Synteny-Guided Resolution of Gene Trees Clarifies the Functional Impact of Whole-Genome Duplications

Elise Parey, Alexandra Louis, Cédric Cabau, Yann Guiguen, Hugues Roest Crollius, Camille Berthelot

Synteny-Guided Resolution of Gene Trees Clarifies the Functional Impact of Whole-Genome Duplications

Elise Parey [1], Alexandra Louis,[1] Cédric Cabau,[2] Yann Guiguen [3], Hugues Roest Crollius [*,1] and Camille Berthelot [*,1]

[1]Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Ecole Normale Supérieure, CNRS, INSERM, Université PSL, Paris, France
[2]SIGENAE, GenPhySE, Université de Toulouse, INRAE, ENVT, Castanet Tolosan, France
[3]INRAE, LPGP, Rennes, France

***Corresponding authors:** E-mails: camille.berthelot@bio.ens.psl.eu; hrc@bio.ens.psl.eu.
Associate editor: Koichiro Tamura

Abstract

Whole-genome duplications (WGDs) have major impacts on the evolution of species, as they produce new gene copies contributing substantially to adaptation, isolation, phenotypic robustness, and evolvability. They result in large, complex gene families with recurrent gene losses in descendant species that sequence-based phylogenetic methods fail to reconstruct accurately. As a result, orthologs and paralogs are difficult to identify reliably in WGD-descended species, which hinders the exploration of functional consequences of WGDs. Here, we present Synteny-guided CORrection of Paralogies and Orthologies (SCORPiOs), a novel method to reconstruct gene phylogenies in the context of a known WGD event. WGDs generate large duplicated syntenic regions, which SCORPiOs systematically leverages as a complement to sequence evolution to infer the evolutionary history of genes. We applied SCORPiOs to the 320-My-old WGD at the origin of teleost fish. We find that almost one in four teleost gene phylogenies in the Ensembl database (3,394) are inconsistent with their syntenic contexts. For 70% of these gene families (2,387), we were able to propose an improved phylogenetic tree consistent with both the molecular substitution distances and the local syntenic information. We show that these synteny-guided phylogenies are more congruent with the species tree, with sequence evolution and with expected expression conservation patterns than those produced by state-of-the-art methods. Finally, we show that synteny-guided gene trees emphasize contributions of WGD paralogs to evolutionary innovations in the teleost clade.

Key words: **whole-genome duplications, synteny, gene phylogeny, molecular evolution.**

Article

Introduction

Whole-genome duplications (WGDs) are dramatic evolutionary events that result in the doubling of a species entire genome. Several ancient WGDs have occurred in land plants, fungi, and animals, in which they represent a major source of functional innovation with long-term impact on the evolution of species (Jaillon et al. 2004; Van de Peer et al. 2017). Genome doubling events have also been uncovered in non-model organisms in recent years (Kenny et al. 2016; Sollars et al. 2017), with more still to be discovered. Although many gene duplicates produced by WGDs are eventually lost, some are retained and thought to provide raw material for evolution to repurpose into new functions (Ohno 1970; Lynch and Conery 2000). For example, 35% of human genes still have ancient duplicates from two WGD events ~550 Ma, which diversified essential multigene families including the MHC (immunity) or the four HOX clusters (anteroposterior development) (Sacerdot et al. 2018; Singh and Isambert 2020). Investigating the fates of duplicate genes in descendant species is crucial to understand how WGDs contribute to phenotypic robustness, adaptation, and evolvability. This however requires the accurate identification of WGD orthologs, that is, genes descended by speciation after the duplication event, from WGD paralogs, that is, genes descended from different duplicates after the duplication event. WGD paralogs, also called ohnologs, must also be distinguished from paralogs that arose by small-scale duplications or retrotransposition (Hahn 2009).

Orthology and paralogy relationships between genes are generally inferred from a phylogenetic tree. This gene tree represents the most likely evolutionary history from a common ancestral gene based on existing gene sequences. Reconciliation with the species phylogeny then labels gene duplication and speciation events. These phylogeny-reconciled gene trees allow rigorous, model-based tests to examine the evolution of gene sequences and functions. However, gene trees are known to contain errors related to methodological, technical, and biological factors (Som 2015). A prominent source of errors is low phylogenetic signal in sequences, that is, insufficient numbers of substitutions to confidently support one gene tree topology over others

(Rasmussen and Kellis 2007). To address this issue, species-tree-aware methods use proximity to the structure of the species tree, known as the reconciliation cost, to select among statistically equivalent gene tree topologies (Durand et al. 2006; Vilella et al. 2009; Rasmussen and Kellis 2011; Szöllősi et al. 2013; Wu et al. 2013; Scornavacca et al. 2015). Because of computational trade-offs, these methods either rely on heuristic exploration of the gene-tree solution space and result in suboptimal trees, or have an intensive computational cost and are not applicable to large data sets. Critically, these limitations are enhanced in the presence of ancient WGDs. Species descended from a WGD frequently have two paralogs or more per gene family, directly increasing the size of the solution space. Further, reconciliation methods do not account for the acceleration of gene loss rates shortly after the WGD, or for rate heterogeneity across species (Zwaenepoel and Van de Peer 2019).

However, we argue that WGDs also possess underexploited characteristics that can be leveraged to improve gene tree modeling. First, the known timing of well-supported WGD(s) can be integrated as prior knowledge to select or constrain gene tree topologies. Second, WGDs result in specific patterns of genome organization, where pairs of sister regions share duplicated genes in conserved order throughout the genome. These double-conserved syntenic (DCS) regions can be readily uncovered by comparison with unduplicated outgroup genomes (Jaillon et al. 2004; Kellis et al. 2004). Because reconciled gene trees are unreliable, DCS patterns are commonly used to identify WGD-duplicated genes with confidence (Byrne and Wolfe 2005; Catchen et al. 2009; Muffato et al. 2010). In multispecies studies, this evidence has typically been used to exclude thousands of gene families where synteny disagrees with the precomputed tree structure (Kassahn et al. 2009; Berthelot et al. 2014; Braasch et al. 2016), an imperfect solution that reduces statistical power and may lead to technical biases. However, synteny has never been used to systematically select among alternative gene tree topologies in the context of a known WGD event. To the best of our knowledge, only two gene tree-building methods are able to leverage evidence from gene organization to correct orthology and paralogy relationships: SYNERGY, a Neighbor-Joining iterative tree-building algorithm that is no longer available (Wapinski et al. 2007), and ParalogyCorrector, which identifies and corrects gene trees inconsistent with synteny information (Lafond et al. 2013). However, ParalogyCorrector is designed to remove unsupported duplication nodes and recover all true orthologs at the expense of paralogs, which makes it unsuited to WGD studies.

Here, we present Synteny-guided CORrection of Paralogies and Orthologies (SCORPiOs), a synteny-guided gene tree-building algorithm for WGD studies. SCORPiOs builds optimized, species-tree-aware gene trees consistent with known WGD events, local synteny context, and gene sequence evolution. We apply SCORPiOs to the teleost-specific genome duplication (TGD), dated 320 Ma (Jaillon et al. 2004) and find that almost one in four WGD-descended teleost gene trees in Ensembl are incorrect. We propose a corrected, synteny-consistent tree for 70% of these gene families (2,387 of 3,394 synteny-inconsistent trees). Then, we show that the corrected trees emphasize how duplicate gene retention after the WGD has shaped developmental and signaling pathways involved in known phenotypic innovations in the teleost lineage.

Results

Synteny Is Informative to Describe Gene Phylogenetic Relationships after WGD

After a WGD, gene positions along chromosomes display a particular conservation signature (DCS) that can be leveraged to identify and differentiate orthologous- and paralogous-duplicated regions across species (Jaillon et al. 2004; Kellis et al. 2004). This genomic organization reveals gene trees that are inconsistent with the trees of other syntenic genes in their local neighborhood (fig. 1*A*). Synteny may therefore be useful to systematically correct erroneous trees by sourcing additional evolutionary information from surrounding genes. To confirm that true WGD orthologs and paralogs can be distinguished based on information from their neighbors, we identified 2,394 reliable WGD gene duplicates in 229 unambiguous DCS regions in zebrafish and medaka covering 23% and 26% of each genome, respectively, using gene homologies from Ensembl as in previous studies (Materials and Methods; Kassahn et al. 2009; Braasch et al. 2016; supplementary table S1, Supplementary Material online). We then tested whether local gene neighborhoods differ between orthologous and paralogous WGD gene copies, in the absence of any correction. In theory, paralogous regions diverge immediately after the WGD event, in particular through massive gene losses, and should be more different in terms of gene retention, loss, and gene sequence evolution across species compared with orthologous genomic segments, which diverge at speciation (fig. 1*B*; Scannell et al. 2006). We found that orthologs indeed share more orthologous syntenic neighbors and have more similar local gene retention and loss patterns than paralogs, as expected (fig. 1*C*, Wilcoxon–Mann–Whitney tests, $P < 2.2e{-}16$). Local syntenic context is therefore informative to distinguish orthologs from paralogs, and can potentially be leveraged to correct erroneous gene trees systematically after a WGD event.

SCORPiOs: Synteny-Guided CORrection of Paralogies and Orthologies in Gene Trees

Integrating sequence and synteny information to build gene phylogenies is challenging, and current methods are not designed to deal with WGDs. To address this issue, we have developed SCORPiOs, an algorithm that improves gene trees in the presence of one or several WGD events, using information from synteny conservation patterns. Because it complements gene sequence evolution with syntenic information, SCORPiOs is suitable to study WGD events where synteny with outgroups species remains detectably conserved, such as the different fish WGDs, the baker's yeast WGD or most plant WGDs (Van de Peer et al. 2017). SCORPiOs is not intended for very ancient events such as the 1R or 2R vertebrate WGDs,

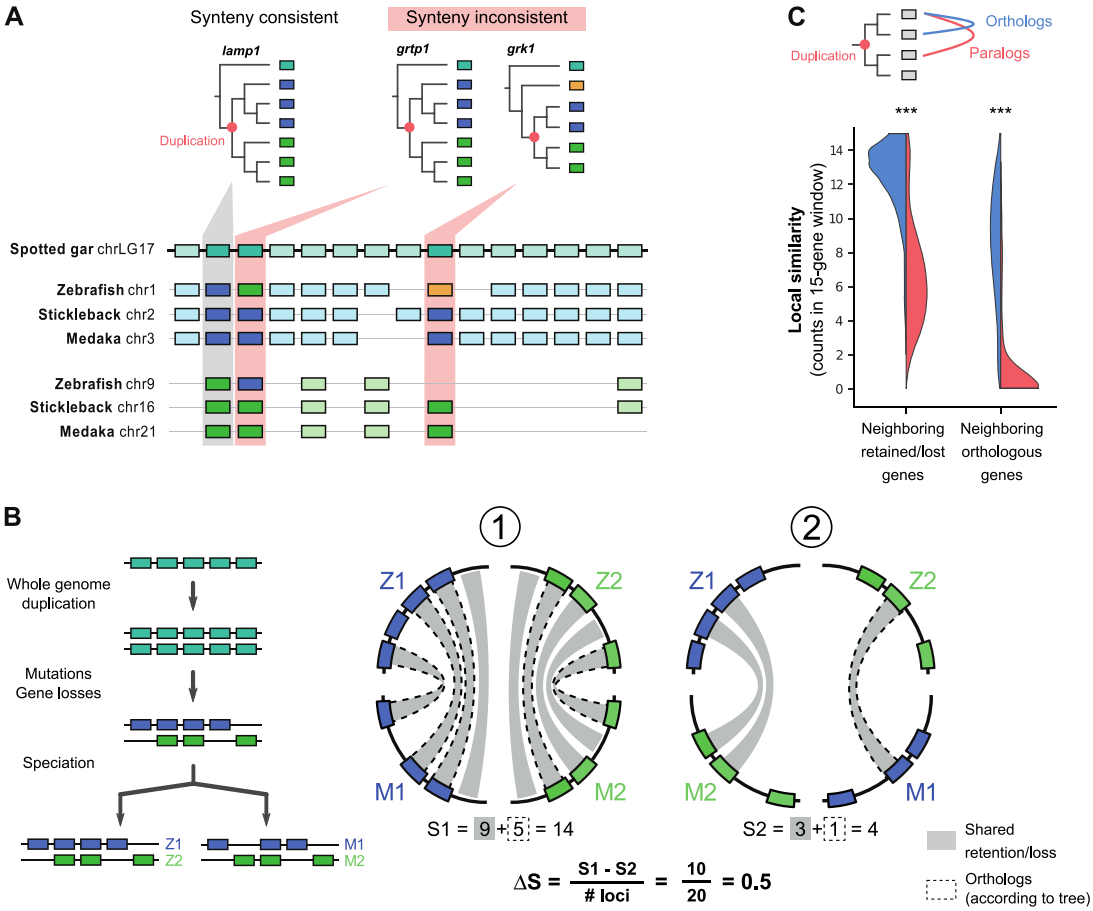

FIG. 1. Local synteny context is informative to sort out gene duplicates after WGD. (*A*) Gene trees and synteny context of the *lamp1*, *grtp1*, and *grk1* gene families in teleosts and their nonduplicated outgroup (spotted gar). Duplication nodes are labeled in red. Blue and green boxes represent members of the two orthology groups in each original gene tree from Ensembl; yellow denotes genes annotated as nonduplicated. The synteny context identifies gene trees where zebrafish homologs are assigned to the wrong orthology groups (highlighted in red). (*B*) Schematic evolution of a genomic segment after WGD, with gene colors as in (*A*). After WGD, the duplicated segments evolve independently and accumulate mutations and gene deletions. Zebrafish and medaka genomic segments are compared under two scenarios: Z1/M1 and Z2/M2 (scenario 1), which corresponds to true orthologs, and Z1/M2 and Z2/M1 (scenario 2), corresponding to paralogs. Under each scenario, genes similarly retained or lost across species are represented by gray links, and genes annotated as orthologs are linked by dotted lines. True orthologous segments (1) share more similar patterns of gene retentions/losses and orthologies than paralogous segments (2), resulting in a high four-way similarity score ΔS. (*C*) Local syntenic context can differentiate orthologs from paralogs. Distributions of the number of shared gene retentions/losses and annotated orthologs in a 15-gene window around zebrafish/medaka orthologs (in blue) and paralogs (in red), based on the original gene trees from Ensembl. Wilcoxon–Mann–Whitney tests, $n = 2{,}394$, $^{***}P < 0.001$.

where synteny with outgroups is highly degraded. SCORPiOs is coded in Python 3, implemented as a snakemake workflow (Köster and Rahmann 2012), and takes as inputs: the full set of 1) phylogenetic gene trees and 2) gene positions from extant, WGD-derived genomes along with one or several unduplicated outgroups; 3) the corresponding gene sequence alignments; and 4) the species phylogeny with the putative WGD position(s). For convenience, SCORPiOs also includes an implementation of TreeBeST (Vilella et al. 2009) and can initialize the process from sequence alignments if precomputed gene trees are not available. The major steps of SCORPiOs are outlined in figure 2 (see supplementary fig. S1, Supplementary Material online, for a detailed flowchart):

The implementation details of these different steps in SCORPiOs are developed below.

Comprehensive Identification of Homologous Families after the WGD Event

To identify duplicated regions within a genome, SCORPiOs takes advantage of patterns of DCS with a nonduplicated outgroup. After the WGD, genomic regions of the nonduplicated outgroup give rise to two orthologous genomic segments in each duplicated ingroup species (fig. 1*B*). However, genomic rearrangements, gene losses or duplications, and errors in the initial gene trees can obscure these 2:1 orthologous relationships in modern genomes. Most WGD studies thus restrain themselves to genomic regions where the WGD event is evident either from the gene trees or from highly conserved DCS patterns (Scannell et al. 2006; Kassahn et al. 2009; Berthelot et al. 2014; Inoue et al. 2015; Braasch et al. 2016). As a consequence, significant subsets of gene families are de facto discarded from analysis (54% not considered in

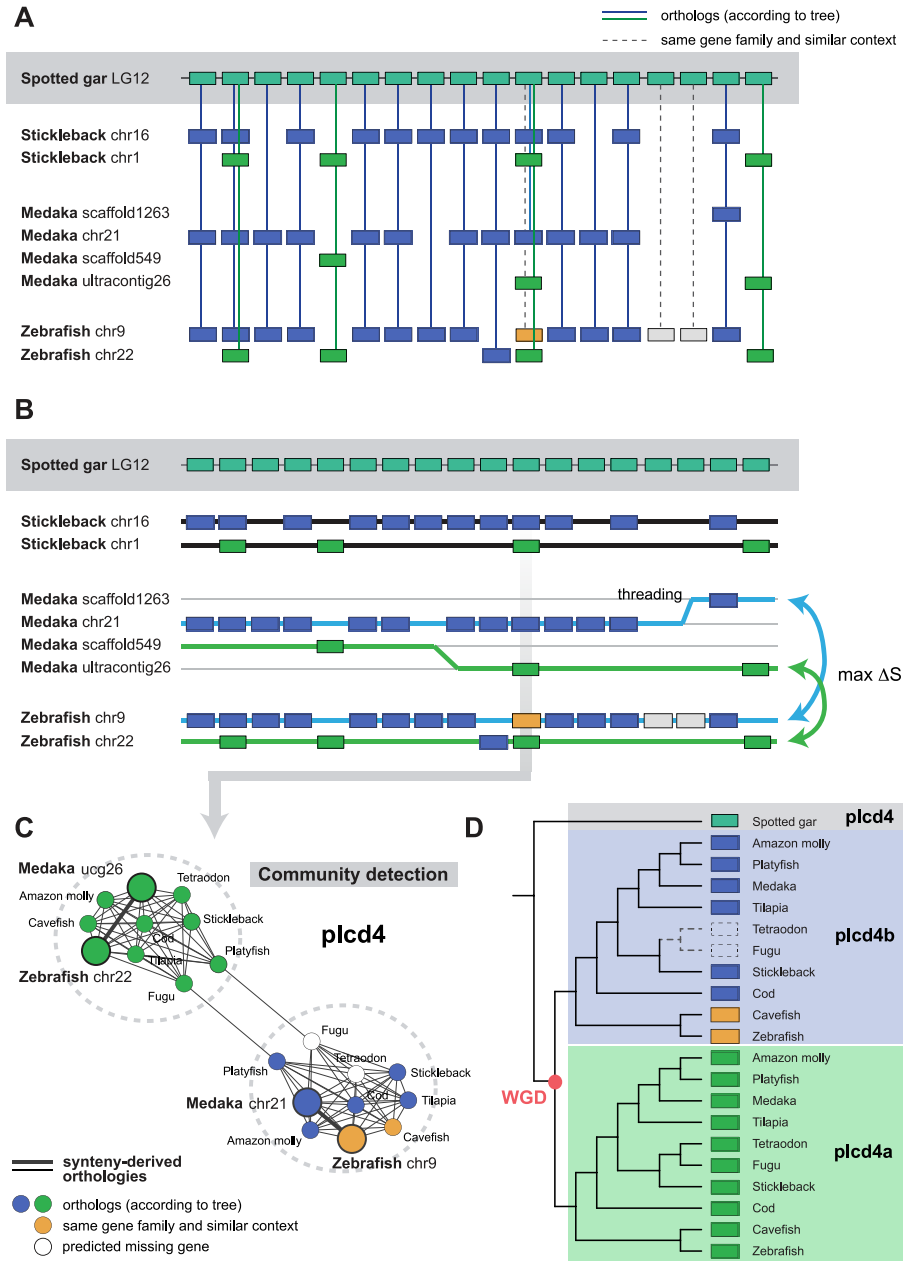

FIG. 2. Overview of the SCORPiOs workflow. (*A*) For every region in the reference nonduplicated genome, SCORPiOs identifies all potential orthologs in the duplicated genomes (orthologs and syntenic homologs) from the original gene trees. Here, a section of the spotted gar linkage group 12, and genomic locations of its potential orthologous genes in three TGD teleosts (out of ten). For each gene family, color represents genes identified as orthologs in the original gene trees from Ensembl. (*B*) SCORPiOs maximizes the synteny context similarity ΔS to thread and pair orthologous genomic segments between duplicated species (see supplementary note 2, Supplementary Material online). Green and blue threads show the highest scoring configuration for the medaka/zebrafish comparison. (*C*) For each individual gene, SCORPiOs integrates orthology links derived from the local synteny context across all species into a gene graph, where orthologous communities are identified. Here, the *plcd4* gene, whose genomic position is highlighted in (*B*). Nodes represent homologs of the *plcd4* gene, colored as in (*A*) and (*B*), where the medaka and zebrafish genes are highlighted as large circles. Links represent orthology relationships deduced from the similarity in syntenic contexts, with orthologies between zebrafish and medaka highlighted with a thicker line. SCORPiOs explicitly models nodes corresponding to missing homologs in the gene graphs (in white). Ucg, ultracontig. (*D*) If the original gene tree is not consistent with the synteny-derived orthologous gene communities, SCORPiOs proposes a corrected gene tree. Here, the corrected gene tree for the *plcd4* gene family, which includes zebrafish and cavefish homologs (in yellow) as part of the *plcd4b* subclade based on their genomic location (as seen in *B*).

Inoue et al.; 43% in Braasch et al. 2016). Omitting such genes from analysis because of incorrect trees or genomic location is highly problematic, and may result in patterns reflecting technical bias rather than biological insight.

SCORPiOs addresses this problem by first identifying all potentially WGD-descended gene duplicates in the initial gene trees. SCORPiOs uses a loose definition of orthology to build a comprehensive list of potential orthologs between

each gene in the nonduplicated reference genome and all duplicated genomes (fig. 2A and supplementary note 1 and fig. S2, Supplementary Material online). This table retains all gene copies since the ingroup/outgroup speciation node, as well as any other homologs from the same gene tree with a loosely similar syntenic context. Unlike other methods, SCORPiOs therefore first strives to define a comprehensive gene set from each gene family to be searched for potential orthology and paralogy relationships, instead of pruning all genes that do not belong to predefined duplicated segments.

Identification of Orthologous Genomic Segments between Pairs of Species

As a second step, SCORPiOs reconstructs ancestral WGD-duplicated regions, which may have become eroded by genome evolution processes. SCORPiOs uses the nonduplicated outgroup as a proxy for the ancestral gene order. It scans the outgroup genome with a sliding window of user-defined length, and uses the comprehensive list of orthologs to identify and thread the ancestral-duplicated segments in each descendant species (fig. 2B and supplementary note 2, Supplementary Material online). These duplicated genomic segments are then compared between pairs of duplicated species to match orthologous segments. SCORPiOs scores the syntenic similarity between segments using the number of genes annotated as orthologs in the initial gene trees, as well as the pattern of gene retention and losses, as described in figure 1. SCORPiOs is flexible and does not require perfect gene order conservation or contiguity to thread genomic segments, instead attempting to maximize the local similarity between both duplicated species (supplementary note 2, Supplementary Material online). This results in a four-way comparison of threaded, scored genomic segments for every pair of duplicated genomes and every window in the outgroup (fig. 2B).

SCORPiOs interprets the differentiation in four-way similarity scores (ΔS) as a confidence index that two genomic segments can be assigned as orthologs between the two duplicated species. This information is then reflected back to the genes within those segments to identify putative orthologs and paralogs using the window of highest ΔS that they belong to (fig. 1B and supplementary note 2 and fig. S4, Supplementary Material online). Of note, SCORPiOs considers that independent lineage-specific duplications are unlikely to result in duplicated genomic segments across different species that could be confused with the WGD. Paralogs in duplicated genomic regions consistent with the known WGD event are therefore considered as WGD paralogs, to the exclusion of other types of paralogs.

Orthology and Paralogy Constraints across All Species

As the previous step assigns orthologs and paralogs between pairs of species, SCORPiOs then integrates those results to define groups of orthologous genes across all species under study. SCORPiOs exhaustively performs all pairwise comparisons between duplicated genomes as described above, and constructs an orthology graph for every gene family (fig. 2C). SCORPiOs orthology graphs are unweighted graphs where two nodes (genes) are linked by an edge when they are inferred as orthologs based on their local synteny similarity. The graphs also include "placeholder" nodes, which correspond to gene copies that have been lost since the WGD but whose existence can be inferred from the synteny pattern (supplementary fig. S5, Supplementary Material online). In each graph, we then expect to find two similar-sized orthologous gene communities that are derived from the WGD. If synteny was perfectly informative and the process was error-free, we would expect these communities to be two independent, fully connected cliques. In practice, we observe that some graphs do not result in two isolated communities due to inconsistencies in the orthology assignments. SCORPiOs then uses the Girvan–Newman algorithm (GN), a community detection algorithm that iteratively removes the most central edges of each graph to separate nodes in two communities (Girvan and Newman 2002). In cases where GN is unable to separate genes of a duplicated species in two communities, we apply the Kerningan–Lin algorithm (KL) (Kernighan and Lin 1970). KL is a heuristic aimed at finding two communities of similar sizes, that require cutting the fewest edges. If the KL solution separates the duplicated genes from the same species more frequently, it is preferred over the GN solution. At the end of this step, SCORPiOs has identified the two groups of orthologous genes across extant species that each descend from a single ancestral gene in the common WGD ancestor, derived from local synteny information. Of note, one of these orthology groups can be entirely made of placeholders, corresponding to the loss of one duplicate gene in all species (typically an early loss before the first speciation, supplementary fig. S5, Supplementary Material online).

Gene Tree Correction and Test

Next, SCORPiOs identifies and attempts to correct gene trees that do not fulfill the orthologies and paralogies from their synteny-derived orthology graph. First, SCORPiOs checks whether each gene tree contains a node from which the unduplicated outgroup gene diverges, followed by either 1) a node corresponding to the WGD, under which each subtree encompasses one of the two orthology communities from the orthology graph, or 2) a single subtree, in cases where the same paralog is missing in all species (i.e., placeholder nodes subgraph; supplementary fig. S5, Supplementary Material online). When this topological constraint is not verified in the original gene tree (supplementary fig. S6, Supplementary Material online), SCORPiOs proposes a corrected, fully resolved gene tree that is both species tree and synteny-consistent. We explore the constrained topologies solution space using the fast distance-based approach implemented by the program ProfileNJ (Noutahi et al. 2016). Briefly, ProfileNJ independently resolves each multifurcated orthology group, so as to minimize the reconciliation cost with the species tree. SCORPiOs then compares the likelihoods of the original and corrected tree and accepts the correction if both trees are statistically equivalent (approximately unbiased test, Shimodaira and Hasegawa 2001; branch lengths fitted with PhyML, Guindon et al. 2010; Materials and Methods). If

ProfileNJ fails to find an adequate solution, SCORPiOs uses the TreeBeST-modified version of PhyML, which fits the gene sequences in each post-WGD subtree using maximum likelihood (ML) optimization while accounting for the species phylogeny topology (Vilella et al. 2009), and compares the corrected tree to the original one as above. We correct gene tree topologies only when the sequence similarity data are explained equally well (or better) by the synteny-aware tree: our correction approach is conservative, in the sense that it gives precedence to the likelihood of the molecular evolution model. Note also that SCORPiOs does not attempt to build a corrected tree for very large multigenic families (>1.5 genes per species in one or both post-WGD communities), as these are often overaggregated families in input trees that SCORPiOs cannot solve. As a final step of the pipeline, SCORPiOs can reinsert the corrected subtrees into the original gene tree depicting the evolution of the whole-gene family, and recompute branch lengths (supplementary note 3 and figs. S7–S9, Supplementary Material online). Additional available options when executing SCORPiOs are described in supplementary note 4, Supplementary Material online.

SCORPiOs Corrects a Significant Fraction of Neopterygii Subtrees in Ensembl

We applied SCORPiOs to gene trees from the Ensembl Compara database, version 89 (Vilella et al. 2009), which includes 70 vertebrate genomes, of which eleven are Neopterygii: ten duplicated teleost genomes, and the spotted gar as an outgroup to the TGD (supplementary fig. S10, Supplementary Material online). This set represents a total of 21,431 gene trees reconciled with the vertebrate species tree. About 50% of genes are syntenic between spotted gar and teleost genomes, which makes it a suitable outgroup and proxy for the ancestral gene order (Materials and Methods; Braasch et al. 2016).

We ran SCORPiOs in iterative mode, which corrects the gene trees the first time, thus improving the quality of the synteny information, and then again until convergence (window size: 15 genes, with a step of 1 gene). For this data set, two iterations were sufficient to reach convergence (see supplementary tables S2–S4, Supplementary Material online, for detailed results by iteration). In brief, in iteration 1, we identified 15,476 Neopterygii loose orthology groups within the 21,431 gene trees. SCORPiOs produced 14,576 orthology graphs, of which 10,172 (70%) were already subdivided into two isolated orthology cliques. The vast majority of families not producing graphs (86%) are small families (<3 genes), for which we do not build graphs, as they would result in the same tree topology (supplementary tables S2–S3, Supplementary Material online). For 3,394 gene trees (22%), the orthologies and paralogies relationships were in disagreement with the synteny graph. In total, at the end of the two iterations, we were able to find a synteny-aware and statistically equivalent gene tree for 2,387 of these genes families, thus correcting 70% of all synteny-inconsistent gene trees (15% of Ensembl Neopterygii subtrees). Interestingly, for 672 of the gene trees that we correct (28%), the new tree is better supported by the sequences and results in a significantly higher likelihood (AU test, $\alpha = 0.05$). This may reflect difficulty of the Ensembl pipeline to explore the tree topology space due to the high number of species in the database, as previously described (Wu et al. 2013). Furthermore, we also applied SCORPiOs to version 94 of the Ensembl Compara database, containing 47 teleost genomes and a total of 43,491 gene trees (supplementary fig. S11, Supplementary Material online). With this increase in phylogenetic coverage, a higher proportion of gene trees was inconsistent with synteny (7,168 synteny-inconsistent gene trees for 15,760 gene families, 45%), and could be improved by SCORPiOs (3,681 corrected gene trees, 23%). Again, this result is in line with the idea that exploration of the topology space becomes limited for larger trees, inducing more errors. Altogether, we demonstrate that insight from synteny can improve a significant fraction of gene trees in the presence of many gene duplicates after a WGD event.

Validation of SCORPiOs Trees and Comparison to Existing Methodologies

SCORPiOs corrects gene trees by providing synteny-derived orthology and paralogy constraints to existing tree-building programs, ProfileNJ and TreeBeST PhyML. To evaluate how SCORPiOs improves over the state of the art, we compared the 2,387 corrected teleost gene trees to alternatives topologies obtained with other methodologies. First, we used RAxML (Stamatakis 2014) to build a pure ML tree for each gene family, as a standard to compare against the likelihood of other trees. Second, we included the original, uncorrected trees from Ensembl Compara (Vilella et al. 2009), which are phylogeny-reconciled consensus trees build with the TreeBeST pipeline. Last, we tested ParalogyCorrector, the only other existing methodology that corrects gene trees using synteny information (Lafond et al. 2013). To compute ParalogyCorrector trees, we provided our synteny-derived orthologies constraints along with the original Ensembl subtree. ParalogyCorrector then finds a new gene tree minimizing the differences to the original tree while satisfying the orthology constraints (but not necessarily paralogy), whereas SCORPiOs fully recomputes each paralog subtree under the corrected duplication node.

We first evaluated whether each tree is well supported by the sequence alignment (AU test, $\alpha = 0.05$; fig. 3A). As expected, RAxML solutions are systematically the best fit to the sequence data. However, SCORPiOs is able to find a solution as good as the pure ML fit for 61% of gene trees, which is significantly better than the original Ensembl trees (47%; proportion test, $P < 2.2e{-}16$) or ParalogyCorrector (51%; proportion test, $P < 2.2e{-}16$). Additionally, the SCORPiOs trees are a better fit to the gene sequence information than either the Ensembl or the ParalogyCorrector trees for 9.5% of the 2,387 test gene families.

We then assessed whether the gene trees are concordant with the known species phylogeny. For each software, we identified the species-tree inconsistent nodes, previously referred to as "dubious nodes" or "nonapparent duplication nodes" (Vilella et al. 2009; Lafond et al. 2014; Materials and Methods). We find that SCORPiOs outperforms both RAxML

and the original Ensembl trees in terms of species-tree congruence (Wilcoxon signed rank test, P values $<2.2e{-}16$, fig. 3B). Both SCORPiOs and ParalogyCorrector produce trees that are highly congruent to the species phylogeny, with a slight advantage to ParalogyCorrector (Wilcoxon signed rank test, $P < 2.2e{-}16$, 88.6% and 99.7% of fully congruent trees, respectively; proportion test $P < 2.2e{-}16$). However, the noncongruent solution from SCORPiOs was better supported by sequence information for 146 of 269 gene families for which ParalogyCorrector reported a fully congruent solution, suggesting that the additional duplication nodes inferred by SCORPiOs are largely correct.

Lastly, we assessed whether gene trees inferred by all four methods are parsimonious by calculating their total number of duplication nodes. We found that SCORPiOs infers the lowest number of duplications in gene trees, compared with all other methods (Wilcoxon signed rank test, P values $< 2.2e{-}16$, fig. 3C). By design, SCORPiOs explicitly replaces the WGD node at its known location in the tree when both gene copies survive in modern genomes (fig. 3D), resulting in fewer inferred duplications downstream of the WGD node. In contrast, ParalogyCorrector places duplications closer to the leaves and wrongly identifies most WGD duplicates as species-specific duplications. For 38% of Ensembl gene trees, the duplication is placed at the Neopterygii node and incorrectly encompasses the spotted gar gene (nonduplicated outgroup). In conclusion, SCORPiOs significantly improves gene trees in an evolutionary context where other methods struggle, due to the high number of gene copies and nonparsimonious tree structures created by WGD events.

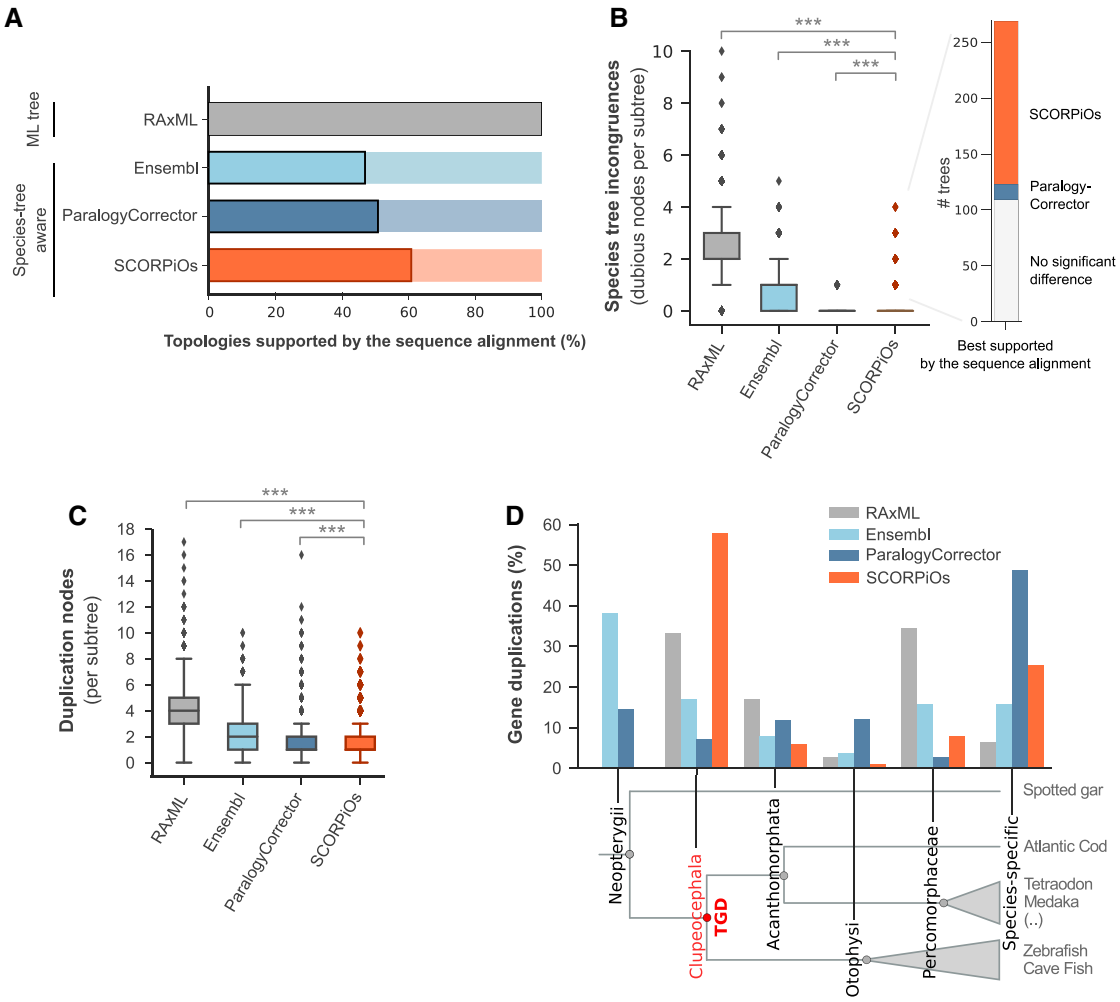

FIG. 3. SCORPiOs performs better than state-of-the-art methods on WGD gene trees. (*A*) SCORPiOs produces gene trees with higher phylogenetic support than either ParalogyCorrector or the Ensembl Compara pipeline. Bars represent the fraction of trees where the topology proposed by either of the three species-tree aware methods is not significantly less likely than the maximum-likelihood, nonreconciled solution from RAxML (AU tests, $P \geq 0.05$). (*B*) SCORPiOs and ParalogyCorrector both produce fewer species tree inconsistencies than RAxML or Ensembl Compara in teleost gene subtrees. Boxplots represent the distributions of dubious, low-support duplication nodes per tree; dots represent the top 5% outlier trees. Right inset, stacked barplot: when SCORPiOs introduces more dubious duplication nodes than ParalogyCorrector, we compared the likelihood of both solutions according to sequence evolution and found that the solution from SCORPiOs is equally or more likely for 255 out of 269 gene families. (*C*) Number of duplication nodes for each method. (*D*) Position of duplication nodes in the species tree for each method. The Clupeocephala ancestor corresponding to the expected TGD node is highlighted in red. Intermediate ancestors with fewer than 5% of duplications for all methods are not displayed.

SCORPiOs Corrections Improve Correlation of Orthologous Gene Expression

In addition to their consistency with sequence evolution and phylogeny, we evaluated whether orthologs corrected by SCORPiOs are functionally more similar than paralogs, as previously reported (Koonin 2005; Altenhoff et al. 2012; Chen and Zhang 2012). We used gene expression data from 11 tissues in zebrafish and medaka (Pasquier et al. 2016) to investigate whether SCORPiOs-corrected zebrafish/medaka orthologs display higher similarity in expression patterns (Materials and Methods). For instance, SCORPiOs modified the orthology relationships in the *cxcl12* gene family for zebrafish and medaka (fig. 4A), grouping together medaka *cxcl12a* and zebrafish *cxcl12b* in one orthology group and medaka *cxcl12b* and zebrafish *cxcl12a* in the other based on their local syntenic surroundings. Indeed, gene expression patterns support the orthology reassignment, with one gene copy expressed in bones, brain, embryo, gills, and liver, and the other expressed at high level, predominantly in kidney, in both species.

Overall, SCORPiOs modified the orthology relationships between zebrafish and medaka for 761 gene families. Of these, 210 correspond to orthology/paralogy reassignments (as in fig. 4A), whereas the remaining 551 correspond to removal of errors or addition of new homology relationships. The correction increased the number of 1-to-1 orthologs between zebrafish and medaka (13,463 vs. 14,008) as well as the total number of genes with an ortholog in the other species (16,150 zebrafish and 15,543 medaka genes with an ortholog before correction, vs. 16,316 and 15,748 after). For the 210 gene families where medaka and zebrafish orthologies were reassigned, we find that orthologs are expressed at closer average levels after correction (Wilcoxon signed rank test, P value = 0.0171, fig. 4B) and also significantly more correlated across tissues (Wilcoxon signed rank test, P value = 0.0050, fig. 4C). These results support that SCORPiOs measurably improves orthology and paralogy relationships. Additionally, they suggest that erroneous orthology relationships may obfuscate functional investigation of gene evolution after genome duplication, especially when their effects get compounded over dozens of species and thousands of gene families, as reported above for teleosts.

SCORPiOs Correction Emphasizes WGD Contributions to Evolutionary Innovations in Teleost Fish

Numerous studies have suggested a link between the function of a gene and retention or loss after a WGD. Yet, in the absence of systematic gene tree correction methods, the fate of TGD duplicates has only been investigated in a restricted set of ~6,000 high-confidence teleost gene families (Kassahn et al. 2009; Inoue et al. 2015; Braasch et al. 2016), potentially introducing biases in subsequent conclusions. Here, we used the full set of 21,431 gene trees from Ensembl corrected by SCORPiOs to investigate gene retention across ten teleost species (zebrafish, cavefish, tetraodon, fugu, stickleback, medaka, tilapia, platyfish, amazon molly, and cod). We classified genes into three categories with respect to their fate after the TGD (Materials and Methods). Briefly, we grouped genes retained in two copies across all ten teleost species ("systematic ohnologs," $n = 1,828$), genes found in single copy in all species ("singletons," $n = 13,895$), and genes retained in two copies in at least one teleost species but not in all ("facultative ohnologs," $n = 7,265$) (supplementary fig. S12, Supplementary Material online). We then used expression levels and functional annotations in zebrafish to explore how gene function relates to evolutionary trajectory after the TGD (Materials and Methods).

Overall, we find that singletons have slightly higher average expression levels and broader expression patterns than both systematic and facultative ohnologs (Wilcoxon–Mann–Whitney test, $P < 0.001$, fig. 5A, supplementary fig. S13, Supplementary Material online, Materials and Methods). This is in line with previous observations on paralogs and may reflect cases of subfunctionalization between the two groups of ohnologs, for which duplicated genes have partitioned the ancestral function, becoming expressed in fewer tissues and/or at lower level (Huminiecki and Wolfe 2004; De Smet et al. 2013; Guschanski et al. 2017). We next investigated whether tissue-specific singletons and ohnologs display preferential expression in different tissues, reflecting different contributions to teleost evolution. We find that tissue-specific systematic ohnologs are overrepresented in brain, heart, and muscle, and depleted in liver, intestine, ovary, and testis, compared with all zebrafish tissue-specific genes (tau > 0.9; hypergeometric tests, corrected $P < 0.05$, fig. 5B). In contrast, tissue-specific singletons genes are overrepresented in liver, kidney, intestine, and testis (fig. 5B). For facultative ohnologs, we observe an enrichment in brain-specific genes, but also in muscle-specific expression (fig. 5B). Interestingly, enrichment in brain- and heart-specific genes, as well as depletion in liver- and testis-specific genes, have been already observed in human ohnologs retained after the 1R and 2R vertebrate WGDs (Guschanski et al. 2017). This result ties in with previous reports that some gene families and functional categories are recurrently amplified by independent WGDs, possibly because they offer adaptive advantages when duplicated en masse (van Hoek and Hogeweg 2009; De Smet and Van de Peer 2012).

Additionally, we investigated whether TGD ohnologs and singletons belong to different biological pathways using Gene Ontology Biological Processes (GO BP) and KEGG pathway enrichment analyses (Materials and Methods). Systematic and facultative ohnologs are enriched in general molecular processes previously found in WGD duplicates, linked to transcriptional regulation and metabolic processes, as well as terms related to the nervous system, consistent with the brain-specific expression patterns reported above (fig. 5C and supplementary tables S5–S8, Supplementary Material online) (Blomme et al. 2006; Inoue et al. 2015; Singh et al. 2015; Li et al. 2016; Pasquier et al. 2017; Roux et al. 2017) In contrast, singletons are enriched in housekeeping functions, with some of the most significant GO terms being "cell cycle," "nucleic acid metabolic process," and "cellular localization," along with the KEGG pathways "Ribosome" and "DNA replication" (fig. 5C and supplementary tables S9 and S10, Supplementary Material online) (De Smet et al. 2013; Li

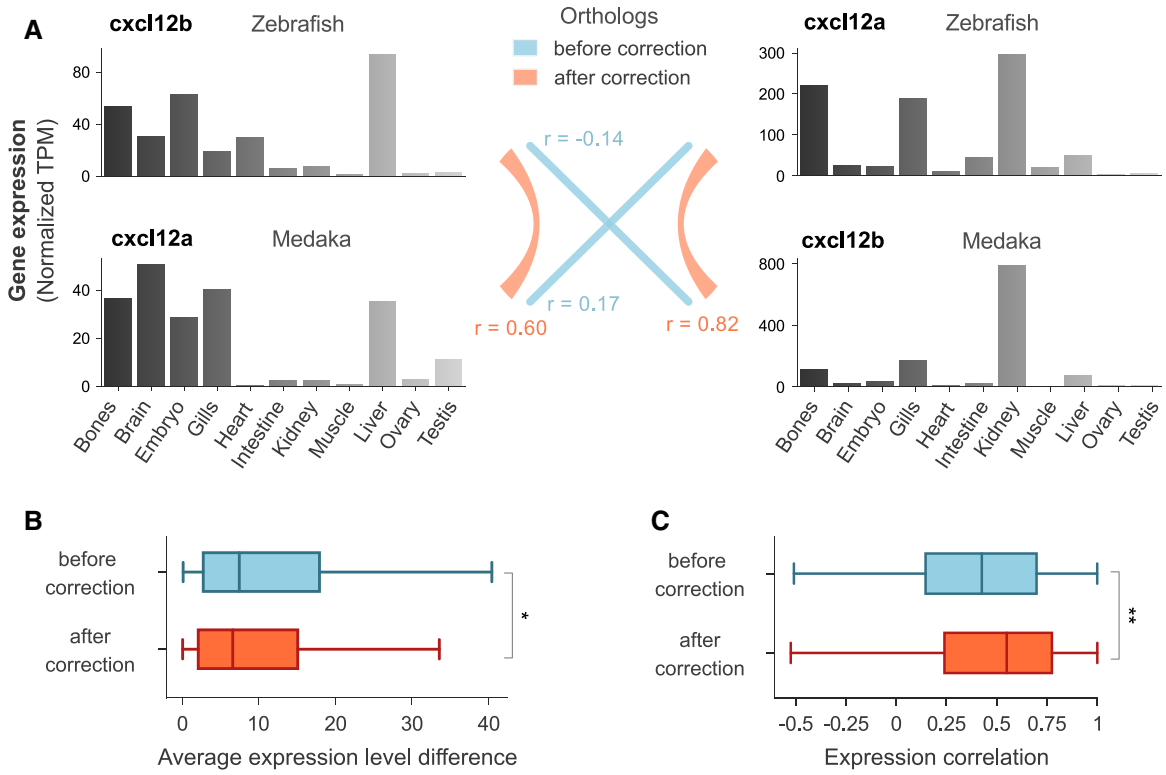

FIG. 4. Functional similarity of orthologous genes after SCORPiOs correction. (*A*) Expression of zebrafish and medaka *cxcl12* homologs in 11 tissues (quantile-normalized transcripts per million, TPM). Orthology relationships and expression level correlations before and after correction by SCORPiOs are noted in color (*r*, Pearson correlation coefficient). (*B*) Difference in average expression levels across 11 tissues between orthologous zebrafish/medaka genes, before and after correction (paired Wilcoxon test, $n = 210$, $^{*}P < 0.05$). (*C*) Average correlation of expression levels for orthologous zebrafish/medaka genes before and after correction (paired Wilcoxon test, $n = 210$, $^{**}P < 0.01$).

et al. 2016). However, we discover here that systematic duplicates are also enriched in more specific functions, especially related to retina physiology. Interestingly, the teleost WGD coincides with functional innovations in the retina specific to this clade, where photoreceptor cells are organized in a regular pattern described as a "cone mosaic" (Lyall 1957; Engström 1963; Sukeena et al. 2016). These results are mirrored in the KEGG analysis with the overrepresentation of the taurine and hypotaurine metabolism pathway, suggested to have a functional role in the teleost retina (Lima et al. 1998; Omura and Inagaki 2000). Our results therefore support that the amplification of retinal genes during the teleost WGD was important in the acquisition of this evolutionary innovation.

Finally, some functional enrichments become prominent only after gene tree correction with SCORPiOs (in bold on fig. 5). In particular, we observe an enrichment for both systematic and facultative ohnologs toward terms related to the circulatory system ("Adrenergic signaling in cardiomyocytes," "Vascular smooth muscle contraction," "VEGF signaling pathway"; fig. 5C, supplementary tables S6 and S8, Supplementary Material online). This extends broader support to a previous report that TGD-derived duplicates, especially those of the *elastin* gene, have led to morphological sophistication of the teleost heart and circulatory system (Moriyama et al. 2016). Lastly, genes in the facultative ohnolog category are enriched in the "Melanogenesis" pathway, also consistent with the expansion of the pigmentation repertoire after the TGD (Lorin et al. 2018). Taken together, our results suggest a strong contribution of TGD duplicates to functional novelty in this clade, mediated by the fixation of specialized duplicated genes. Importantly, many of these enrichments were fully obscured by errors in the gene evolutionary histories downloaded from as respected a reference database as Ensembl, which is widely sourced for comparative and evolutionary studies (Alföldi and Lindblad-Toh 2013; Herrero et al. 2016). SCORPiOs therefore fulfills its purpose in the arsenal of tree-building tools and has potential to dramatically further investigations into the evolutionary and functional outcomes of WGD events.

Discussion

Gene retention and loss after a WGD is a poorly understood interplay between functional redundancy, increased evolvability, and mutational cost. Several complementary models of gene evolution have been proposed to account for duplicate retention after WGD, including neofunctionalization, where one copy acquires a new function, whereas the other maintains the original one (Ohno 1970; Lynch and Conery 2000); subfunctionalization, where both copies partition the ancestral function between themselves (Force et al. 1999); dosage balance, where subunits of macromolecular complexes are maintained as duplicates to ensure proper stoichiometry between interacting partners (Blomme et al. 2006; Veitia et al. 2008; Makino and McLysaght 2010); and cost of

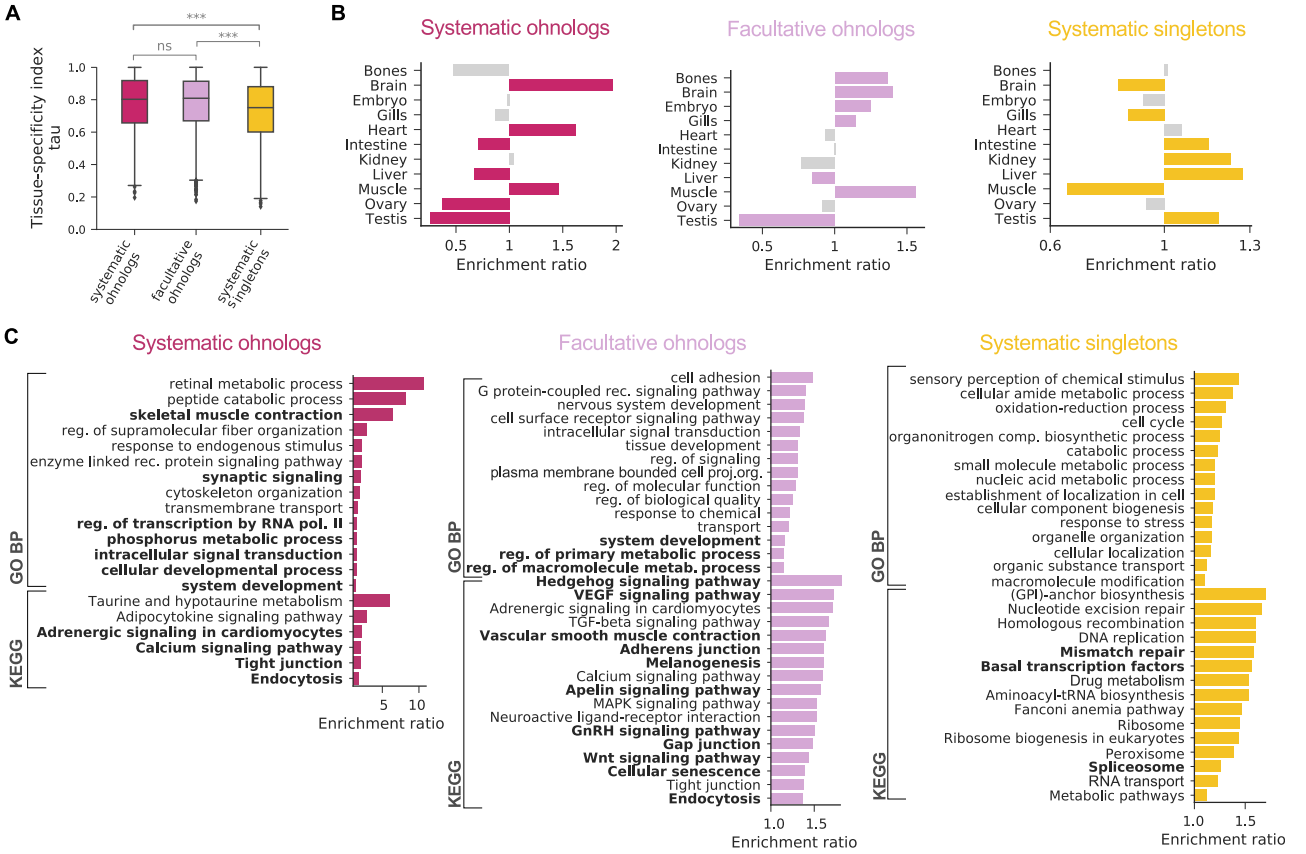

FIG. 5. Functional analysis of genes with different evolutionary trajectories after the TGD. (*A*) Tissue specificity of zebrafish genes with different evolutionary trajectories. Systematic ohnologs: WGD duplicates retained in two copies in all ten teleost species under study ($n = 1,828$). Facultative ohnologs: WGD duplicates retained in two copies in at least one species ($n = 7,265$). Systematic singletons: WGD duplicates returned to single-copy state in all ten species ($n = 13,895$). (*B*) Preferential tissue of expression for tissue-specific genes (tau > 0.9) from each evolutionary trajectory. Colors denote statistical significance (hypergeometric test with BH correction, $P < 0.05$). (*C*) Gene ontology biological process and KEGG pathways enrichments for tissue-specific genes from each evolutionary trajectory (hypergeometric test with BH correction, $P < 0.05$). Bold, only enriched after gene tree correction with SCORPiOs.

deleterious mutations, impeding pseudogenization, and loss (Gout et al. 2010; Singh et al. 2012). Overall, the relative contributions of these processes to short- and long-term gene evolution remain unclear, although it is generally agreed that the evolutionary fate of genes following WGDs is tightly intertwined with their ancestral functions.

Polyploidizations are widespread through eukaryotic evolution, representing many independent opportunities to characterize gene evolution after WGD. Yet, WGDs represent a serious challenge to current gene tree reconstruction methods due to the high volume of duplicates and gene losses that they produce. As a result, incertitudes in gene phylogenies have been a limiting factor to all WGDs studies, whether they investigate the incidence and timing of ancient WGDs (Van de Peer et al. 2010; Ruprecht et al. 2017; Zwaenepoel and Van de Peer 2019), biases in duplicate gene retention (Scannell et al. 2006; Kassahn et al. 2009; Inoue et al. 2015), or genome organization evolution (Varadharajan et al. 2018). Studying WGD occurrences and consequences calls for the development of specific methodologies to characterize the complex histories of WGD genes. Illustrating on-going efforts, the recently published WHALE approach accounts for incertitude in gene tree reconciliations when identifying plausible WGDs in a species tree (Zwaenepoel and Van de Peer 2019). SCORPiOs fills another methodological gap by integrating insights from genome evolution to improve reconciled gene trees.

Genome evolution operates through three major mechanisms: nucleotide substitutions, gene duplications and losses, and genomic rearrangements. To date, integrating these different evolutionary events into a unified framework remains an open challenge (Chauve et al. 2013). The inference of reconciled gene trees, which jointly depict the history of substitutions and gene gains and losses, has been growingly addressed in recent years (Szöllősi et al. 2015). Synteny conservation has the potential to neatly complement sequence similarity in gene evolution studies, because gene order evolves via independent mechanisms, and is more resilient to saturation at deep evolutionary times (Rokas and Holland 2000). Genome organization information still remains difficult to incorporate in gene phylogenies, largely due to the lack of well-supported evolutionary models (Chauve et al. 2013) and the need for contiguous genome assemblies. The most notable effort to use extant synteny to correct gene trees in a general context showed mixed results (ParalogyCorrector within the RefineTree framework; Lafond et al. 2013;

Noutahi et al. 2016). However, we show here that in contexts where additional priors on genome organization can be leveraged, synteny patterns can be highly informative and effectively improve reconciled gene trees. As high-quality reference genome assemblies are becoming affordable and straightforward, we expect that synteny will become increasingly useful to gene history resolution in an ever more complex comparative genomics landscape. For instance, synteny-aware methods will allow the investigation of other significant biological events, such as gene conversions, which introduce discordances in the history of a gene sequence and the history of its locus.

Lastly, assessing the quality of gene trees remains a challenging task, simply because the true evolutionary history of a gene is unknown. Although statistical likelihood is widely used as a goodness-of-fit criteria to evaluate gene trees, the tree of ML according to sequence evolution is frequently incorrect (Shimodaira 2002; Szöllősi et al. 2015). It is generally assumed that the correct tree falls within an interval of equally supported trees, but numerous factors can invalidate this hypothesis, ranging from errors in sequences or their alignment to unrealistic assumptions of evolutionary models. Consequently, other goodness-of-fit metrics have been introduced, including measures of species-gene tree discordance, parsimony of the duplication and loss scenario, distances to gold standard or simulated trees, functional similarity of orthologs, and power to reconstruct ancestral genomes (Altenhoff et al. 2016; Noutahi et al. 2016). Their use has been heterogeneous across studies, guided by specific aims, relevance, and feasibility. Here, we validate SCORPiOs on real data, taking full advantage of computable metrics to demonstrate the improved quality of SCORPiOs corrected trees. In the future, efforts toward standardized benchmarking, as led by the Quest for Orthologs community, will be instrumental in producing ever more accurate gene phylogenetic trees.

Materials and Methods

Synteny Similarity of WGD Orthologs and Paralogs in the Absence of Tree Correction

Gene trees constructed with TreeBeST were downloaded from Ensembl v.89 (Vilella et al. 2009) and used to define initial orthology and paralogy relationships before correction. We extracted a set of 2,394 high-confidence, WGD-descended homologous gene pairs in zebrafish and medaka using synteny criteria similar to approaches in previous studies (Kassahn et al. 2009; Braasch et al. 2016). Well-defined WGD duplicated regions were identified in the medaka and zebrafish genomes, where one contiguous window of 15 genes or more in the spotted gar genome has homologs on exactly two different chromosomes in both medaka and zebrafish. In total, these 229 regions include 5959 genes in zebrafish (23%) and 5193 in medaka (26%). Medaka and zebrafish gene pairs were considered high-confidence WGD duplicates when they are located at the midpoint of one such 15-gene window. The orthology and paralogy relationships between those gene copies were extracted from the original gene trees (1696 orthologous and 698 paralogous gene pairs). For each medaka-zebrafish gene pair (orthologs or paralogs), we counted across the 15-gene window: 1) the number of orthologs genes between medaka and zebrafish, according to the original trees, and 2) the number of homologs to spotted gar genes similarly retained or lost in both species.

Genome-wide synteny conservation was calculated between spotted gar (used as the outgroup) and other teleost genomes in the absence of tree correction using PhylDiag with default parameters (Lucas et al. 2014).

Gene Tree Comparisons

Nucleotide sequence alignments containing teleost fish sequences were downloaded from Ensembl v.89 and pruned of non-Neopterygii sequences. For each tree topology inferred by either TreeBeST, ParalogyCorrector or SCORPiOs, phylogenetic likelihood was computed with PhyML using the HKY85 model (Guindon et al. 2010). Likelihoods for alternative topologies were compared using the AU test implemented in Consel, at $\alpha = 0.05$ (Shimodaira and Hasegawa 2001).

Dubious nodes (or nonapparent duplication nodes) are gene duplications inferred by the reconciliation procedure where no descendant species actually contains two genes copies. They correspond to inconsistencies between the gene and species trees and are likely errors in the gene tree topology. To find dubious nodes, we used treebest sdi to reconcile gene trees with the species tree and identified all duplication nodes with a confidence score of 0 (Vilella et al. 2009).

Gene Expression Analysis

We used RNA-seq data sets from the PhyloFish database (Pasquier et al. 2016), which provides transcriptomes in fish for the following tissues: bones, brain, embryo, gills, heart, intestine, kidney, liver, muscle, ovary, and testis. We used kallisto (Bray et al. 2016) with default parameters to quantify transcript abundances for the full set of Ensembl transcripts in zebrafish and medaka. We summed transcripts per million (TPM) values of alternative transcripts to obtain a quantification of the expression of their corresponding gene. Finally, TPM values were quantile normalized within each species to obtain equivalent distributions of gene expression levels across tissues (Bolstad et al. 2003).

Functional Similarity of Orthologs

We assessed the functional similarity of zebrafish-medaka orthologs before and after gene tree correction. From the 2,387 corrected gene families, we selected the subset of 210 trees where zebrafish and medaka orthologies were reassigned by SCORPiOs. Orthologous gene expression levels were compared before and after correction using Pearson correlation and differences in mean expression across tissues. Average correlation and average expression difference before and after correction were compared using Wilcoxon signed rank tests. All tests are paired to ensure that results are unbiased by heterogeneous evolutionary rates across gene families.

Evolutionary Categories of Genes

We used the corrected set of gene trees to classify zebrafish genes with respect to their evolutionary fate across species after the TGD. We extracted all teleost gene clades from the trees and used the duplication status of the root node to determine the fate of the descending genes across species. If the root node is not a duplication, then all genes returned to a single-copy state after the TGD and we classify descending zebrafish genes as "systematic singletons." If the root node is a duplication (corresponding to the TGD), and all descending species retained both duplicated copies (duplication confidence score = 1), we defined them as "systematic ohnologs." Finally, if more than one but not all species retained the two copies (duplication confidence score < 1), we classify genes as "facultative ohnologs." We excluded from this classification 2,086 zebrafish genes with no other teleost homolog in their respective subtrees.

Expression of Genes with Different Trajectories after the TGD

We used the tau index (Yanai et al. 2005) to assess the degree of tissue specificity of the expression of zebrafish genes. Tau varies between 0 and 1, where 0 means broad expression and 1 specific expression. We defined tissue-specific genes as genes with a tau index >0.9 and tested for enrichment of genes specific to particular tissues using hypergeometric tests with Benjamini and Hochberg correction for multiple testing.

Functional Enrichment of Genes with Different Trajectories after the TGD

We used WebGestalt (Liao et al. 2019) to search for functional enrichment in each evolutionary category of genes, with all zebrafish protein-coding genes as background. For systematic ohnologs, 786 and 496 out of 1,828 genes were mapped to an annotation in GO BP and KEGG pathway, respectively, 5,496 and 3,392 out of 13,895 in systematic singletons, and 2,698 and 1,625 out of 7,265 in facultative ohnologs. WebGestalt uses hypergeometric tests, corrected for multiple testing with the Benjamini and Hochberg procedure, to test for significant enrichment. We report significant enrichments at a threshold of corrected *P* value <0.01. For visualization purposes (fig. 5C), we reduced redundancy of functional GO terms to a maximum of 15, using the weighted set cover method implemented in Webgestalt. In all cases, the reduced set covers >97% of the total genes in each category. We repeated the analysis starting from the uncorrected Ensembl gene tree set to determine if the enriched annotations differ after applying SCORPiOs.

Availability and Implementation

SCORPiOs is coded in Python 3 and implemented as a snakemake workflow, supported on Linux and macOS. Code is publicly available on Github at https://github.com/DyogenIBENS/SCORPIOS. SCORPiOs is distributed under the GNU GPLv3 license. Teleost gene trees, before and after correction, have been deposited to Zenodo (doi:10.5281/zenodo.3727519).

Supplementary Material

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

Acknowledgments

We thank Pierre Vincens for the coordination of computing resources and all members of the GenoFish consortium for fruitful discussions. This work is funded by ANR GenoFish (Grant No. ANR-16-CE12-0035), 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* Université Paris).

References

Alföldi J, Lindblad-Toh K. 2013. Comparative genomics as a tool to understand evolution and disease. *Genome Res*. 23(7):1063–1068.

Altenhoff AM, Boeckmann B, Capella-Gutierrez S, Dalquen DA, DeLuca T, Forslund K, Huerta-Cepas J, Linard B, Pereira C, Pryszcz LP, Quest for Orthologs Consortium, et al. 2016. Standardized benchmarking in the quest for orthologs. *Nat Methods*. 13(5):425–430.

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(5):e1002514.

Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noël B, Bento P, Silva CD, 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(1):10.

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.

Bolstad BM, Irizarry RA, Åstrand M, Speed TP. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19(2):185–193.

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(4):427–437.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*. 34(5):525–527.

Byrne KP, Wolfe KH. 2005. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res*. 15(10):1456–1461.

Catchen JM, Conery JS, Postlethwait JH. 2009. Automated identification of conserved synteny after whole-genome duplication. *Genome Res*. 19(8):1497–1505.

Chauve C, El-Mabrouk N, Guéguen L, Semeria M, Tannier E. 2013. Duplication, rearrangement and reconciliation: a follow-up 13 years later. In: Chauve C, El-Mabrouk N, Tannier E, editors. Models and algorithms for genome evolution. computational biology. London: Springer London. p. 47–62. Available from: https://doi.org/10.1007/978-1-4471-5298-9_4.

Chen X, Zhang J. 2012. The ortholog conjecture is untestable by the current gene ontology but is supported by RNA sequencing data. *PLoS Comput Biol*. 8(11):e1002784.

De Smet R, Adams KL, Vandepoele K, Van Montagu MCE, 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.

De Smet R, Van de Peer Y. 2012. Redundancy and rewiring of genetic networks following genome-wide duplication events. *Curr Opin Plant Biol*. 15(2):168–176.

Durand D, Halldórsson BV, Vernot B. 2006. A hybrid micro-macroevolutionary approach to gene tree reconstruction. *J Comput Biol*. 13(2):320–335.

Engström K. 1963. Cone types and cone arrangements in teleost retinae1. *Acta Zool*. 44(1–2):179–243.
Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151(4):1531–1545.
Girvan M, Newman M. 2002. Community structure in social and biological networks. *Proc Natl Acad Sci U S A*. 99(12):7821–7826.
Gout J-F, Kahn D, Duret L, Paramecium Post-Genomics Consortium. 2010. The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. *PLoS Genet*. 6(5):e1000944.
Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*. 59(3):307–321.
Guschanski K, Warnefors M, Kaessmann H. 2017. The evolution of duplicate gene expression in mammalian organs. *Genome Res*. 27(9):1461–1474.
Hahn MW. 2009. Distinguishing among evolutionary models for the maintenance of gene duplicates. *J Hered*. 100(5):605–617.
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* 2016:bav096.
Huminiecki L, Wolfe KH. 2004. Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. *Genome Res*. 14(10a):1870–1879.
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.
Kassahn KS, Dang VT, Wilkins SJ, Perkins AC, Ragan MA. 2009. Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates. *Genome Res*. 19(8):1404–1418.
Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 428(6983):617–624.
Kenny NJ, Chan KW, Nong W, Qu Z, Maeso I, Yip HY, Chan TF, Kwan HS, Holland PWH, Chu KH, et al. 2016. Ancestral whole-genome duplication in the marine chelicerate horseshoe crabs. *Heredity* 116(2):190–199.
Kernighan BW, Lin S. 1970. An efficient heuristic procedure for partitioning graphs. *Bell Syst Tech J*. 49(2):291–307.
Koonin EV. 2005. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet*. 39(1):309–338.
Köster J, Rahmann S. 2012. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* 28(19):2520–2522.
Lafond M, Chauve C, Dondi R, El-Mabrouk N. 2014. Polytomy refinement for the correction of dubious duplications in gene trees. *Bioinformatics* 30(17):i519–i526.
Lafond M, Semeria M, Swenson KM, Tannier E, El-Mabrouk N. 2013. Gene tree correction guided by orthology. *BMC Bioinformatics* 14(S15):S5.
Li Z, Defoort J, Tasdighian S, Maere S, Van de Peer Y, Smet RD. 2016. Gene duplicability of core genes is highly consistent across all angiosperms. *Plant Cell* 28(2):326–344.
Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. 2019. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res*. 47(W1):W199–W205.
Lima L, Obregón F, Matus P. 1998. Taurine, glutamate and GABA modulate the outgrowth from goldfish retinal explants and its concentrations are affected by the crush of the optic nerve. *Amino Acids* 15(3):195–209.
Lorin T, Brunet FG, Laudet V, Volff J-N. 2018. Teleost fish-specific preferential retention of pigmentation gene-containing families after whole genome duplications in vertebrates. *G3 (Bethesda)* 8:1795–1806.
Lucas JM, Muffato M, Crollius HR. 2014. PhylDiag: identifying complex synteny blocks that include tandem duplications using phylogenetic gene trees. *BMC Bioinformatics* 15(1):268.
Lyall AH. 1957. Cone arrangements in teleost retinae. *J Cell Sci*. s3–s98:189–201.
Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290(5494):1151–1155.
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.
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.
Muffato M, Louis A, Poisnel C-E, Crollius HR. 2010. Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. *Bioinformatics* 26(8):1119–1121.
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(8):e0159559.
Ohno S. 1970. Polyploidy: duplication of the entire genome. In: Ohno S, editor. Evolution by gene duplication. Heidelberg (Berlin): Springer. p. 98–106. Available from: https://doi.org/10.1007/978-3-642-86659-3_17.
Omura Y, Inagaki M. 2000. Immunocytochemical localization of taurine in the fish retina under light and dark adaptations. *Amino Acids* 19(3–4):593–604.
Pasquier J, Braasch I, Batzel P, Cabau C, Montfort J, Nguyen T, Jouanno E, Berthelot C, Klopp C, Journot L, et al. 2017. Evolution of gene expression after whole-genome duplication: new insights from the spotted gar genome. *J Exp Zool Mol Dev Evol*. 328(7):709–721.
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(1):368.
Rasmussen MD, Kellis M. 2007. Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes. *Genome Res*. 17(12):1932–1942.
Rasmussen MD, Kellis M. 2011. A Bayesian approach for fast and accurate gene tree reconstruction. *Mol Biol Evol*. 28(1):273–290.
Rokas A, Holland P. 2000. Rare genomic changes as a tool for phylogenetics. *Trends Ecol Evol*. 15(11):454–459.
Roux J, Liu J, Robinson-Rechavi M. 2017. Selective constraints on coding sequences of nervous system genes are a major determinant of duplicate gene retention in vertebrates. *Mol Biol Evol*. 34(11):2773–2791.
Ruprecht C, Lohaus R, Vanneste K, Mutwil M, Nikoloski Z, Van de Peer Y, Persson S. 2017. Revisiting ancestral polyploidy in plants. *Sci Adv*. 3(7):e1603195.
Sacerdot C, Louis A, Bon C, Berthelot C, Roest Crollius H. 2018. Chromosome evolution at the origin of the ancestral vertebrate genome. *Genome Biol*. 19(1):166.
Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH. 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* 440(7082):341–345.
Scornavacca C, Jacox E, Szöllősi GJ. 2015. Joint amalgamation of most parsimonious reconciled gene trees. *Bioinformatics* 31(6):841–848.
Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. *Syst Biol*. 51(3):492–508.
Shimodaira H, Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17(12):1246–1247.
Singh PP, Affeldt S, Cascone I, Selimoglu R, Camonis J, Isambert H. 2012. On the expansion of "dangerous" gene repertoires by whole-genome duplications in early vertebrates. *Cell Rep*. 2(5):1387–1398.
Singh PP, Arora J, Isambert H. 2015. Identification of ohnolog genes originating from whole genome duplication in early vertebrates, based on synteny comparison across multiple genomes. *PLoS Comput Biol*. 11(7):e1004394.
Singh PP, Isambert H. 2020. OHNOLOGS v2: a comprehensive resource for the genes retained from whole genome duplication in vertebrates. *Nucleic Acids Res*. 48:D724–D730.

Sollars ESA, Harper AL, Kelly LJ, Sambles CM, Ramirez-Gonzalez RH, Swarbreck D, Kaithakottil G, Cooper ED, Uauy C, Havlickova L, et al. 2017. Genome sequence and genetic diversity of European ash trees. *Nature* 541(7636):212–216.

Som A. 2015. Causes, consequences and solutions of phylogenetic incongruence. *Brief Bioinform*. 16(3):536–548.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.

Sukeena JM, Galicia CA, Wilson JD, McGinn T, Boughman JW, Robison BD, Postlethwait JH, Braasch I, Stenkamp DL, Fuerst PG. 2016. Characterization and evolution of the spotted gar retina. *J Exp Zool Mol Dev Evol*. 326(7):403–421.

Szöllősi GJ, Rosikiewicz W, Boussau B, Tannier E, Daubin V. 2013. Efficient exploration of the space of reconciled gene trees. *Syst Biol*. 62(6):901–912.

Szöllősi GJ, Tannier E, Daubin V, Boussau B. 2015. The inference of gene trees with species trees. *Syst Biol*. 64(1):e42–e62.

Van de Peer Y, Maere S, Meyer A. 2010. 2R or not 2R is not the question anymore. *Nat Rev Genet*. 11(2):166–166.

Van de Peer Y, Mizrachi E, Marchal K. 2017. The evolutionary significance of polyploidy. *Nat Rev Genet*. 18(7):411–424.

van Hoek MJA, Hogeweg P. 2009. Metabolic adaptation after whole genome duplication. *Mol Biol Evol*. 26(11):2441–2453.

Varadharajan S, Sandve SR, Gillard GB, Tørresen OK, Mulugeta TD, Hvidsten TR, Lien S, Asbjørn Vøllestad L, 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.

Veitia RA, Bottani S, Birchler JA. 2008. Cellular reactions to gene dosage imbalance: genomic, transcriptomic and proteomic effects. *Trends Genet*. 24(8):390–397.

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(2):327–335.

Wapinski I, Pfeffer A, Friedman N, Regev A. 2007. Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics* 23(13):i549–i558.

Wu Y-C, Rasmussen MD, Bansal MS, Kellis M. 2013. TreeFix: statistically informed gene tree error correction using species trees. *Syst Biol*. 62(1):110–120.

Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E, et al. 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* 21(5):650–659.

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.