A high-resolution comparative atlas across 74 fish genomes illuminates teleost evolution after whole-genome duplication

Elise Parey[1,2], Alexandra Louis[1], Jérôme Monfort[2], Yann Guiguen[2], Hugues Roest Crollius[1*], Camille Berthelot[1,3*]

[1] Institut de Biologie de l'Ecole normale supérieure (IBENS), Département de Biologie, Ecole normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France.

[2] INRAE, LPGP, 35000, Rennes, France.

[3] *Current address*: Institut Pasteur, Université de Paris, CNRS UMR 3525, INSERM UA12, Comparative Functional Genomics group, F-75015 Paris, France.

* Corresponding authors

Abstract

Teleost fish are one of the most species-rich and diverse clades amongst vertebrates, which makes them an outstanding model group for evolutionary, ecological and functional genomics. Yet, despite a growing number of sequence reference genomes, large-scale comparative analysis remains challenging in teleosts due to the specifics of their genomic organization. As legacy of a whole genome duplication dated 320 million years ago, a large fraction of teleost genomes remain in duplicate paralogous copies. This ancestral polyploidy confounds the detailed identification of orthologous genomic regions across teleost species. Here, we combine tailored gene phylogeny methodology together with the state-of-the art ancestral karyotype reconstruction to establish the first high resolution comparative atlas of paleopolyploid regions across 74 teleost fish genomes. We show that this atlas represents a unique, robust and reliable resource for fish genomics. We then use the comparative atlas to study the tetraploidization and rediploidization mechanisms that affected the ancestor of teleosts. Although the polyploid history of teleost genomes appears complex, we uncover that meiotic recombination persisted between duplicated chromosomes for over 60 million years after polyploidization, suggesting that the teleost ancestor was an autotetraploid.

Keywords – whole genome duplication, chromosome evolution, rediploidization, gene phylogenies, comparative genomics

Introduction

Since the first teleost fish genome sequence was published in 2002 (Aparicio et al. 2002), fish genomics has massively contributed to our understanding of vertebrate genome function and evolution. As an early-diverging vertebrate clade, teleosts are at an ideal phylogenetic position to study deeply rooting vertebrate features. Notably, the conservation of regulatory circuits and developmental pathways has turned zebrafish, medaka and - to a lesser extend - platyfish into model species for human diseases (Wittbrodt et al. 2002; Lieschke and Currie 2007; Schartl 2014). In addition, fish have become popular in evolutionary, ecological and physiological genomics, illuminating processes such as environmental adaptation, species diversification, social behavior or sex determination (Rittschof et al. 2014; Capel 2017; Salzburger 2018; Kim et al. 2019; Xie et al. 2019; Greenway et al. 2020). Around two hundred fish species have reference genome assemblies, and many more are expected to become available in the next years (Rhie et al. 2021), bringing up the necessity to design comparative frameworks that will effectively permit to dissect the evolution and function of teleost genomes.

All teleost fish species descend from an ancient round of whole genome duplication (WGD), dated 320 Mya (Jaillon et al. 2004). This dramatic evolutionary event, referred to as the teleost-specific genome duplication (TGD), doubled all chromosomes and genes present in the teleost ancestor. The TGD has left a significant imprint on extant teleost genomes: while most duplicated genes have returned to a single-copy state, an important fraction of teleost genes remain in two copies, called ohnologs. For instance, 26% of all zebrafish genes are still retained as duplicated ohnologs (Howe et al. 2013). Evidence suggests that TGD duplicates have been involved in the evolution of functional innovations (Zakon et al. 2006; Moriyama et al. 2016; Escobar-Camacho et al. 2020), but it remains unclear to which extent differential gene retention has sustained the impressive phenotypic diversity of the teleost clade.

Importantly, the redundancy in fish genomes can be appreciated at the macrosyntenic level, where remnants of ancestrally duplicated chromosomes form runs of large duplicated regions (Postlethwait et al. 2000; Taylor et al. 2003; Jaillon et al. 2004). These pairs of sister regions have been called homeologs or double-conserved syntenic regions (DCS). The precise identification and delimitation of fish DCS regions remains elusive, as WGDs present severe challenges to both gene phylogeny and ancestral genome reconstruction methodologies (Parey et al. 2020). Previous DCS characterizations were therefore limited to regions of highly conserved gene order or small species sets: the largest multi-species dataset comprised eight fish species and included ~29% of all genes in DCS (Conant 2020). A high-resolution, genome-wide identification of duplicated regions across a large number of fish genomes is therefore desirable, not only as an additional support for orthology predictions, but also as a lever to investigate the genomic and functional implications of the TGD.

WGD can arise through two mechanisms: autopolyploidization (polyploidization within or between populations of a single species) or allopolyploidization (polyploidization through hybridization of related species), with different consequences on subsequent genome evolution (Stebbins 1947; Mason and Wendel 2020). In particular, autopolyploidization and allopolyploidization differentially shape the rediploidization process, i.e. how the polyploid genome returns to a largely diploid state over time. Due to initial similarity amongst the different copies of duplicated chromosomes, young autopolyploid genomes are characterized by multivalent meiotic behaviour and tetrasomic inheritance. As a result, recombination and gene conversion can occur over whole homeologous chromosomes for millions of years following genome duplication in ancient autopolyploid lineages (Robertson et al. 2017; Gundappa et al. 2021), but these events are rarer and more localized after allopolyploidization, and probably dependent on genomic divergence between the parental genomes (Lloyd et al. 2018; Li et al. 2021). The retention of redundant gene copies is also affected by the nature of the polyploidization event. After allopolyploidization, one of the two subgenomes often loses more genes than the other (Garsmeur et al. 2014; Session et al. 2016; Cheng et al. 2018), although not always (Sun et al. 2017). This unequal gene retention has been linked to differences in transposable element repertoires and epigenetic silencing in the two subgenomes. Such differences can lead to reduced expression levels and relaxed selective pressure biased towards one of the subgenomes, which then accumulates more gene losses. In contrast, in the case of autopolyploidy, gene losses are expected to affect both homeologs equally due to their high similarity. It is unclear at this time whether the teleost whole-genome duplication corresponds to an ancestral autotetraploidization or an allotetraploidization event (Martin and Holland 2014; Conant 2020).

Here, we apply a phylogenetic pipeline specifically developed for WGD events (Parey et al. 2020) to retrace the evolutionary history of the genes and chromosomes of 74 teleost species encompassing most of the major fish clades. We combine these phylogenetic trees with the latest ancestral reconstruction of the pre-TGD ancestral karyotype (Nakatani and McLysaght 2017) to build a comprehensive comparative atlas of TGD-duplicated regions in teleost fish genomes. We then leverage this comparative atlas of teleost genomes to investigate the mechanisms of polyploidization and rediploidization that underlie the teleost genome duplication.

Results

Construction of TGD-aware teleost gene trees

We collected a dataset of 101 vertebrate genomes, including 74 teleost fish, 2 non-teleost fish (bowfin and spotted gar, which did not undergo the teleost-specific genome duplication event), 20 other vertebrates, of which 6 mammals, and 5 non-vertebrate genomes (Supplementary Figure S1; Supplementary Table S1). We used TreeBeST to reconstruct the phylogenetic relationships of 26,692 gene families across those 106 genomes (Methods; (Vilella et al. 2009; Herrero et al. 2016)). We then applied SCORPiOs to improve the placement of the TGD in these gene trees, using the bowfin and the spotted gar as reference outgroups. Briefly, SCORPiOs leverages syntenic gene neighborhoods to distinguish WGD-descended orthologs from paralogs when sequence information is inconclusive, and accurately positions WGD duplication nodes in genes trees (Parey et al. 2020). We identified 8,144 teleost gene subtrees out of 17,493 (47%) that were inconsistent with synteny information, of which 5,611 could be corrected (32% of all subtrees). We note that the corrected-to-inconsistent tree ratio (69%) is comparable to our previous application of SCORPiOs to fish data. Similarly, the proportion of errors in initial sequence-based gene trees is in line with our previous application of SCORPiOs to a dataset of 47 teleost species (Parey et al. 2020). These 26,692 WGD-aware teleost gene trees predict that the ancestral genome of teleost fish contained 46,206 genes after the duplication event, in line with the latest estimates from the Ensembl database (49,255 ancestral teleost genes in release v100; Methods).

A high-resolution atlas of the TGD duplication across 74 teleost genomes

Teleost fish genomes are mosaics of duplicated regions, formed through rearrangements of duplicated ancestral chromosomes (Supplementary Figure S2A). Long-standing efforts have been made to reconstruct the ancestral teleost karyotype before the whole-genome duplication (Jaillon et al. 2004; Kasahara et al. 2007; Nakatani and McLysaght 2017). According to the state-of-the-art reference, this ancestral teleost karyotype comprised 13 chromosomes (Nakatani and McLysaght 2017). Nakatani and McLysaght delineated between 353 and 690 megabase-scale genomic regions that descend from these 13 ancestral chromosomes in zebrafish, tetraodon, stickleback and medaka.

We combined this pre-TGD ancestral karyotype with our gene trees to identify regions and genes that descend from sister duplicated chromosomes across all 74 teleosts in our dataset (Methods; Supplementary Figure S2). First, we transformed the reference segmentations of the zebrafish, tetraodon, stickleback and medaka genomes from 13 to 26 ancestral chromosomes after the duplication (1a, 1b, …, 13a, 13b). In each genome, we iteratively grouped regions from an ancestral chromosome into two post-duplication copies by minimizing

intragroup paralogs (Methods; Supplementary Figure S2D). To assess robustness, we performed 100 groupings with random restart and found that genes were assigned to the same chromosome copy in 80% of iterations on average. We then identified orthologous ancestral chromosomes across all four species using gene orthologies, and arbitrarily named one of each pair `a` and `b` (Supplementary Figure S2C).

Next, we propagated these ancestral chromosome annotations to the other 70 teleost genomes using gene orthology relationships (Supplementary Figure S2D). For 1,303 gene trees (7%), orthologs from the four reference species were not assigned to the same ancestral chromosome, and we resolved these inconsistencies by assigning all orthologous genes, including those of the reference species, to the most represented ancestral chromosome. This process results in a 74-species comparative genomic atlas with genes annotated to post-duplication chromosomes, along with fully resolved orthology and paralogy links between all included species (Supplementary Figure S2E). This comparative atlas integrates 70% to 90% of each genome into 24,938 post-duplication gene families (Figure 1A-B), and greatly improves upon the state-of-the-art both in terms of species and genomic coverage. The atlas is available on the Genomicus database webserver (see Data availability and implementation).

Quality checking the teleost genome comparative atlas

As a quality check, we assessed discrepancies between the pre-TGD karyotype reconstruction by (Nakatani and McLysaght 2017) and our ancestral chromosome assignations. Both should be globally congruent because the ancestral karyotype provides the groundwork for the comparative atlas, but discrepancies can arise when orthologous genes are assigned to different pre-duplication chromosomes between the four reference species. In this case, we resolve the inconsistency by reassigning all orthologs to the most represented pre-duplication chromosome, which will be different from the original assignation in (Nakatani and McLysaght 2017) for at least one species. We found that 9% of zebrafish genes differ in ancestral chromosome assignation between our comparative map and the pre-TGD karyotype, versus 2-3% in medaka, tetraodon or stickleback. This likely reflects small-scale rearrangements in zebrafish captured by our individual gene trees but missed by the macrosyntenic approach of (Nakatani and McLysaght 2017). Alternatively, zebrafish genes may be more frequently placed incorrectly in our gene trees, which may lead to erroneous ancestral chromosome reassignations. To explore this possibility, we identified gene trees that remain synteny-inconsistent after correction by SCORPiOs, and potentially contain errors. Zebrafish genes reassigned to a different ancestral chromosome in our comparative map are not over-represented in synteny-inconsistent trees (7% vs 14% for all zebrafish genes), suggesting that their orthology relationships and chromosomal reassignations are overall well-supported by their sequences and syntenic gene neighborhoods.

175

Additionally, we used random noise simulations to demonstrate that the comparative atlas is robust to potential uncertainty or errors in the original ancestral genome reconstruction. The delineation of genomic regions descended from each pre-duplication chromosome from
180 (Nakatani and McLysaght 2017) relies on the identification of conserved synteny blocks and their breakpoints between teleost genomes and outgroups. Because breakpoint locations are challenging to determine - as also attested by lower posterior probabilities close to breakpoints in (Nakatani and McLysaght 2017) - the region boundaries can vary in precision. We shifted the positions of these boundaries with increasingly large errors, mimicking situations where up
185 to 25% of genes change pre-duplication chromosome assignations in each of the reference genomes (Methods). We found that even large errors in region boundaries did not majorly affect the final atlas, with only 11% of genes changing ancestral chromosome assignations at the highest noise settings (Supplementary Figure S3).

Finally, we report that correcting gene trees with SCORPiOs had a decisive impact on the
190 establishment of the comparative atlas, enabling the inclusion of a significantly larger fraction of teleost genes (83% vs 61%, Supplementary Figure S4). The teleost comparative genomic atlas represents therefore a reliable, comprehensive and robust resource for fish genomics.

The comparative atlas enhances teleost gene and genome annotations

195 We next investigated how the comparative atlas may improve crucial resources for fish evolutionary, ecological and functional genomics. The Zebrafish Information Network (ZFIN) provides manually curated, high-quality reference annotations for zebrafish genes and implements rigorous conventions for gene naming (Ruzicka et al. 2019). These gene names are then propagated to orthologous genes in other teleost genomes, providing the basis of the
200 entire teleost gene nomenclature and functional annotation transfers. In an effort to convey evolutionary meaning within the gene names, zebrafish paralogs descended from the TGD are identified with an ‘a’ or ‘b’ suffix. ZFIN guidelines recommend that adjacent genes should carry the same suffix when they belong to a continuous syntenic block inherited since the TGD, sometimes called syntelogs (Zhao et al. 2017). This aspiration has however been difficult to
205 implement in the absence of a high-resolution map of zebrafish duplicated regions and their ancestral chromosomes of origin. To assess the consistency in consecutive zebrafish gene names, we extracted and compared zebrafish ‘a’ and ‘b’ gene suffixes along duplicated regions from the comparative atlas (Figure 2). We find that despite previously mentioned efforts, current zebrafish ‘a’ and ‘b’ suffixes are not consistent with the polyploid history of the zebrafish
210 genome (Figure 2A), and that 43% of ‘a’ and ‘b’ gene suffixes should be reassigned in order to reflect the shared history of syntenic genes (Methods). Additionally, the ZFIN nomenclature

currently does not impart a suffix to singleton genes, which correspond to TGD-duplicated genes where one of the copies was eventually lost. As a result, only 26% of zebrafish genes are annotated with an 'a' or 'b' suffix in ZFIN. Using the comparative atlas, we were able to
215 annotate which paralogs were retained for 84% of all zebrafish genes, significantly extending the evolutionary annotation of gene histories in this species (Figure 2B). The comparative atlas also provides similar annotations for all 74 studied teleosts, as shown for medaka, stickleback and tetraodon (Supplementary Figure S5). How ancient tetraploids return to a mostly diploid state is an active area of research, where distinguishing which paralog gene has been retained
220 can be essential (Inoue et al. 2015; Robertson et al. 2017; Conant 2020; Simakov et al. 2020; Gundappa et al. 2021). The comparative atlas opens the possibility to systematically extend suffix annotations to singleton genes and formally identify which ancestral copies have been retained and lost along the evolution of teleosts. Altogether, the comparative atlas is a key resource to include biologically meaningful, historically accurate insight into reference gene
225 annotations and support further investigations of teleost genome evolution.

The teleost duplication was followed by delayed meiosis resolution

We took advantage of the comparative genomic atlas to analyze early genome evolution after the TGD. In particular, we investigated whether meiosis was fully resolved with bivalent
230 chromosomal pairing by the time Osteoglossiform and Clupeocephala lineages diverged 267 million years ago (Kumar et al. 2017), or whether some duplicated genomic regions still recombined during meiosis. Sequence exchanges due to prolonged recombination are detectable because they delay the divergence of duplicated regions until after meiosis is fully resolved. Therefore, if recombination persisted locally when speciation occurred, ohnologs
235 from these genomic regions will be mis-identified as clade-specific duplications, as they are more similar in sequence within clades than across clades (Robertson et al. 2017; Gundappa et al. 2021).

For this analysis, we selected a subset of six teleost genomes *(*paramormyrops*,* arapaima*,* Asian arowana*,* zebrafish*,* medaka and stickleback*)* and all previously integrated outgroups to
240 ensure an equal representation of Osteoglossiform and Clupeocephala genomes in the dataset (Methods, Supplementary Figure S6). We leveraged the gene tree correction applied by SCORPiOs to identify genes that experienced delayed meiosis resolution, where sequence analysis identifies duplication events in each clade after speciation while synteny suggests a more ancient, common gene duplication from the WGD event. Such sequence-synteny
245 conflicts are directly revealed by SCORPiOs and correspond to gene trees for which synteny-guided correction induces a significant drop in likelihood estimated from sequence divergence. We grouped gene trees with sequence-synteny conflicts according to their ancestral chromosome of origin in the comparative atlas. Three anciently duplicated chromosome pairs

(3, 10 and 11) are significantly enriched in sequence-synteny conflicts, thus hinting towards
250 prolonged recombination between these homeologs after the TGD (Figure 3A; $p < 0.05$, hypergeometric tests corrected for multiple testing).

Manual inspection of these homeologs indeed suggested delayed meiosis resolution after the Osteoglossiform/Clupeocephala lineage split. We therefore explicitly confronted the likelihoods for the tree topologies expected under early (before the lineage split) or late (after the split)
255 meiosis resolution for 5,557 gene trees which retain both ohnologs in at least one of the descending lineages (Figure 3B; Methods). For 638 gene trees, the early resolution topology was significantly more likely, while the late resolution topology was favored for 1,361 trees (likelihood AU-tests, alternative topology rejected at $\alpha=0.05$; no significant differences for the remaining 3,558 trees). For example, the *col12a1* paralogs had stopped recombining and
260 accumulated independent substitutions by the time Osteoglossiforms and Clupeocephala diverged (Figure 3C). On the other hand, the *map1a* TGD paralogs presumably still recombined when the taxa diverged, and both paralogs diverged independently later on in each lineage. We visualized the location of these genes on the medaka karyotype, which remains close to the ancestral teleost karyotype, revealing chromosomal regions of early and
265 late meiosis resolution (Figure 3D). This snapshot of the rediploidization status at the time of the Osteoglossiforms/Clupeocephala lineage split confirms that ancestral chromosome pairs 3, 10 and 11 were subjected to prolonged recombination (Figure 3D). In summary, we provide the first evidence that entire chromosomes experienced delayed meiosis resolution and rediploidization in teleosts and continued to exchange genetic material between homeologs for
270 at least 60 million years after the teleost whole genome duplication.

The teleost genome experienced biased gene retention after duplication

The prolonged exchange of genetic material between duplicated chromosomes after the TGD suggests an autopolyploid origin for teleosts. To challenge this hypothesis, we tested whether
275 one chromosome copy systematically retained more genes than the other, as observed after several instances of allopolyploidization (Garsmeur et al. 2014). Indeed, previous observations on a limited subset of genes suggest that the ancestral teleost genome may have experience such biased retention after the TGD (Conant 2020). We computed gene retention in non-overlapping 10-gene windows along duplicated chromosome copies in several teleost
280 genomes, using an outgroup fish as proxy for the ancestral gene order (Figure 4A-B, Supplementary Figures S7-S8, Methods) and total gene retention on homeologs in the ancestor (*Osteoglossocephalai*) of all 74 teleosts (Supplementary Table S2). We uncover a consistent, significant bias in gene retention on ancestral chromosome pairs 3, 4, 7, 11 and 13, where genes were preferentially retained on one homeolog over the other, regardless of
285 study species ($p < 0.05$, Wilcoxon paired tests corrected for multiple testing). We find additional

but weaker evidence for biased gene retention on ancestral chromosome pairs 2, 5, 6, 8 and 9 in some combinations of genome comparisons. We however do not find evidence of retention bias between ancestrally duplicated copies for chromosome pairs 1, 10, 12. This preferential gene retention is not an artifact of the atlas construction due to genes unassigned to either duplicated chromosome: indeed, conservatively assigning such genes to the homeolog with the lowest retention did not offset the retention imbalance (Supplementary Table S3). In summary, and against expectations from results on delayed meiotic recombination, we find that genes were preferentially retained on one duplicated chromosome in a subset of ancestral chromosomes, a character more frequently observed in allopolyploids (Garsmeur et al. 2014; Cheng et al. 2018). Of note, the chromosome copy with highest gene retention is typically annotated as 'a' in the atlas as a consequence of the construction process – all 'a' chromosomes may not originate from the same parental subgenome and should not be interpreted as such.

Differences in selective pressures and gene expression do not explain the observed bias in genes retention

Previous observations have suggested that biased gene retention in allopolyploid genomes reflects partial epigenetic silencing of one parental subgenome (Freeling et al. 2012; Bird et al. 2021). Genes on silenced chromosome copies are expressed at lower levels, and therefore subjected to lower selective pressures and ultimately to pseudogenization. We therefore investigated whether anciently duplicated chromosomes copies that retained fewer genes have decayed because of unequal selective pressure. We estimated dN/dS ratios for 1,269 pairs of TGD-derived paralogs conserved in at least 40 species, and tested whether ancient duplicated chromosomes display significant differences in dN/dS (Methods). Ancestral chromosome pairs 2, 3, 4, 11 and 13 experienced significantly faster gene evolution on one chromosome copy compared to the other (Figure 4C). Four out of five of these chromosome pairs also exhibit strongly biased genes retention ($n = 13$; $p = 0.03$; Fisher's exact test); however, genes under lower selective pressure (high dN/dS) were located on the homeolog with lower retention for ancestral chromosome pairs 3 and 4, as predicted if low selection promotes gene loss, while the opposite is true for ancestral chromosome pairs 11 and 13. Thus, differences in selective pressures are inconsistent with the observed bias in gene retention.

In addition, we investigated whether ancestral chromosome copies exhibit differences in gene expression, which would have been established after the TGD and have perdured since. We find no bias in ohnolog gene expression between ancestral chromosomes from each pair. This result is consistent whether looking at average gene expression across 11 tissues in medaka

(Figure 4D), zebrafish (Supplementary Figure S9), or tissue by tissue in either medaka and zebrafish (Supplementary Figure S10-S11).

In conclusion, our findings are consistent with previous reports that gene retention on ancestral chromosomes was biased following the teleost duplication (Conant 2020). However, while this observation was previously interpreted as evidence for an allopolyploid origin of teleost fish, our additional results suggest that genome evolution after the TGD does not display other correlated traits classically expected in allopolyploids. Together with our evidence that ancestral duplicated chromosomes were similar enough to maintain meiotic recombination over millions of years, our results are more consistent with an autopolyploid origin of teleosts.

Discussion

As sequencing technologies progress, data analysis and integration are becoming the major roadblocks to extract the rich information that comparative genomics can provide on species evolution and genome function. Teleost fish have a long-standing history as tractable model species for vertebrate development and human disease (Ohno et al. 1968; Streisinger et al. 1981; Haffter et al. 1996; Lieschke and Currie 2007; Schartl 2014), and have contributed major breakthroughs in ecological, evolutionary and functional genomics over the years (Braasch et al. 2016; Capel 2017; Xie et al. 2019). As such, they have become a high priority taxon for several large-scale projects aiming to extend phylogenetic coverage of vertebrate genome resources (Fan et al. 2020; Rhie et al. 2020). However, functional annotations of these genomes lag behind those of intensely studied organisms such as human and mouse, where diverse datasets including tissue-specific transcriptomes, epigenomic information and gene knock-out data are available to dissect how the genome is deployed across the whole organism and how genes contribute to phenotypes. Because these datasets remain costly and time-consuming to generate, it is not expected that most non-model teleost genomes will benefit from deep functional characterization in the near future. Being able to transfer functional annotations from one species to another is therefore all the more crucial to exploit their shared genomic resources. While this solution is imperfect, there is extensive evidence that orthologous genes and regions often maintain conserved functions and characteristics across long evolutionary scales, so that comparative approaches can supply a rich source of functional annotations where none exist (Altenhoff et al. 2012; Irimia et al. 2013; Polychronopoulos et al. 2017). Transferring functional annotations is however particularly difficult in fish, where many regions are duplicated remnants of the teleost-specific whole genome duplication and orthologous copies are difficult to identify across species. We expect that the genome-scale, clade-wide paralogy and orthology resources we provide here will allow

functional annotations to be extrapolated across species and propel studies of functional characterization of teleost genomes.

360 It is important to note that the teleost comparative atlas comes with a number of limitations that directly stem from the way it was built. First, the assignations to ancestral chromosomes before the TGD are only as good as the state-of-the-art ancestral genome reconstruction. There is a general consensus that the ancestral teleost genome contained 13 chromosomes (Kasahara et al. 2007; Nakatani and McLysaght 2017) – however, the precise delineation of genes 365 descending from each of those 13 groups may be subject to modifications as the field evolves, which may lead to updates in the ancestral assignations of the regions in the comparative atlas. Second, while we show that the comparative atlas is resilient to species-specific errors in ancestral chromosome or gene orthology groups assignations due to the redundant information provided by multiple species, the atlas is ultimately based on gene family tree 370 models, which are sometimes inaccurate or incomplete (Hahn 2007; Som 2015). Inaccurate trees may result in local misspecifications of gene paralogs or ancestral assignations in the comparative atlas. We have previously shown that SCORPiOs significantly improves gene tree accuracy after a WGD event (Parey et al. 2020), and we found only few discrepancies between megabase-scale regions predicted to descend from the same ancestral chromosome from 375 (Nakatani and McLysaght 2017) and paralogy relationships predicted at gene-to-gene resolution by our gene trees. This suggests that the atlas is generally accurate. However, we note that SCORPiOs flagged 2,832 gene trees where sequence identity relationships are inconsistent with their local syntenic context, which may represent areas where the comparative atlas is either less reliable or biologically less informative.

380 We leveraged the comparative atlas to investigate how the post-duplication rediploidization process has shaped extant teleost genomes. One long-standing question with respect to the teleost duplication is whether the tetraploid teleost ancestor arose via allopolyploidization or autopolyploidization (Martin and Holland 2014; Conant 2020). Previous studies have highlighted the contrasting genomic implications of allopolyploidization and 385 autopolyploidization, where the two subgenomes of allopolyploids are often – although not systematically – subjected to asymmetrical evolution, while in autopolyploids all chromosome copies are highly similar and equally affected by the rediploidization process (Garsmeur et al. 2014; Cheng et al. 2018). We find evidence for prolonged recombination between whole homeologous chromosomes after the TGD, thus strongly suggesting an autopolyploid origin. 390 As such, the annotations of ‘a’ and ‘b’ chromosomal copies in the comparative atlas should not be interpreted as two distinct subgenomes where all ‘a’ chromosomes (or ‘b’) descend from a single parental genome. We also note that the comparative atlas is impacted by this delayed meiosis resolution: ‘a’ and ‘b’ assignments are not informative for genomic regions that maintained meiotic recombination until after species divergence. Indeed, in this evolutionary

395 scenario, gene sequences only start diverging once recombination is suppressed, which implies that there are no strict 1:1 orthology relationships between duplicated genes in these species.

Our results give broader support to a previous study on the rediploidization of the *Hox* genes clusters in teleosts (Martin and Holland 2014), since we demonstrate that delayed meiosis 400 resolution after the Osteoglossiform-Clupeocephala lineage split did not only affect these specific *Hox* clusters segments but do extend to whole ancestral chromosomes. As a plausible alternative to autopolyploidization, we cannot rule out that the TGD is the result of a segmental allopolyploidization (Stebbins 1947) following hybridization between parental genomes where most homeologs have immediate bivalent pairing but a few display tetravalent meiosis 405 behaviour, like in allotetraploid loaches (Li et al. 2011). Further genome-wide studies including even more basally branching teleosts, especially elopomorph genomes (Hughes et al. 2018), will be crucial to estimate the proportion of the ancestral tetraploid teleost genome subjected to delayed meiosis resolution. To date, genes sequences relationships between Elopomorpha and Clupeocephala have only been evaluated for *Hox* clusters and confirm prolonged 410 recombination between these duplicated genomic segments (Henkel et al. 2012). In addition, we find a significant bias in genes retention for a subset of duplicated chromosomes pairs, as also previously observed at a more local scale by (Conant 2020). We lack an explanatory mechanism for these differences, which have been classically linked to allopolyploid genome rediploidization, where one parental subgenome is more expressed, under stronger selective 415 pressure and retains more genes. However, we find no correlation linking together differences in genes retention, expression level and selective pressure; and 'autopolyploid-like' homeologs do not all correspond to homeologs unaffected by the retention bias. The rediploidization process is well-documented in plants, but it remains poorly understood in vertebrates and its long-term consequences are unclear. While the well-studied examples of salmonids and carps 420 genome duplications seem to recapitulate plant-based models of autopolyploid and allopolyploid rediploidization respectively (Robertson et al. 2017; Xu et al. 2019; Gundappa et al. 2021; Li et al. 2021), a more complex rediploidization history may have shaped teleost genomes. As a striking counter-example to the classical allopolyploid rediploidization model, asexual allotetraploid loaches present a bias in genes retention that is surprisingly in the 425 opposite direction to the expression bias, i.e. the most expressed subgenome accumulated more gene losses (Janko et al. 2021). Taken together, these results suggest that underappreciated factors may govern genes retention following genome duplication in vertebrates.

Methods

430

Libraries and packages

Scripts to build the comparative atlas were written in Python 3.6.8 and assembled together in a pipeline with Snakemake version 5.13.0 (Köster and Rahmann 2012). The ete3 package (version 3.1.1) was used for phylogenetic gene tree manipulation and drawing. Other python
435 package dependencies used for plots and analyses include matplotlib (version 3.1.1), seaborn (version 0.9.0), numpy (version 1.18.4), pandas (version 0.24.2), pingouin (version 0.3.4) for Wilcoxon paired tests and multiple testing correction, and scipy (version 1.4.1) for Fisher's exact tests. Chromosomes painting and synteny comparisons were drawn with the RIdeogram R package (Hao et al. 2020).

440

Genomic resources

Genome assemblies and annotations for the 101 vertebrate genomes were downloaded from various sources, including the NCBI, Ensembl version 95 and GigaDB. The source and assembly version used for each genome are listed in Supplementary Table S1. The gene
445 coordinates files for the 74 teleost genomes in the comparative atlas have been deposited to Zenodo (https://doi.org/10.5281/zenodo.5776772).

Phylogenetic gene trees

Initial gene trees were built using the Ensembl Compara pipeline (Vilella et al. 2009). Briefly,
450 starting from the sets of proteins derived from the longest transcripts in each genome, we performed an all-against-all BLASTP+ (Altschul et al. 1990), with the following parameters '-seg no -max_hsps 1 -use_sw_tback -evalue 1e-5'. We then performed clustering with hcluster_sg to define gene families, using parameters '-m 750 -w 0 -s 0.34 -O'. We built multiple alignments using M-Coffee (Wallace et al. 2006), with the command 't-coffee -type=PROTEIN
455 -method mafftgins_msa, muscle_msa,kalign_msa,t_coffee_msa -mode=mcoffee'. Next, we conducted phylogenetic trees construction and reconciliation with TreeBeST, using default parameters (Vilella et al. 2009). Because TreeBeST is systematically biased towards inferring gene duplication nodes that are overly old (Hahn 2007), we pre-processed gene trees to edit nodes with a very low duplication confidence score (score < 0.1), using ProfileNJ (Noutahi et
460 al. 2016). Finally, we ran SCORPiOs (version v1.1.0) to account for several whole genome duplication events in the species phylogeny and correct gene trees accordingly: the teleost 3R WGD, using bowfin and gar as outgroups; the carps 4R WGD, using zebrafish as outgroup; and the salmonids 4R WGD, using Northern pike as outgroup. SCORPiOs reached convergence after 5 iterative rounds of correction.

465

Ancestral gene statistics

We calculated the predicted number of genes in the post-duplication ancestral teleost genome using our set of 26,692 gene trees, and compared to 60,447 state-of-the art gene trees stored in Ensembl Compara v100. This ancestral gene number is an indirect but accurate approximation of the quality of inferred gene trees, since the major challenge is to accurately position duplications at this ancestral node. TreeBeST phylogeny-reconciled gene trees were recursively browsed and gene copies inferred in the teleost ancestor (*Osteoglossocephalai*) were collected.

Comparative genomic atlas

The FishComparativeAtlas pipeline annotates teleost genes to post-TGD duplicated chromosomes. It takes as input genomic regions annotated to pre-duplication chromosomes for 4 reference species (zebrafish, medaka, stickleback and tetraodon), and the set of gene trees described above. Segmentation of the four teleost species with respect to the 13 ancestral chromosomes were extracted from Nakatani and McLysaght (2017) and genomic interval coordinates converted to lists of genes. All genomes were then reduced to ordered lists of genes. We first converted input genomic intervals from zebrafish genome assembly Zv9 to GRCz11 and from medaka genome assembly MEDAKA1 to ASM223467v1, by transferring boundary genes between assemblies using Ensembl gene ID histories. Next, we identified putative WGD sister regions within each of the 4 reference species, as regions sharing a high fraction of TGD-descended paralogs (Supplementary Figure S2B). We grouped regions descended from the same pre-duplication ancestral chromosome, and iteratively annotated pairs of regions into internally consistent sets of `a` and `b` post-duplication sister regions as follows: for each ancestral chromosome, we started from the largest descendant region and arbitrarily defined it as `a`. All regions sharing 50% ohnologs or more with this `a` region are identified its `b` paralog(s). Additional search rounds were then conducted to extend the ‘a’ and ‘b’ annotations in a stepwise manner to all regions descended from this ancestral chromosome. The required ohnolog fraction was decreased at each round, and the search was stopped when remaining blocks shared less than 5% ohnologs with previously annotated regions. Since this identification of duplicated regions was performed independently in each of the 4 species, `a` and `b` regions annotations were homogenized to ensure consistency across species (Supplementary Figure S2C). The homogenization step uses orthology relationships from the gene trees and stickleback as an arbitrary guide species: annotations were switched in other species when `a` segments shared more orthologous genes with the `b` region of stickleback. Where conflicts remained for individual genes annotations, we used a majority vote procedure across all four species to define the post-duplication chromosome. Finally, annotations were propagated to all teleost genomes in the gene trees using orthologies.

Simulation of ancestral chromosome boundary shifts

To simulate incertitude in interval boundaries in the original ancestral genome reconstruction from Nakatani and McLysaght (2017), we randomly drew new boundaries in the vicinity of their original location according to a Gaussian distribution with standard deviation σ varying in [5, 10, 25, 50, 75, 100] genes. These boundary shifts were independently generated for each of the 4 reference species, with n=100 random noise simulations for each σ value and each reference species. In total, simulations generated 600 noisy input datasets that were processed with the FishComparativeAtlas pipeline in order to assess its robustness to noise. Each of the 600 produced outputs were then compared to the comparative atlas, by counting the proportion of gene families with a reassigned ancestral chromosome (Supplementary Figure S3).

Early and late rediploidization gene tree topologies

Phylogenetic gene trees were built for the reduced set of 33 genomes as follows. We first filtered the CDS multiple alignments previously computed for the 74 teleosts and 27 non-teleost outgroup genomes set as described above, to retain only genes of all 27 outgroups and 6 teleost genomes, including 3 Osteoglossiforms: paramormyrops (Paramormyrops kingsleyae), arapaima (*Arapaima gigas*) and arowana (*Scleropages formosus*) and 3 Clupeocephala: zebrafish *(Danio rerio),* medaka (*Oryzias latipes)* and stickleback (*Gasterosteus aculeatus)* (Supplementary Figure S6). We used these reduced multiple alignments to build phylogenetic gene trees with TreeBeST (Vilella et al. 2009), using default parameters and the option "-X 10". This resulted in a set of 14,391 gene trees containing genes of the 33 retained genomes. We then ran SCORPiOs to correct trees for the teleost-specific duplication using the gar and bowfin genomes as outgroups.

To investigate the occurrence of delayed rediploidization, we implemented an extension to the SCORPiOs pipeline (SCORPiOs "LORelEi" for "Lineage-specific Ohnolog Resolution Extension) to: (i) identify gene trees characterized by sequence-synteny prediction conflicts and (ii) perform likelihood AU-tests (Shimodaira and Hasegawa 2001) to evaluate how early and late rediploidization tree topologies are supported by gene sequence evolution. For (i), we identify gene trees that SCORPiOs attempts to correct based on syntenic information, but whose correction is rejected due to low sequence-based likelihood. For (ii), we selected the 5,557 gene families containing a gene in at least one reference outgroup (bowfin or gar) and resulting in distinct tree topologies under AORe and LORe. In practice, these topologies can be distinguished when at least one teleost group (Osteoglossiform or Clupeocephala) retained both ohnologs, although not necessarily in the same species. For each of these 5,557 families, we built 3 gene trees using RaxML 8.2.12 (Stamatakis 2014), with the GTRGAMMA model:

540 the unconstrained maximum likelihood (ML) tree, the constrained early rediploidization topology and the constrained late rediploidization topology (Figure 3B). We then used CONSEL (Shimodaira and Hasegawa 2001) to test for significant differences in log-likelihood reported by RAxML (Stamatakis 2014). A tree topology was rejected when it is significantly less likely than the ML tree at α=0.05.

545

Genes retention on homeologous chromosomes

Because ancestral gene order is particularly difficult to reconstruct in teleosts due to an elevated rate of microsyntenic rearrangements and gene copy losses (Inoue et al. 2015; Nakatani and McLysaght 2017), we take advantage of a non-duplicated outgroup fish genome
550 as a proxy for the ancestral gene order. Here, we make the assumption that consecutive genes on the outgroup genome, all assigned to the same pre-duplication chromosome, represent the ancestral gene order. Using the gene trees, we identify 10,629 1-to-1 orthologies between spotted gar genes and teleost pre-duplication gene families, and 11,599 with bowfin genes. We then used the gene order in the outgroup genomes as an approximation of the ancestral
555 teleost gene order. We reduced outgroup genomes to these genes, and extracted all blocks of consecutive genes annotated to the same pre-duplication chromosome. We computed the percentage of gene copies retained on ‘a’ and ‘b’ homeologs in each extant duplicated genome, using non-overlapping windows of 10 genes along these blocks.

560 **dN/dS ratio for ohnologous gene copies**

We considered 1,269 teleost gene families annotated in the comparative atlas, that contain 2 ohnologous gene copies for at least 40 out of 64 teleost genomes (all studied genomes excluding the 10 teleost with an additional WGD)and no subsequent gene duplication. We pruned each gene tree topology to only retain ohnologous genes. We next divided each tree
565 to build two separate subtrees: one subtree containing the ‘a’ gene copies and the orthologs in bowfin and spotted gar, and a second subtree with the ‘b’ gene copies and the same outgroup genes. For each ‘a’ and ‘b’ subfamily, we used translatorX vLocal (Abascal et al. 2010) to (i) translate coding sequences, (ii) align resulting amino acid sequences with Mafft v7.310 (Katoh and Standley 2013) using option ‘--auto’, (iii) trim poorly aligned regions with
570 Gblocks version 0.91b (Castresana 2000) using parameters ‘-b4=2 -b5=h’ and (iv) back-translate the sequences into codon alignments. We estimated the dN/dS ratio by maximum likelihood with CodeML (Yang 2007) using model M0 as configured in the ete3-evol wrapper (Huerta-Cepas et al. 2016), which estimates a single unconstrained dN/dS ratio for each subtree. Finally, we used Wilcoxon paired tests to identify significant differences in dN/dS
575 between ‘a’ and ‘b’ genes from each duplicated chromosome pair, corrected for multiple testing using the Benjamini-Hochberg procedure.

Expression level of ohnologous genes

We used RNA-seq datasets across 11 tissues (bones, brain, embryo, gills, heart, intestine, kidney, liver, muscle, ovary, and testis) in zebrafish and medaka from the PhyloFish database (Pasquier et al. 2016), normalized into TPM (transcripts per million transcripts) and quantile normalized across tissues as previously described (Parey et al. 2020). Gene IDs were then converted from Ensembl version 89 to version 95 using conversion tables downloaded from BioMart (Smedley et al. 2009). Average expression across tissue (Figure 4C, Supplementary Figure S9) and by-tissue expression (Supplementary Figure S10-S11) were calculated for ohnologs grouped by their ancestral chromosome of origin.

ZFIN gene names

Zebrafish ZFIN gene names were extracted using BioMart (Smedley et al. 2009) from the Ensembl database (version 95). We extracted the last letter of gene names, which represents 'a' and 'b' ohnology annotations in ZFIN. We then computed the minimal number of 'a' and 'b' ZFIN gene name reassignments that would be necessary to be consistent with the comparative atlas. In the comparative atlas, 'a' and 'b' labels are arbitrarily assigned to duplicated chromosomes, i.e. genes descended from chromosomes 1a and 1b could be swapped to 1b and 1a. In order to not artificially overestimate discordances, we first swapped such arbitrary 'a' and 'b' annotations to minimize differences with ZFIN. Finally, we counted the remaining number of 'a' and 'b' disagreement for zebrafish genes in the comparative atlas that were also annotated in ZFIN.

Data availability and implementation

Gene homology relationships and local synteny conservation between teleosts can be interactively browsed through the Genomicus webserver (Nguyen et al. 2022), accessible at https://www.genomicus.bio.ens.psl.eu/genomicus-fish-03.01/cgi-bin/search.pl. The homology relationships from the comparative atlas can be downloaded as flat or HTML files via an ftp server (ftp://ftp.biologie.ens.fr/pub/dyogen/genomicus-fish/03.01/ParalogyMap), along with the gene trees in New Hampshire eXtended (.nhx) format (ftp://ftp.biologie.ens.fr/pub/dyogen/genomicus-fish/03.01/protein_trees.nhx),

All code for the FishComparativeAtlas pipeline is publicly available on GitHub (https://github.com/DyogenIBENS/FishComparativeAtlas). An archive containing a stable version of the code along with all input data (including gene trees) and the final atlas has been deposited to Zenodo (https://doi.org/10.5281/zenodo.5776772), to reproduce the generation of the comparative atlas or directly inspect it. The SCORPiOs LORelEi extension is available on GitHub (https://github.com/DyogenIBENS/SCORPIOS) and has also been deposited to

Zenodo (https://doi.org/10.5281/zenodo.3727519), along with all input data, environments and outputs.

Acknowledgements

We thank Pierre Vincens for the coordination of computing resources and all members of the GenoFish consortium for fruitful discussions.

Funding

This work is funded by ANR GenoFish (grant number ANR-16-CE12-0035-02), and was supported by grants from the French Government and implemented by ANR [ANR–10–LABX–54 MEMOLIFE and ANR–10–IDEX–0001–02 PSL* Research University]. This project received funds from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 817923 (AQUA-FAANG).

References

Abascal F, Zardoya R, Telford MJ. 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* 38:W7–W13.

Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C. 2012. Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Comput. Biol.* 8:e1002514.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–410.

Aparicio S, Chapman J, Stupka E, Putnam N, Chia J-M, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, et al. 2002. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* 297:1301–1310.

Bird KA, Niederhuth CE, Ou S, Gehan M, Pires JC, Xiong Z, VanBuren R, Edger PP. 2021. Replaying the evolutionary tape to investigate subgenome dominance in allopolyploid Brassica napus. *New Phytol.* 230:354–371.

Braasch I, Gehrke AR, Smith JJ, Kawasaki K, Manousaki T, Pasquier J, Amores A, Desvignes T, Batzel P, Catchen J, et al. 2016. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nat. Genet.* 48:427–437.

Capel B. 2017. Vertebrate sex determination: evolutionary plasticity of a fundamental switch. *Nat. Rev. Genet.* 18:675–689.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17:540–552.

Conant GC. 2020. The lasting after-effects of an ancient polyploidy on the genomes of teleosts. *PLOS ONE* 15:e0231356.

Cheng F, Wu J, Cai X, Liang J, Freeling M, Wang X. 2018. Gene retention, fractionation and subgenome differences in polyploid plants. *Nat. Plants* 4:258–268.

Escobar-Camacho D, Carleton KL, Narain DW, Pierotti MER. 2020. Visual pigment evolution in Characiformes: the dynamic interplay of teleost whole-genome duplication, surviving opsins and spectral tuning. *Mol. Ecol.* 29:2234–2253.

Fan G, Song Y, Yang L, Huang X, Zhang S, Zhang M, Yang X, Chang Y, Zhang H, Li Y, et al. 2020. Initial data release and announcement of the 10,000 Fish Genomes Project (Fish10K). *GigaScience* [Internet] 9. Available from: https://academic.oup.com/gigascience/article/9/8/giaa080/5893976

Freeling M, Woodhouse MR, Subramaniam S, Turco G, Lisch D, Schnable JC. 2012. Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. *Curr. Opin. Plant Biol.* 15:131–139.

Garsmeur O, Schnable JC, Almeida A, Jourda C, D'Hont A, Freeling M. 2014. Two Evolutionarily Distinct Classes of Paleopolyploidy. *Mol. Biol. Evol.* 31:448–454.

Greenway R, Barts N, Henpita C, Brown AP, Rodriguez LA, Peña CMR, Arndt S, Lau GY, Murphy MP, Wu L, et al. 2020. Convergent evolution of conserved mitochondrial pathways underlies repeated adaptation to extreme environments. *Proc. Natl. Acad. Sci.* 117:16424–16430.

Gundappa MK, To T-H, Grønvold L, Martin SAM, Lien S, Geist J, Hazlerigg D, Sandve SR, Macqueen DJ. 2021. Genome-wide reconstruction of rediploidization following autopolyploidization across one hundred million years of salmonid evolution. *Mol. Biol. Evol.*:msab310.

Haffter P, Granato M, Brand M, Mullins MC, Hammerschmidt M, Kane DA, Odenthal J, Kelsh RN, Furutani-Seiki M, Vogelsang E, et al. 1996. The identification of genes with unique and essential functions in the development of the zebrafish, Danio rerio. Development 123:1–36.

Hahn MW. 2007. Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biol.* 8:R141.

Hao Z, Lv D, Ge Y, Shi J, Weijers D, Yu G, Chen J. 2020. RIdeogram: drawing SVG graphics to visualize and map genome-wide data on the idiograms. *PeerJ Comput. Sci.* 6:e251.

Henkel CV, Burgerhout E, Wijze DL de, Dirks RP, Minegishi Y, Jansen HJ, Spaink HP, Dufour S, Weltzien F-A, Tsukamoto K, et al. 2012. Primitive Duplicate Hox Clusters in the European Eel's Genome. *PLOS ONE* 7:e32231.

Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, Vilella AJ, Searle SMJ, Amode R, Brent S, et al. 2016. Ensembl comparative genomics resources. *Database J. Biol. Databases Curation* 2016.

Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, Collins JE, Humphray S, McLaren K, Matthews L, et al. 2013. The zebrafish reference genome sequence and its relationship to the human genome. *Nature* 496:498–503.

Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* 33:1635–1638.

Hughes LC, Ortí G, Huang Y, Sun Y, Baldwin CC, Thompson AW, Arcila D, Betancur-R R, Li C, Becker L, et al. 2018. Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. Proceedings of the National Academy of Sciences 115:6249–6254.

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.* 112:14918–14923.

Irimia M, Maeso I, Roy SW, Fraser HB. 2013. Ancient cis-regulatory constraints and the evolution of genome architecture. *Trends Genet.* 29:521–528.

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* 431:946–957.

Janko K, Bartoš O, Kočí J, Roslein J, Drdová EJ, Kotusz J, Eisner J, Mokrejš M, Štefková-Kašparová E. 2021. Genome Fractionation and Loss of Heterozygosity in Hybrids and Polyploids: Mechanisms, Consequences for Selection, and Link to Gene Function. *Mol. Biol. Evol.* [Internet]. Available from: https://doi.org/10.1093/molbev/msab249

Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, Yamada T, Nagayasu Y, Doi K, Kasai Y, et al. 2007. The medaka draft genome and insights into vertebrate genome evolution. *Nature* 447:714–719.

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* 30:772–780.

Köster J, Rahmann S. 2012. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* 28:2520–2522.

Kim B-M, Amores A, Kang S, Ahn D-H, Kim J-H, Kim I-C, Lee JH, Lee SG, Lee H, Lee J, et al. 2019. Antarctic blackfin icefish genome reveals adaptations to extreme environments. *Nat. Ecol. Evol.* 3:469–478.

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol Biol Evol 34:1812–1819.

Li J-T, Wang Q, Huang Yang M-D, Li Q-S, Cui M-S, Dong Z-J, Wang H-W, Yu J-H, Zhao Y-J, Yang C-R, et al. 2021. Parallel subgenome structure and divergent expression evolution of allo-tetraploid common carp and goldfish. *Nat. Genet.* 53:1493–1503.

Li Y-J, Yu Z, Zhang M-Z, Qian C, Abe S, Arai K. 2011. The origin of natural tetraploid loach Misgurnus anguillicaudatus (Teleostei: Cobitidae) inferred from meiotic chromosome configurations. *Genetica* 139:805.

Lieschke GJ, Currie PD. 2007. Animal models of human disease: zebrafish swim into view. *Nat. Rev. Genet.* 8:353–367.

Lloyd A, Blary A, Charif D, Charpentier C, Tran J, Balzergue S, Delannoy E, Rigaill G, Jenczewski E. 2018. Homoeologous exchanges cause extensive dosage-dependent gene expression changes in an allopolyploid crop. *New Phytol.* 217:367–377.

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.* 31:2592–2611.

Mason AS, Wendel JF. 2020. Homoeologous Exchanges, Segmental Allopolyploidy, and Polyploid Genome Evolution. *Front. Genet.* [Internet] 11. Available from: https://www.frontiersin.org/articles/10.3389/fgene.2020.01014/full

Moriyama Y, Ito F, Takeda H, Yano T, Okabe M, Kuraku S, Keeley FW, Koshiba-Takeuchi K. 2016. Evolution of the fish heart by sub/neofunctionalization of an elastin gene. *Nat. Commun.*7:1–10.

Nakatani Y, McLysaght A. 2017. Genomes as documents of evolutionary history: a probabilistic macrosynteny model for the reconstruction of ancestral genomes. *Bioinformatics* 33:i369–i378.

Nguyen NTT, Vincens P, Dufayard JF, Roest Crollius H, Louis A. 2022. Genomicus in 2022: comparative tools for thousands of genomes and reconstructed ancestors. Nucleic Acids Res. 50:D1025–D1031.

Noutahi E, Semeria M, Lafond M, Seguin J, Boussau B, Guéguen L, El-Mabrouk N, Tannier E. 2016. Efficient Gene Tree Correction Guided by Genome Evolution. *PLOS ONE* 11:e0159559.

Ohno S, Wolf U, Atkin NB. 1968. Evolution from fish to mammals by gene duplication. *Hereditas* 59:169–187.

Parey E, Louis A, Cabau C, Guiguen Y, Roest Crollius H, Berthelot C. 2020. Synteny-Guided Resolution of Gene Trees Clarifies the Functional Impact of Whole-Genome Duplications. *Mol. Biol. Evol.* 37:3324–3337.

Pasquier J, Cabau C, Nguyen T, Jouanno E, Severac D, Braasch I, Journot L, Pontarotti P, Klopp C, Postlethwait JH, et al. 2016. Gene evolution and gene expression after whole genome duplication in fish: the PhyloFish database. *BMC Genomics* 17:368.

Polychronopoulos D, King JWD, Nash AJ, Tan G, Lenhard B. 2017. Conserved non-coding elements: developmental gene regulation meets genome organization. *Nucleic Acids Res.* 45:12611–12624.

Postlethwait JH, Woods IG, Ngo-Hazelett P, Yan YL, Kelly PD, Chu F, Huang H, Hill-Force A, Talbot WS. 2000. Zebrafish comparative genomics and the origins of vertebrate chromosomes. *Genome Res.* 10:1890–1902.

Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Kim J, et al. 2021. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* 592:737–746.

Rittschof CC, Bukhari SA, Sloofman LG, Troy JM, Caetano-Anollés D, Cash-Ahmed A, Kent M, Lu X, Sanogo YO, Weisner PA, et al. 2014. Neuromolecular responses to social challenge: Common mechanisms across mouse, stickleback fish, and honey bee. *Proc. Natl. Acad. Sci. U. S. A.* 111:17929–17934.

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.* 18:111.

Ruzicka L, Howe DG, Ramachandran S, Toro S, Van Slyke CE, Bradford YM, Eagle A, Fashena D, Frazer K, Kalita P, et al. 2019. The Zebrafish Information Network: new support for non-coding genes, richer Gene Ontology annotations and the Alliance of Genome Resources. *Nucleic Acids Res.* 47:D867–D873.

Salzburger W. 2018. Understanding explosive diversification through cichlid fish genomics. *Nat. Rev. Genet.* 19:705–717.

Schartl M. 2014. Beyond the zebrafish: diverse fish species for modeling human disease. *Dis. Model. Mech.* 7:181–192.

Session AM, Uno Y, Kwon T, Chapman JA, Toyoda A, Takahashi S, Fukui A, Hikosaka A, Suzuki A, Kondo M, et al. 2016. Genome evolution in the allotetraploid frog Xenopus laevis. *Nature* 538:336–343.

Shimodaira H, Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinforma. Oxf. Engl.* 17:1246–1247.

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.* 4:820–830

Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A. 2009. BioMart – biological queries made easy. BMC Genomics 10:22.

Som A. 2015. Causes, consequences and solutions of phylogenetic incongruence. *Brief. Bioinform.* 16:536–548.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinforma. Oxf. Engl.* 30:1312–1313.

Stebbins GL. 1947. Types of Polyploids: Their Classification and Significance. In: Demerec M, editor. Advances in Genetics. Vol. 1. Academic Press. p. 403–429.

Streisinger G, Walker C, Dower N, Knauber D, Singer F. 1981. Production of clones of homozygous diploid zebra fish (Brachydanio rerio). Nature 291:293–296.

Sun H, Wu S, Zhang G, Jiao C, Guo S, Ren Y, Zhang J, Zhang H, Gong G, Jia Z, et al. 2017. Karyotype Stability and Unbiased Fractionation in the Paleo-Allotetraploid Cucurbita Genomes. *Mol. Plant* 10:1293–1306.

Taylor JS, Braasch I, Frickey T, Meyer A, Peer YV de. 2003. Genome Duplication, a Trait Shared by 22,000 Species of Ray-Finned Fish. *Genome Res.* 13:382–390.

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* 19:327–335.

Wallace IM, O'Sullivan O, Higgins DG, Notredame C. 2006. M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res.* 34:1692–1699.

Wittbrodt J, Shima A, Schartl M. 2002. Medaka — a model organism from the far east. *Nat. Rev. Genet.* 3:53–64.

Xie KT, Wang G, Thompson AC, Wucherpfennig JI, Reimchen TE, MacColl ADC, Schluter D, Bell MA, Vasquez KM, Kingsley DM. 2019. DNA fragility in the parallel evolution of pelvic reduction in stickleback fish. *Science* 363:81–84.

Xu P, Xu J, Liu G, Chen L, Zhou Z, Peng W, Jiang Y, Zhao Z, Jia Z, Sun Y, et al. 2019. The allotetraploid origin and asymmetrical genome evolution of the common carp Cyprinus carpio. *Nat. Commun.* 10:4625.

Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* 24:1586–1591.

Zakon HH, Lu Y, Zwickl DJ, Hillis DM. 2006. Sodium channel genes and the evolution of diversity in communication signals of electric fishes: Convergent molecular evolution. *Proc. Natl. Acad. Sci.* 103:3675–3680.

Zhao M, Zhang B, Lisch D, Ma J. 2017. Patterns and Consequences of Subgenome Differentiation Provide Insights into the Nature of Paleopolyploidy in Plants. *Plant Cell* 29:2974–2994.

Figures

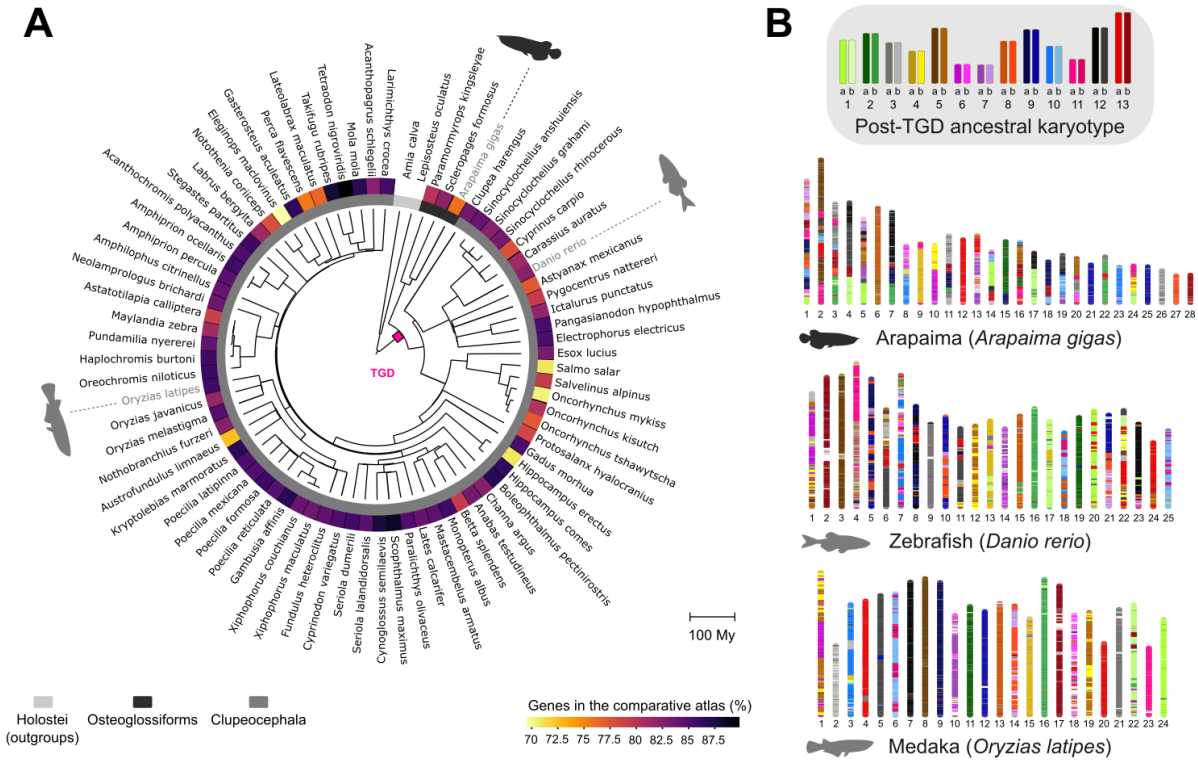

Figure 1: Comparative atlas of WGD-duplicated regions across 74 teleosts A. Phylogenetic tree of the 74 teleost genomes in the comparative atlas and 2 non-duplicated fish outgroups. Clades are annotated in the inner circle: Holostei (outgroups, grey), Osteoglossiforms (yellow) and Clupeocephala (orange). The proportion of genes from each species annotated in the comparative atlas is represented in the outer circle. Divergence times were extracted from TimeTree (Kumar et al. 2017). **B.** Karyotype paintings using the comparative atlas. At the top, we show the inferred ancestral karyotype after the teleost whole-genome duplication (TGD). Below, karyotypes of three teleost genomes are colored by their ancestral chromosome of origin according to the comparative atlas (1a, 1b, ..., 13a, 13b).

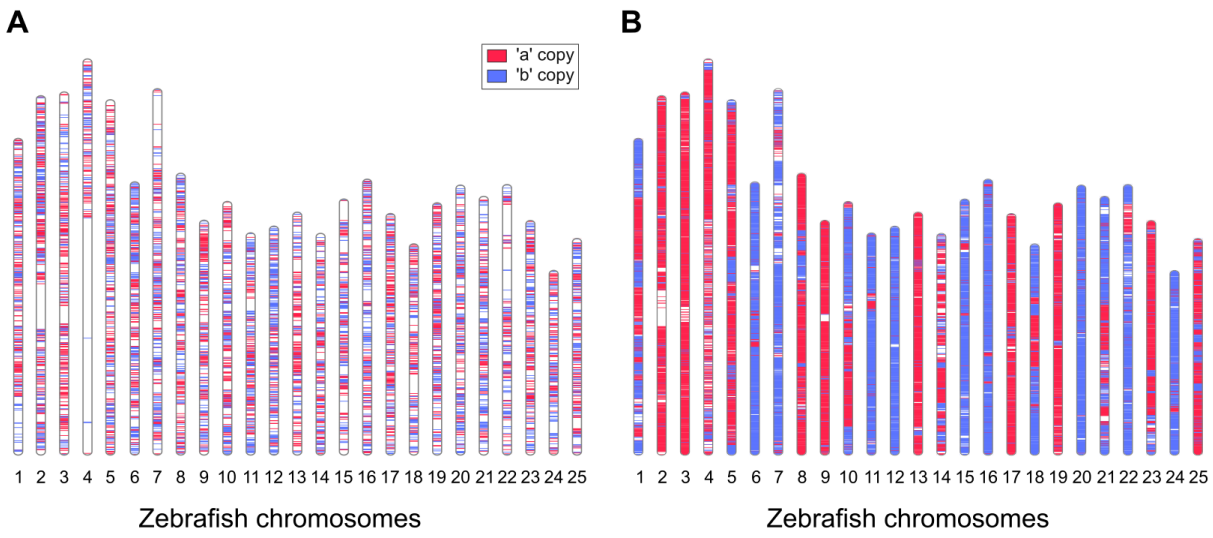

Figure 2: Zebrafish gene names are not evolutionary-consistent. A. Karyotypic localization of zebrafish 'a' and 'b' TGD paralogs, according to the ZFIN annotation. **B.** Full re-annotation of zebrafish 'a' and 'b' gene copies using the comparative atlas (84% of zebrafish genes annotated, including genes without a WGD paralog).

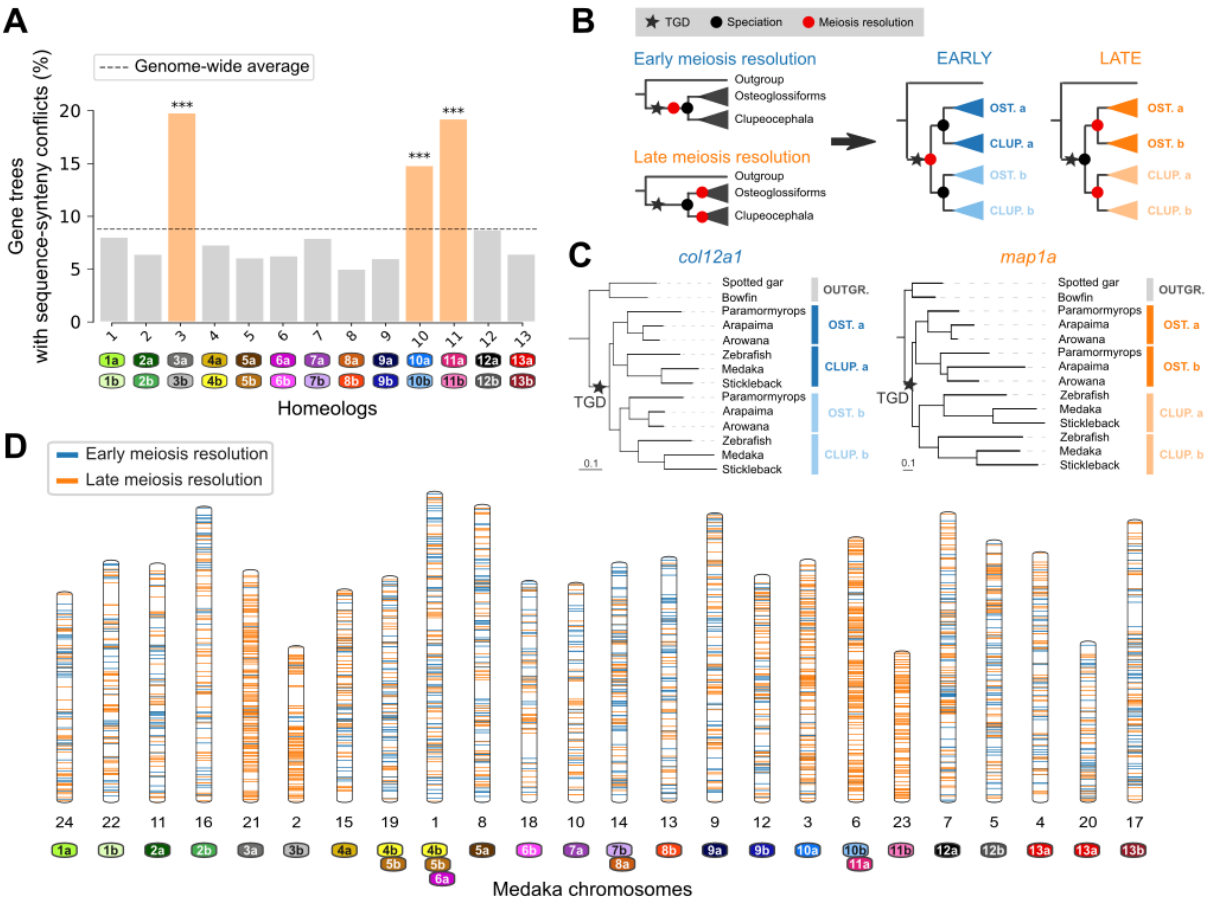

Figure 3: Delayed rediploidization following the TGD. A. Ancestral chromosomes 3, 10 and 11 are enriched in sequence-synteny conflicts (Methods, *** $p < 0.001$, hypergeometric tests with Benjamini-Hochberg correction for multiple testing). Color labels identify ancestrally duplicated chromosomes as in Figure 1A. Dotted line represents the genome-wide average. **B.** Gene tree topologies expected under early and late meiosis resolution hypotheses. The early rediploidization tree topology assumes that rediploidization was complete before the divergence of Osteoglossiform and Clupeocephala species, initiating 'a' and 'b' gene sequence divergence before speciation. The late rediploidization tree topology assumes that rediploidization was completed only after the divergence of Osteoglossiform and Clupeocephala species, delaying 'a' and 'b' duplicated sequences divergence to after speciation. **C.** Examples of early (*col12a1* family) and late (*map1a* family) rediploidized gene trees. For the *col12a1* family, the late rediploidization topology is inconsistent with gene sequence evolution ($p = 4e\text{-}09$, AU-test). For the *map1a* family, early rediploidization was similarly rejected (p-value = 0.001, AU-test). **D.** Early and late rediploidization gene families visualized on the medaka karyotype. Medaka chromosomes are annotated as numbers, while color labels represent ancestral chromosomes, as in (A). Homeologs 3, 10 and 11 almost entirely rediploidized later than the Osteoglossiform/Clupeocephala divergence.

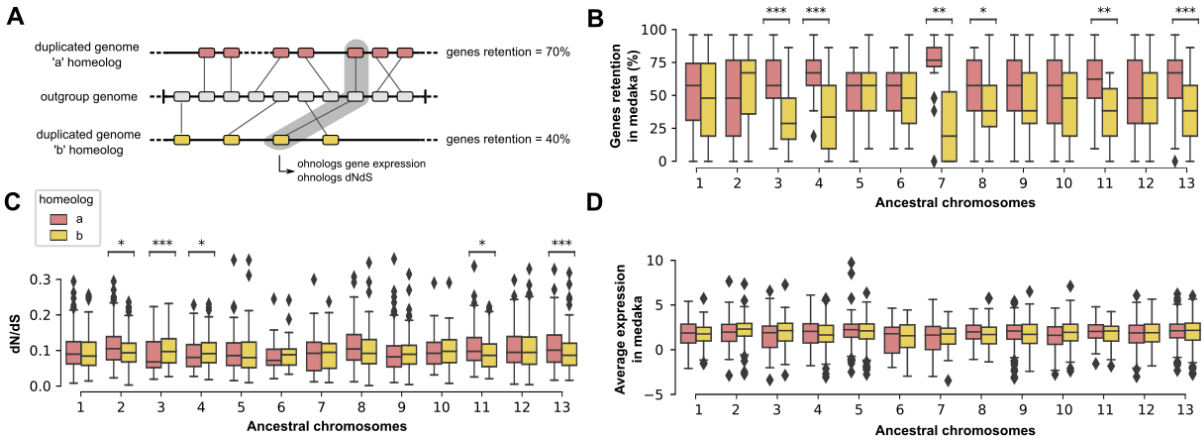

Figure 4: Differences in genes retention, selective pressure and gene expression between duplicated chromosomes. A. Schematic example of genes retention calculation. Using an outgroup genome as an approximation of the ancestral gene order, we assess genes retention on each duplicated chromosome in teleost genomes, by 10-gene bins, regardless of their genomic location (Methods). **B.** Gene retention on anciently duplicated chromosome copies in medaka, using the spotted gar genome as a proxy for ancestral gene order. Ancestral chromosomes with a significant bias in gene retention on one of the two copies are highlighted (*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, Wilcoxon paired tests with Benjamini-Hochberg correction for multiple testing). **C.** Differences in selective pressure (dN/dS ratio) between homeologs (Methods; p-values and tests as in (B)). **D.** Average expression across tissues in medaka. No significant differences in expression were detected between genes of duplicated chromosome copies (Wilcoxon paired tests with Benjamini-Hochberg correction for multiple testing, at $\alpha=0.05$).