genome Alignment Capturing Ss4R Ohnologs

We developed an approach to circumvent the issue that genome alignment tools are geared toward orthologous regions and not designed to capture multispecies ohnolog variation (summarized in [supplementary fig. 14, Supplementary Material online](#)). This approach leverages prior knowledge of collinear/syntenic ohnolog blocks retained from WGD and inputs ohnolog sequence variation to the alignment algorithm as different “species,” allowing multispecies alignments to be generated inclusive of ohnologs. For AORE regions, genome assemblies for Atlantic salmon (Lien et al. 2016), European grayling (Varadharajan et al. 2018), huchen (this study), Chinook salmon (*O. tshawytscha*) (Christensen, Leong, et al. 2018), rainbow trout (Pearse et al. 2019), and northern pike (Rondeau et al. 2014) were used. For SSR regions, we added a *Salvelinus* genome (NCBI accession: GCA\_002910315.2) to increase scope to capture LORe outcomes.

The first step was to identify sequences homologous to the two established ohnolog blocks in Atlantic salmon (Lien et al. 2016) ([supplementary data 1 and 2, Supplementary Material online](#) for AORE and SSR regions, respectively) separately for each target species. Although chromosome-anchored sequence was used for the Atlantic salmon reference, we used scaffolds for other species to recover maximal data, either because no chromosome level assembly was available (e.g., huchen and European grayling), or a large number of scaffolds were not chromosome anchored. We generated BLASTn (Altschul et al. 1997) databases (using the makeblastdb module) for each defined ohnolog block in the Atlantic salmon genome (Lien et al. 2016) and performed per species BLASTn searches using genome scaffolds as queries (*e*-value cutoff: 0.001, maximum target sequences: 3, max\_hsps = 20,000, word size of 40 and minimum 90% sequence identity) before filtering the hits for minimum alignment length (<3,000 bp) and linearity using a published script (Christensen et al. 2018). This step captures high-quality scaffolds sharing close homology to the two Atlantic salmon

reference ohnolog sequences for each species, which were retrieved as fasta files using `fasta_tools` within MAKER v3.0 (Cantarel et al. 2008).

The next step involved splitting the recovered scaffolds in each species homolog set into two files representing the distinct ohnolog sequences. This is crucial to allow genome alignment tools to accept two ohnologs per species within the same alignment. For AORE regions, we performed an initial step to categorize scaffolds on the basis of putative 1:1 orthology to each Atlantic salmon ohnolog, which is possible due to ancestral rediploidization (Macqueen and Johnston 2014; Robertson et al. 2017). This was done by aligning all retrieved scaffolds per ohnolog block to a reference containing both Atlantic salmon ohnolog sequences using Mugsy v1.2.3 (Angiuoli and Salzberg 2011). We retrieved the number of alignment locations per scaffold to each Atlantic salmon ohnolog from the resultant MAF file using a custom script (supplementary methods 2, Supplementary Material online). This allows us to identify the likely orthologous scaffold to each Atlantic salmon ohnolog under the rationale that orthologous sequences share more alignment locations. In most cases, all scaffold alignment locations matched to a single Atlantic salmon ohnolog as expected. We excluded any (possibly chimeric) scaffolds where <70% of the alignment locations matched to a single Atlantic salmon ohnolog. At the end of this step, we split each set of scaffolds per species into two fasta files representing two ohnolog sets and renamed the fasta headers to represent orthology with one of the two Atlantic salmon ohnolog regions (i.e., “species abbreviation\_Atlantic salmon chromosome arm name\_scaffold name”).

For SSR regions, it is either challenging (due to recent ancestral rediploidization) or impossible (due to LORe) to identify 1:1 orthology for ohnolog sequences across salmonid species (Robertson et al. 2017). Consequently, we modified our approach to bin scaffolds homologous to the Atlantic salmon ohnolog blocks (supplementary data 2, Supplementary Material online) into two groups that individually contain no more than one ohnolog per species, allowing them each to be included in the genome alignment. To achieve this, all scaffolds homologous to each Atlantic salmon ohnolog pair were aligned to one of the Atlantic salmon sequences using Minimap2 v2.16 (Li 2018) (kmer size: 27;  $\leq 5\%$  sequence divergence allowed between subject and reference; other parameters default). Alignments were visualized in Tablet v.1.17.08.17 (Milne et al. 2010) allowing scaffolds to be binned into two groups of ohnolog per species based on shared overlap of two distinct ohnologs to a single Atlantic salmon reference. Where a single alignment was present, potentially due to assembly collapse (Varadharajan et al. 2018), scaffolds were randomly binned into one of the two ohnolog groups. We renamed the ohnologous scaffolds in the two bins as for AORE regions, except using either “Chr-A” or “Chr-B” designations to replace Atlantic salmon chromosome arm names.

The final step was to align all sequences homologous to each ohnolog block in Atlantic salmon across the different species. This was done using Mugsy v1.2.3 (Angiuoli and

Salzberg 2011), setting one of the Atlantic salmon ohnolog sequences as the reference, and then adding separate fasta files to the alignment for 1) the other Atlantic salmon ohnolog sequence, 2) two distinct sets of ohnologous scaffolds per salmonid species, and 3) a single set of coorthologous scaffolds for northern pike. The maximum distance along a single sequence to chain the anchors into a single local collinear block (LCB) was set to 2,000 bp, and the minimum span of aligned regions within LCBs was set to 100 bp (other parameters default). Summary statistics for the final MAF files produced for AORE and SSR regions (supplementary data 3 and 4, Supplementary Material online, respectively) were extracted using `mafStats` within `mafTools` v01 (Earl et al. 2014).

### Alignment Processing and Phylogenetics

We processed the alignment blocks within each MAF file to capture maximal useful information on ohnolog evolution for phylogenetic analysis. This was done by filtering on the basis of the number of represented sequences (a product of the different taxa plus Ss4R ohnologs captured) using a custom script (supplementary methods 2, Supplementary Material online). For AORE regions, alignments were filtered for a minimum of 9 out of 11 possible sequences in any block, meaning most salmonid species were required to retain two ohnologs, leading to a total of 92.2 Mb alignment blocks of which 34,603, 43,329, and 28,613 included 11, 10, and 9 sequences, respectively (supplementary data 5, Supplementary Material online). For SSR regions, we allowed a more inclusive filtering strategy to capture data in regions where rediploidization was most delayed (i.e., ohnologs have diverged the least) and assembly collapse (Varadharajan et al. 2018) and fragmentation is common. In SSR regions, up to 13 sequences were possible across different taxa and retained ohnologs; we recovered 15,313 alignments with >11 sequences represented, and a further 7,805 alignments with >9 sequences represented (supplementary data 6, Supplementary Material online).

The parsed MAF files were converted to fasta format using `Maffilter` v1.3.1 (Dutheil et al. 2014). The `fasta-splitter` script (Lam et al. 2015) was then used to split each alignment block per MAF file into individual fasta files. Each of the split fasta files was processed through `GBlocks` v0.91b (Castresana 2000) using default parameters to filter low-quality alignment regions. These finished alignments were used to construct maximum-likelihood phylogenetic trees in `IQTREE` v1.6.8 (Nguyen et al. 2015), using the best fitting substitution model (Kalyaanamoorthy et al. 2017) (selected separately for each alignment by comparing the fit of 88 different models) and ultrafast bootstrapping (Minh et al. 2013) with 1,000 iterations to obtain branch support values.

### Reconstructing Rediploidization History in SSR Regions

We matched expectations of LORe (Robertson et al. 2017) against empirical data on ohnolog divergence captured by phylogenetic trees sampled from SSR regions. All trees were assigned to one of the five possible SSR categories (SSR1-5; fig. 1b), facilitated using scripts executed in R (R core team,

2020) (supplementary methods 2, Supplementary Material online) followed by manual checking of every tree. All trees were rooted to northern pike, or European grayling when pike was absent from the alignment, or in rare situations where both species were absent, using midpoint rooting. To initially assign trees into different SSR categories, we exploited predicted monophyly for included species (fig. 1b), along with the *Dupfinder* function (Varadharajan et al. 2018) to confirm the position of nested ohnolog clades. This allowed us, for example, to initially assign trees to SSR1 on the criteria of 1) monophyly of European grayling and 2) a duplication node shared by all Salmoninae members; or to SSR2 on the criteria of 1) separate monophyly of both European grayling and huchen and 2) a duplication node shared by all ancestrally anadromous Salmoninae members (and so on for SSR3–5). After binning the trees into the five categories, all trees were plotted and manually visualized to check the automatic assignments. At this point, we removed trees showing unexpected branching and/or bootstrap values <50 along inferred rediploidization nodes. In some cases, we observed trees where a species showed monophyletic ohnolog sequences with strong bootstrap support, yet clustered within one of two ohnolog clades present elsewhere in the tree. This occurred commonly for trees assigned to SSR2 (i.e., two monophyletic huchen sequences branched as a sister to one of two ohnolog clades representing the ancestrally anadromous Salmoninae species). We accepted such topologies as consistent with SSR2. We present examples of accepted trees for each SSR category in supplementary figs. 15–19, Supplementary Material online, whereas all 11,136 trees used in fig. 1 are provided in supplementary data 7, Supplementary Material online.

A custom script (supplementary methods 2, Supplementary Material online) was used to retrieve Atlantic salmon chromosome coordinates for all SSR trees (supplementary data 8, Supplementary Material online). The data were visualized as circos plots using OmicCircos v1.26.0 (Hu et al. 2014) or Circlize v0.4.11 (Gu et al. 2014). Supplementary methods 3, Supplementary Material online describes the positional mapping of phylogenetic trees with different SSR topologies against brown trout chromosomes homologous to Atlantic salmon ohnolog blocks 03–06.

### Estimating Rediploidization Age across the Genome

To estimate rediploidization age across defined ohnolog blocks within the Atlantic salmon genome (Lien et al. 2016), we used a concatenation approach to maximize the available data. For AORE regions, all alignments per defined Ss4R ohnolog block (Lien et al. 2016) were concatenated using SeqKit v0.8.0 (Shen et al. 2016) to generate a single alignment file (described in supplementary data 10, Supplementary Material online). For SSR regions, we concatenated alignments across defined ohnolog blocks to capture each inferred SSR category (fig. 1b and supplementary figs. 2–10, Supplementary Material online), that is, separate concatenations were generated for regions inferred as SSR1, 2, 3, 4, and 5 (supplementary data 10, Supplementary Material online). Each sequence file was aligned using Mafft v7.0 (Katoh and Standley 2016) with default parameters. MCMCTree (within

PAML-v4.9h) (Yang 2007) was used to estimate rediploidization age, represented by the divergence time of ohnolog sequences (Macqueen and Johnston 2014), using approximate likelihood, an independent rate clock model (clock = 2) and the general time reversible substitution model (model = 7), allowing independent rates for all nucleotide substitutions. An input tree topology was used to set the appropriate rediploidization history (e.g., AORE, SSR1, etc.) and temporally constrained using uniform distribution priors (after Macqueen and Johnston 2014) (visualized in supplementary fig. 20, Supplementary Material online). Each analysis was allowed to run for 2,030,000 iterations with a burn-in value of 30,000 and sample frequency of 20 for both AORE and SSR alignments, leading to a final sample size of 100,000. The mean Bayesian divergence times and 95% credibility intervals estimated by MCMCTree for each rediploidization node were plotted using ggplot2 v3.3.2 (Wickham et al. 2016).

### Gene Set Enrichment Analyses

We used an R script (Supplementary methods 2, Supplementary Material online) to: 1) extract all annotated genes in the Atlantic salmon ICSASG\_v2 genome GFF file according to coordinates defining regions with distinct rediploidization ages—generating sets of genes combined across different AORE, SSR1 and SSR2 regions, and combining across all SSR3, 4, and 5 regions (respective coordinates for AORE and SSR regions given in supplementary data 1 and data 10, Supplementary Material online) and 2) cross-reference each gene set with Ss4R ohnologs and singletons defined in Bertolotti et al. (2020), removing any nonmatching genes. GO enrichment tests on these gene sets were done (pipeline described in https://gitlab.com/cigene/R/Ssa.RefSeq.db/-/wikis/go-slim), against all Atlantic salmon genes as the background. Enriched GOslim terms from each gene set were parsed to generate a list of unique and overlapping GOslim terms using an R script (supplementary method 2, Supplementary Material online), and the results visualized through a network graph generated in ggnetwork v0.5.8 (Briatte 2020).

### Gene Expression Analysis

TPM values for 13 northern pike tissues were retrieved from Lien et al. (2016). Unduplicated pike orthologs to a defined list of Atlantic salmon ohnologs and singletons (from Bertolotti et al. 2020) were parsed using orthogroup data available at https://gitlab.com/sandve-lab/defining\_duplicates and a custom R script was edited to capture details from the pike genome and different Ss4R rediploidization categories (script provided in supplementary method 2, Supplementary Material online). For the global analyses (fig. 4a–c), expression levels across 13 tissues were added together per gene separately for 1) pike orthologs to Atlantic salmon singletons, 2) pike orthologs to Atlantic salmon ohnologs, 3) Atlantic salmon singletons, 4) Atlantic salmon ohnologs, and 5) both ohnologs in each Atlantic salmon pair. The data were plotted using ggplot2 (Wickham et al. 2016) in R and compared across all different gene set combinations using Wilcoxon signed-rank tests in R. For the tissue-specific analyses (fig. 4d), TPM values for 15 Atlantic salmon tissues (Lien



et al. 2016) were retrieved using salmonfisher (<https://gitlab.com/sandve-lab/salmonfisher>). This analysis used the same 13 tissues as the comparison between pike and salmon orthologs, adding two tissues (ovary and skin), missing in the pike data. Otherwise, the analyses were treated in the same way, except treating the TPM values per tissue separately.

### Positive Selection Analyses

Curated protein-coding alignments and trees were generated using a published pipeline (Bertolotti et al. 2020) (scripts available at [https://gitlab.com/sandve-lab/salmonid\\_synteny](https://gitlab.com/sandve-lab/salmonid_synteny)). Briefly, Orthofinder v.2.4.0 (Emms and Kelly 2019) was used to generate orthogroups inclusive of Atlantic salmon, brown trout, rainbow trout, coho salmon, Arctic char, huchen, European grayling, northern pike, three spined stickleback, medaka, and Nile tilapia. Coding sequences were extracted from each orthogroup and aligned using Macse v.2.03 (Ranwez et al. 2011) before gene trees were generated using TreeBeST v.1.9.2 (<https://github.com/Ensembl/treebest>). The alignments were split into subtrees according to the presence of pike as an outgroup to salmonids. Trees falling within AOR regions (supplementary data 1, Supplementary Material online) were filtered by additional criteria: 1) the presence of the ancestral Ss4R node with bootstrap support >70 and 2) the branching of one or both the European grayling and huchen salmon orthologs in the correct position with bootstrap support >70. Trees representing SSR regions (supplementary data 2, Supplementary Material online) were manually categorized as SSR1 or SSR2. The codon alignments were used in adaptive Branch-Site Random Effects Likelihood tests (Smith et al. 2015) within command line HyPhy v2.5.9 (Kosakovsky Pond et al. 2005). Branch-specific  $d_N/d_S$  values and  $P$  values indicative of positive selection (corrected  $P < 0.05$ ) derived from a likelihood ratio test (Smith et al. 2015) were retrieved in tabular format for branches of interest (i.e., data in Fig. 5) using a custom R script (supplementary method 2, Supplementary Material online).

### Supplementary Material

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

### Acknowledgments

This work was supported by the Biotechnology and Biological Sciences Research Council (BBS/E/D/10002070) and the Frimmedbio program of the Research Council of Norway (241016). M.K.G. received studentship funding from a University of Aberdeen Elphinstone scholarship with additional support from the Government of Karnataka. We thank Dr Sebastian Beggel, Dr Bernhard C. Stoeckle, Jens-Eike Täuber, and Ms Haiyu Ding at the Aquatic Systems Biology Unit, Technical University of Munich for their support in sampling huchen. We thank Dr Torfinn Nome for supporting bioinformatic analyses. We thank Madhusudhan Gundappa (Twitter: @fish\_lines) for providing species illustrations in figure 1. We also thank Dr Gareth Gillard (Norwegian University of Life Sciences) for support with the RNA-Seq data. The

Earlham Institute performed library preparation and sequencing used in the huchen genome assembly.

### Author Contributions

M.K.G. and D.J.M. designed the research with inputs from S.R.S., D.H., S.A.M., and S.L. J.G. coordinated the huchen sampling and dissected fish. M.K.G. sampled huchen embryos and led the huchen genome assembly. M.K.G. developed the genome alignment approach and performed phylogenomic analyses with help from T.T.H. and L.G. M.K.G. and D.J.M. designed figures, tables, and the supplementary information. M.K.G. and D.J.M. cowrote the manuscript with inputs from all authors leading to the submitted manuscript.

### Data Availability

All data supporting the findings of this study are available within the paper and its supplementary files, including supplementary data provided through Figshare at <https://doi.org/10.6084/m9.figshare.14724684>. The huchen genome assembly is available through NCBI (accession number: GCA\_003317085.1; [https://www.ncbi.nlm.nih.gov/assembly/GCA\\_003317085.1/](https://www.ncbi.nlm.nih.gov/assembly/GCA_003317085.1/)) and the Ensembl Genome Browser ([https://www.ensembl.org/Hucho\\_hucho](https://www.ensembl.org/Hucho_hucho)). RNA-Seq data produced for huchen samples is available through NCBI BioProject: PRJNA480959 (<https://www.ncbi.nlm.nih.gov/bioproject/PRJNA480959>). Scripts used in the study are provided in supplementary method 2, Supplementary Material online or accessible through links in the Materials and Methods section.

### References

- Alexandrou MA, Swartz BA, Matzke NJ, Oakley TH. 2013. Genome duplication and multiple evolutionary origins of complex migratory behavior in Salmonidae. *Mol Phylogenet Evol.* 69(3):514–523.
- Allendorf FW, Bassham S, Cresko WA, Limborg MT, Seeb LW, Seeb JE. 2015. Effects of crossovers between homeologs on inheritance and population genomics in polyploid-derived salmonid fishes. *J Hered.* 106(3):217–227.
- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25(17):3389–3402.
- Angiuoli SV, Salzberg SL. 2011. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics* 27(3):334–342.
- Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noël B, Bento P, Da Silva C, Labadie K, Alberti A, et al. 2014. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat Commun.* 5:3657.
- Bertolotti AC, Layer RM, Gundappa MK, Gallagher MD, Pehlivanoglu E, Nome T, Robledo D, Kent MP, Røsaeg LL, Holen MM, et al. 2020. The structural variation landscape in 492 Atlantic salmon genomes. *Nat Commun.* 11(1):5176.
- Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y. 2006. The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol.* 7(5):R43.
- Blumstein DM, Campbell MA, Hale MC, Sutherland BJC, McKinney GJ, Stott W, Larson WA. 2020. Comparative genomic analyses and a novel linkage map for Cisco (*Coregonus artedii*) provide insights into chromosomal evolution and rediploidization across salmonids. *Genes Genom Genet.* 10:2863–2878.

- Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27(4):578–579.
- Briatte F. 2020. ggnetwork: Geometries to plot networks with “ggplot2.” Available from: <https://cran.r-project.org/package=ggnetwork>
- Campbell MA, Hale MC, McKinney GJ, Nichols KM, Pearse DE. 2019. Long-term conservation of ohnologs through partial tetrasomy following whole-genome duplication in salmonidae. *Genes Genom Genet.* 9:2017–2028.
- Campbell MA, López JA, Sado T, Miya M. 2013. Pike and salmon as sister taxa: detailed intraclade resolution and divergence time estimation of Esociformes+Salmoniformes based on whole mitochondrial genome sequences. *Gene* 530(1):57–65.
- Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Alvarado AS, Yandell M. 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* 18(1):188–196.
- Carretero-Paulet L, Van de Peer Y. 2020. The evolutionary conundrum of whole-genome duplication. *Am J Bot.* 107(8):1101–1105.
- Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol.* 17(4):540–552.
- Christensen KA, Leong JS, Sakhrani D, Biagi CA, Minkley DR, Withler RE, Rondeau EB, Koop BF, Devlin RH. 2018. Chinook salmon (*Oncorhynchus tshawytscha*) genome and transcriptome. *PLoS One.* 13(4):e0195461.
- Christensen KA, Rondeau EB, Minkley DR, Leong JS, Nugent CM, Danzmann RG, Ferguson MM, Stadnik A, Devlin RH, Muzzerall R, et al. 2018. The Arctic charr (*Salvelinus alpinus*) genome and transcriptome assembly. *PLoS One.* 13(9):e0204076.
- Cifuentes M, Eber F, Lucas M-O, Lode M, Chèvre A-M, Jenczewski E. 2010. Repeated polyploidy drove different levels of crossover suppression between homoeologous chromosomes in *Brassica napus* allohaploids. *Plant Cell.* 22(7):2265–2276.
- Clark JW, Donoghue PCJ. 2017. Constraining the timing of whole genome duplication in plant evolutionary history. *Proc R Soc B Biol Sci.* 284:20170912.
- Clavijo BJ, Accinelli GG, Wright J, Heavens D, Barr K, Yanes L, Di-Palma F. 2017. W2RAP: a pipeline for high quality, robust assemblies of large complex genomes from short read data. *bioRxiv.* 110999.
- Conant GC, Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet.* 9(12):938–950.
- De-Kayne R, Feulner PGD. 2018. A European whitefish linkage map and its implications for understanding genome-wide synteny between salmonids following whole genome duplication. *Genes Genom Genet.* 8:3745–3755.
- Du K, Stöck M, Kneitz S, Klopp C, Woltering JM, Adolfs MC, Feron R, Prokopov D, Makunin A, Kichigin I, et al. 2020. The sterlet sturgeon genome sequence and the mechanisms of segmental rediploidization. *Nat Ecol Evol.* 4(6):841–852.
- Dutheil JY, Gaillard S, Stukenbrock EH. 2014. Maffilter: a highly flexible and extensible multiple genome alignment files processor. *BMC Genomics.* 15:53.
- Earl D, Nguyen N, Hickey G, Harris RS, Fitzgerald S, Beal K, Seledtsov I, Molodtsov V, Raney BJ, Clawson H, et al. 2014. Alignathon: a competitive assessment of whole-genome alignment methods. *Genome Res.* 24(12):2077–2089.
- Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20(1):238.
- Furlong RF, Holland PWH. 2002. Were vertebrates octoploid? *Phil Trans R Soc Lond B.* 357(1420):531–544.
- Geist J, Kolahsa M, Gum B, Kuehn R. 2009. The importance of genetic cluster recognition for the conservation of migratory fish species: the example of the endangered European huchen *Hucho hucho* (L.). *J Fish Biol.* 75(5):1063–1078.
- Gillard GB, Grønvold L, Røsæg LL, Holen MM, Monsen Ø, Koop BF, Rondeau EB, Gundappa MK, Mendoza J, Macqueen DJ, et al. 2021. Comparative regulomics supports pervasive selection on gene dosage following whole genome duplication. *Genome Biol.* 22(1):103.
- Gu Z, Gu L, Eils R, Schlesner M, Brors B. 2014. Circlize implements and enhances circular visualization in R. *Bioinformatics* 30(19):2811–2812.
- Han Y, Li X, Cheng L, Liu Y, Wang H, Ke D, Yuan H, Zhang L, Wang L. 2016. Genome-wide analysis of soybean JmjC domain-containing proteins suggests evolutionary conservation following whole-genome duplication. *Front Plant Sci.* 7:1800.
- Heavens D, Accinelli GG, Clavijo B, Clark MD. 2015. A method to simultaneously construct up to 12 differently sized Illumina Nextera long mate pair libraries with reduced DNA input, time, and cost. *BioTechniques* 59(1):42–45.
- Houston RD, Bean TP, Macqueen DJ, Gundappa MK, Jin YH, Jenkins TL, Selly SLC, Martin SAM, Stevens JR, Santos EM, et al. 2020. Harnessing genomics to fast-track genetic improvement in aquaculture. *Nat Rev Genet.* 21(7):389–409.
- Houston RD, Macqueen DJ. 2019. Atlantic salmon (*Salmo salar* L.) genetics in the 21st century: taking leaps forward in aquaculture and biological understanding. *Anim Genet.* 50(1):3–14.
- Hu Y, Yan C, Hsu C-H, Chen Q-R, Niu K, Komatsoulis GA, Meerzaman D. 2014. OmicCircos: R simple-to-use R package for the circular visualization of multidimensional omics data. *Cancer Inform.* 13:13–20.
- Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet.* 11(2):97–108.
- Inoue J, Sato Y, Sinclair R, Tsukamoto K, Nishida M. 2015. Rapid genome reshaping by multiple-gene loss after whole-genome duplication in teleost fish suggested by mathematical modeling. *Proc Natl Acad Sci U S A.* 112(48):14918–14923.
- Jaillon O, Aury J-M, Brunet F, Petit J-L, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, et al. 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 431(7011):946–957.
- Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermini LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods.* 14(6):587–589.
- Katoh K, Standley DM. 2016. A simple method to control over-alignment in the MAFFT multiple sequence alignment program. *Bioinformatics* 32(13):1933–1942.
- Kucinski M, Fopp-Bayat D, Liszewski T, Svinger VW, Lebeda I, Kolman R. 2015. Genetic analysis of four European huchen (*Hucho hucho* Linnaeus, 1758) broodstocks from Poland, Germany, Slovakia, and Ukraine: implication for conservation. *J Appl Genet.* 56(4):469–480.
- Lam K-K, LaButti K, Khalak A, Tse D. 2015. FinisherSC: a repeat-aware tool for upgrading de novo assembly using long reads. *Bioinformatics* 31(19):3207–3209.
- Lecaudey LA, Schlieven UK, Osinov AG, Taylor EB, Bernatchez L, Weiss SJ. 2018. Inferring phylogenetic structure, hybridization and divergence times within Salmoninae (Teleostei: Salmonidae) using RAD-sequencing. *Mol Phylogenet Evol.* 124:82–99.
- Leebens-Mack JH, Barker MS, Carpenter EJ, Deyholos MK, Gitzendanner MA, Graham SW, Grosse I, Li Z, Melkonian M, Mirarab S, et al. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574:679–685.
- Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34(18):3094–3100.
- Li X, Guo B. 2020. Substantially adaptive potential in polyploid cyprinid fishes: evidence from biogeographic, phylogenetic and genomic studies. *Proc R Soc B Biol Sci.* 287:20193008.
- Li Z, McKibben MTW, Finch GS, Blischak PD, Sutherland BL, Barker MS. 2021. Patterns and processes of diploidization in land plants. *Annu Rev Plant Biol.* 72:387–410.
- Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, Hvidsten TR, Leong JS, Minkley DR, Zimin A, et al. 2016. The Atlantic salmon genome provides insights into rediploidization. *Nature* 533(7602):200–205.
- Macqueen DJ, Johnston IA. 2014. A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. *Proc Biol Sci.* 281(1778):20132881.

- Makino T, McLysaght A. 2010. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc Natl Acad Sci U S A*. 107(20):9270–9274.
- Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ. 2017. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics* 33(4):574–576.
- Martin KJ, Holland PWH. 2014. Enigmatic orthology relationships between Hox clusters of the African butterfly fish and other teleosts following ancient whole-genome duplication. *Mol Biol Evol*. 31(10):2592–2611.
- Mason AS, Wendel JF. 2020. Homoeologous exchanges, segmental allopolyploidy, and polyploid genome evolution. *Front Genet*. 11:1014.
- Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, Marshall D. 2010. Tablet—next generation sequence assembly visualization. *Bioinformatics* 26(3):401–402.
- Minh BQ, Nguyen MAT, von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. *Mol Biol Evol*. 30(5):1188–1195.
- Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 32(1):268–274.
- Ohno S. 1970. The enormous diversity in genome sizes of fish as a reflection of nature's extensive experiments with gene duplication. *Trans Am Fish Soc*. 99(1):120–130.
- Parey E, Louis A, Cabau C, Guiguen Y, Roest Crolius H, Berthelot C. 2020. Synteny-guided resolution of gene trees clarifies the functional impact of whole-genome duplications. *Mol Biol Evol*. 37(11):3324–3337.
- Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23(9):1061–1067.
- Pearse DE, Barson NJ, Nome T, Gao G, Campbell MA, Abadía-Cardoso A, Anderson EC, Rudio DE, Williams TH, Naish KA, et al. 2019. Sex-dependent dominance maintains migration supergene in rainbow trout. *Nat Ecol Evol*. 3(12):1731–1742.
- Pond SLK, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21(5):676–679.
- Qiao X, Li Q, Yin H, Qi K, Li L, Wang R, Zhang S, Paterson AH. 2019. Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome Biol*. 20(1):38.
- Ranwez V, Harispe S, Delsuc F, Douzery EJP. 2011. MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. *PLoS One*. 6(9):e22594.
- Robertson FM, Gundappa MK, Grammes F, Hvidsten TR, Redmond AK, Lien S, Martin SAM, Holland PWH, Sandve SR, Macqueen DJ. 2017. Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification. *Genome Biol*. 18(1):111.
- Rondeau EB, Minkley DR, Leong JS, Messmer AM, Jantzen JR, Schallburg KR, von Lemon C, Bird NH, Koop BF. 2014. The genome and linkage map of the Northern pike (*Esox lucius*): conserved synteny revealed between the salmonid sister group and the neoteleostei. *PLoS One*. 9(7):e102089.
- Rozenfeld C, Blanca J, Gallego V, García-Carpintero V, Herranz-Jusado JG, Pérez L, Asturiano JF, Cañizares J, Peñaranda DS. 2019. De novo European eel transcriptome provides insights into the evolutionary history of duplicated genes in teleost lineages. *PLoS One*. 14(6):e0218085.
- Sandve SR, Rohlfs RV, Hvidsten TR. 2018. Subfunctionalization versus neofunctionalization after whole-genome duplication. *Nat Genet*. 50(7):908–909.
- Schranz M, Mohammadin S, Edger PP. 2012. Ancient whole genome duplications, novelty and diversification: the WGD radiation lag-time model. *Curr Opin Plant Biol*. 15(2):147–153.
- Shen W, Le S, Li Y, Hu F. 2016. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One*. 11(10):e0163962.
- Simakov O, Marlétaz F, Yue J-X, O'Connell B, Jenkins J, Brandt A, Calef R, Tung C-H, Huang T-K, Schmutz J, et al. 2020. Deeply conserved synteny resolves early events in vertebrate evolution. *Nat Ecol Evol*. 4(6):820–830.
- Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31(19):3210–3212.
- Smet RD, Adams KL, Vandepoele K, Montagu MCEV, Maere S, Van de Peer Y. 2013. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proc Natl Acad Sci U S A*. 110(8):2898–2903.
- Smit AFA, Hubley R. 2015. RepeatModeler Open-1.0. 2008–2015. Seattle, USA: Institute for Systems Biology. Available from: <http://www.repeatmasker.org>
- Smit AFA, Hubley R, Green P. 2015. RepeatMasker Open-4.0. 2013–2015 289–300. Available from: <http://www.repeatmasker.org>
- Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Kosakovsky Pond SL. 2015. Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol Biol Evol*. 32(5):1342–1353.
- Soltis PS, Marchant DB, Van de Peer Y, Soltis DE. 2015. Polyploidy and genome evolution in plants. *Curr Opin Genet Dev*. 35:119–125.
- Thomas GWC, Ather SH, Hahn MW. 2017. Gene-tree reconciliation with MUL-trees to resolve polyploidy events. *Syst Biol*. 66(6):1007–1018.
- Tiley GP, Barker MS, Burleigh JG. 2018. Assessing the performance of Ks plots for detecting ancient whole genome duplications. *Genome Biol Evol*. 10(11):2882–2898.
- Van de Peer Y, Ashman T-L, Soltis PS, Soltis DE. 2021. Polyploidy: an evolutionary and ecological force in stressful times. *Plant Cell*. 33(1):11–26.
- Van de Peer Y, Mizrahi E, Marchal K. 2017. The evolutionary significance of polyploidy. *Nat Rev Genet*. 18(7):411–424.
- Vanneste K, Van de Peer Y, Maere S. 2013. Inference of genome duplications from age distributions revisited. *Mol Biol Evol*. 30(1):177–190.
- Varadarajan S, Sandve SR, Gillard GB, Tøresen OK, Mulugeta TD, Hvidsten TR, Lien S, Vøllestad LA, Jentoft S, Nederbragt AJ, et al. 2018. The grayling genome reveals selection on gene expression regulation after whole-genome duplication. *Genome Biol Evol*. 10(10):2785–2800.
- Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. 2017. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* 33(14):2202–2204.
- Waterhouse RM, Seppely M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. 2018. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol*. 35(3):543–548.
- Weiss H, Maluszynska J. 2000. Chromosomal rearrangement in autotetraploid plants of *Arabidopsis thaliana*. *Hereditas* 133(3):255–261.
- Wickham H, Chang W, Henry L, Pedersen TL, Takahashi K, Wilke C, Woo K. 2016. ggplot2: elegant graphics for data analysis. New York: Springer-Verlag. Available from: <https://ggplot2.tidyverse.org>
- Wolfe KH. 2001. Yesterday's polyploids and the mystery of diploidization. *Nat Rev Genet*. 2(5):333–341.
- Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24(8):1586–1591.
- Zwaenepoel A, Van de Peer Y. 2019. Inference of ancient whole-genome duplications and the evolution of gene duplication and loss rates. *Mol Biol Evol*. 36(7):1384–1404.