

# Genes Encoding Teleost Fish Ligands and Associated Receptors Remained in Duplicate More Frequently than the Rest of the Genome

Anna Grandchamp\*, Benoît Piégu, and Philippe Monget\*

PRC, UMR85, INRA, CNRS, IFCE, Université de Tours, Nouzilly, France

\*Corresponding authors: E-mails: [anna.grandchamp@neuf.fr](mailto:anna.grandchamp@neuf.fr); [philippe.monget@inra.fr](mailto:philippe.monget@inra.fr).

Accepted: April 5, 2019

## Abstract

Signaling through ligand/receptor interactions is a widespread mechanism across all living taxa. During evolution, however, there has been a diversification in multigene families and changes in their interaction patterns. Among the events that led to the creation of new genes is the whole-genome duplication, which made possible some major innovations. Teleost fishes descended from a common ancestor which underwent one such whole-genome duplication.

In our study, we investigated the effect of complete genome duplication on the evolution of ligand–receptor pairs in teleosts. We selected ten teleost species and used bioinformatics programs and phylogenetic tools in order to study the evolution of the human ligands and receptors that have orthologous genes in fishes, as well as the rest of the fish genomes.

We established that since the complete duplication of the fish genomes, the conservation in duplicate copy of ligand and receptor genes is higher than expected. However, the ligand/receptor pair partners did not necessarily evolve in the same way, and a lot of situations occurred in which one of the partners returned in singleton copy when the other one was maintained in duplicate. This suggests that changes in interaction partners may have taken place during the evolution of teleosts. Moreover, the fate of the ligands and receptor coding genes is partly congruent with the phylogeny of teleosts. However, some incongruences can be observed. We suggest that these incongruences are correlated to the environment.

**Key words:** ligand, receptor, phylogeny, whole-genome duplication, duplicate, singleton, coevolution.

## Introduction

Since the appearance of their last common ancestor, some multicellular organisms have grown in complexity and number of genes (Carroll 2001). The origin of these new genes involves different mechanisms (Long et al. 2003). After the appearance of the new genetic material, if the gene is not directly advantageous, it may evolve either by pseudogenization or toward the acquisition of a new utility (Innan and Kondrashov 2010).

Gene duplication may involve a single gene, several genes, or the entire genome. In the latter case, the entire genome of the individual is doubled to the next generation (Ohno et al. 1968; Ramsey and Schemske 2002; Adams and Wendel 2005; Hufton and Panopoulou 2009). The duplication of a complete genome paves the way for important genetic innovations (Olmo 1983), such as heterosis (Butruille and Boiteux 2000), and the appearance of a large amount of new genetic material (Van de Peer et al. 2009). Whole-genome duplication

(WGD) phenomena have been observed in a wide diversity of taxonomic groups: in plants (Adams and Wendel 2005), bacteria (Kuroda et al. 2001), unicellular eukaryotes (Kellis et al. 2004), and vertebrates (Mable et al. 2011).

Among vertebrates, the duplication in teleost fishes is well documented (Christoffels et al. 2004; Jaillon et al. 2004; Meyer and Van de Peer 2005; Brunet et al. 2006; Glasauer and Neuhaus 2014). Complete duplication occurred in the ray finned fishes lineage. The fish clade showing complete duplication, clupeocephala, diverged about 255 Ma (Betancur-R et al. 2017). The ancestral teleost prior to WGD is believed to have had 12 or 13 chromosomes, following its divergence from the lineage of the lobe-finned fishes, whereas current fishes have an average of 24–25 chromosomes (Kohn et al. 2006; Kasahara et al. 2007; Glasauer and Neuhaus 2014). Fishes are a very diverse taxonomic group. Their diversity was already high before the duplication (Martín-Abad and Poyato-Ariza 2013; Khosla and Lucas

2016) but was amplified afterward (Brunet et al. 2006). The high diversity of fish species combined with a recent complete duplication makes Clupeocephala a group of great interest in the study of complete genome duplication in the animal kingdom.

Following complete genome duplication, all genes do not remain in duplicate in the same way. Most models predict a rapid return of some of the duplicates to singleton status (Maere et al. 2005), the deleterious copies being rapidly pseudogenized (Sankoff et al. 2010). In particular for the rainbow trout, whose genome underwent an extra duplication compared with other teleosts about 100 Ma, it is estimated that about 48% of the genome now remains in duplicate, when the remaining 52% quickly returned to the singleton category (Berthelot 2014).

On the contrary, certain gene types are more likely to remain as duplicates in all taxonomic groups studied. This is the case of transcription factors, protein kinases, enzymes, and transporters (Conant and Wolfe 2008). Several reasons have been given to explain the fact that these genes are more often preserved in duplicate (Comai 2005; Byrne and Wolfe 2007; Innan and Kondrashov 2010; Albalat and Cañestro 2016). The increase of protein production can be advantageous, as for enzymes and transcription factors (Riechmann et al. 2000). Doubling the dosage could entail a better fitness of the species (Papp et al. 2003; Conant and Wolfe 2008; Makino and McLysaght 2010). The presence of the duplicate copy can compensate for a deleterious mutation in the first copy, as a functional redundancy (Clark 1994; Lynch et al. 2001; O'Hely 2006). Another reason could be the neofunctionalization of the new gene, allowing the development of a new function (Taylor and Raes 2004; Byrne and Wolfe 2007). Subfunctionalization also occurs, corresponding to the repartition of the function between the two copies. A set of possibilities has been described in detailed reviews (Comai 2005; Innan and Kondrashov 2010).

Among the interacting molecules are ligands and receptors, which are found in all taxonomic groups. Their interactions are the first steps of signal transduction and involve all functions of the organism. A single receptor may have several ligands (e.g., integrins), and these ligands may have multiple receptors. The interactions are characterized by a broad spectrum of more or less specific affinities (Cuatrecasas and Hollenberg 1976; Bongrand 1999). Teleost fishes share many ligands and receptors with humans, whose complete lists of ligand and receptor interactions are available. In addition, recent studies have provided extensive clarifications on fish phylogeny (Near et al. 2012; Betancur-R et al. 2017).

Understanding which evolutionary factors and strengths allow certain genes to remain in duplicates, whereas others return to singleton status is a real challenge. In our study, we built a list of ligands and receptors shared by ten fish species and orthologous to humans. We studied the evolution of these genes in order to understand how the WGD event

that happened in the ancestor of the teleost fishes affected the evolution of the ligands and their receptors.

## Materials and Methods

We studied ten species of fish: Amazon molly (*Poecilia formosa*), cave fish (*Astyanax mexicanus*), cod (*Gadus morhua*), fugu (*Takifugu rubripes*), medaka (*Oryzias latipes*), platy fish (*Xiphophorus maculatus*), stickleback (*Gasterosteus aculeatus*), tetraodon (*Tetraodon nigroviridis*), tilapia (*Oreochromis niloticus*), and zebrafish (*Danio rerio*). These ten fish species were the only ones available on the Ensembl site at the time of our experiment and are the first entirely sequenced fishes, with the best coverage. These species diverged after the complete duplication of the teleost genome.

There is no list of fish genes coding for ligand/membrane receptors. The only complete list of ligands/membrane receptors in vertebrates available is a list of human genes. This ligand/receptor list was retrieved from Ramiłowski et al. (2015). It contains 2,500 redundant human ligand and receptor interactions, corresponding to genes encoding 706 different ligands and 691 different receptors. We developed a methodology that allowed us to retrieve the fish orthologs of all these human genes.

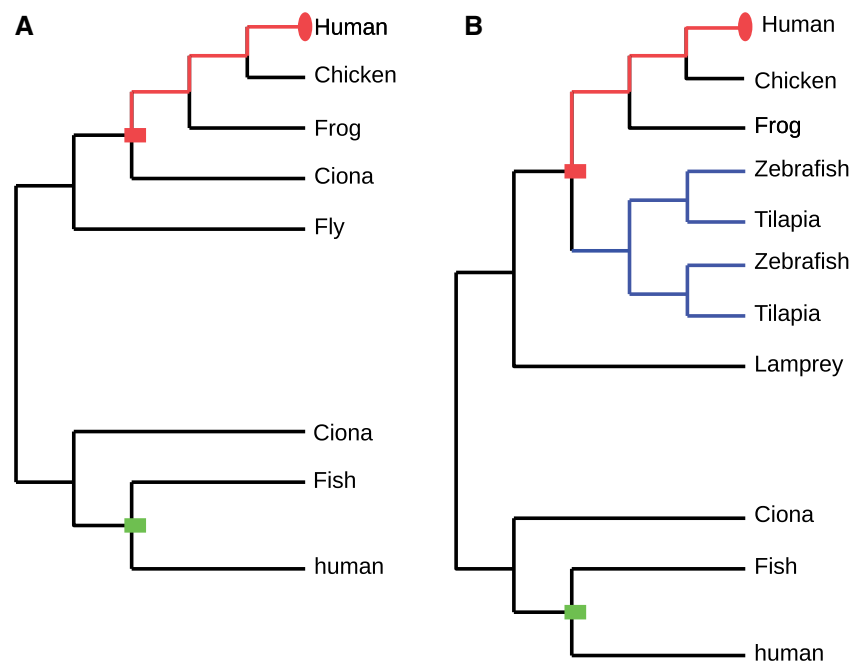
The methodology we used is the following:

1. The 19,888 phylogenetic trees corresponding to the 19,888 human genes present in Ensembl were retrieved.
2. The trees were modified so as to better showcase duplications.
3. The orthologs of each human gene were recovered in each of the ten fish species studied.
4. Among all human orthologs genes present in teleosts, teleost ligands and receptors were identified and analyses were then conducted.

This approach eliminates the ligands and receptors that appeared in fishes and are not existing in humans, but by considering human orthologs we got the biggest lists of fishes ligand receptors interactions.

### 1 and 2: Extraction and Modification of Phylogenetic Trees

Phylogenetic trees were extracted from Ensembl release 82. The phylogenetic trees of Ensembl are based on phylogenetic methods that trace the evolution of a gene family, to find orthology and paralogy relationships. In addition, they indicate the presence of duplications. However, these trees were edited according to a methodology used for the construction of the Genomicus database (Louis et al. 2015). This method provides a better stringency of duplication nodes and ensures that genes that evolve very quickly are taken into account. Duplication nodes with a duplication consistency score (Vilella et al. 2009) below a threshold (here 0.30) were selected and shifted toward terminal branches, unless they were stopped



**FIG. 1.**—Description of the algorithm. (A) Case in which the human gene has no ortholog in fish. The program followed the red line. Once the human gene was found in the tree, the program moved from node to node. At the node of the chordates, no teleosts were encountered. In this example, there were fish paralogous genes in the tree (green square), but they did not correspond to the orthologous gene of the studied human gene. (B) Case in which the human gene is orthologous to the common ancestor of all genes from clupeocephala (purple branch). At the node of euteleostomii (red square) are fish genes that descended from this root. The algorithm retrieved the clupeocephala branch from this node (purple), and the information on duplication/singleton status was noted for the each species.

by a strong intermediate duplication node. The phylogenetic trees corresponding to the 19,888 human genes present in Ensembl were recovered and transformed to Newick format.

### 3: Algorithm

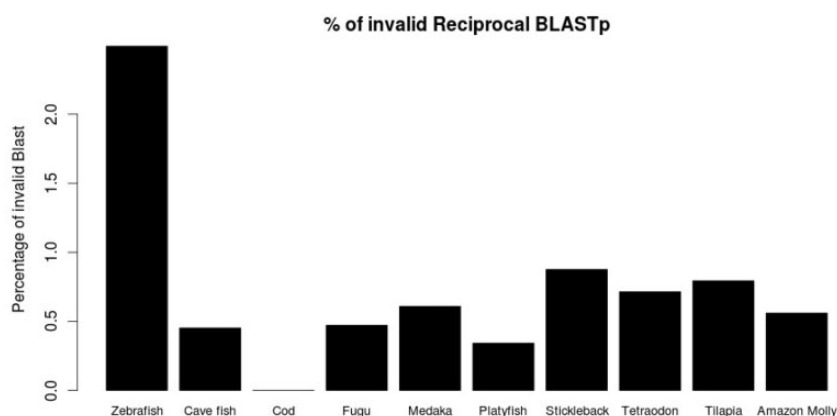
The phylogenetic trees corresponding to each of the human genes were identified. Once the human gene was located in the tree, the algorithm looked for the first encountered node in the branch (fig. 1). For each newly encountered node, the algorithm tested two conditions. First, the encountered node must be more ancestral than *euteleostomii* without crossing a teleost branch. If this condition was respected, the program stopped (fig. 1a), because it meant that this particular human gene did not have an ortholog in fishes. Second, if the algorithm encountered a node including a teleost branch, the program left the loop and retrieved this branch (fig. 1b). In this teleost branch, for each of the ten fish species, the algorithm then determined whether the orthologous gene was present in a single copy in the fish species, in several copies, or not at all. If none of the two conditions were fulfilled, the algorithm moved on to the next encountered node. For each human gene with one or several orthologous genes in fishes, the ortholog singleton or duplicates were recovered in each fish. Among all of these genes, the information regarding the ligand and receptors was specifically retrieved for each fish.

The information concerning the entire collection of orthologs between fishes and human was saved, in order to later compare the evolution of ligand receptors with that of the human orthologs in fishes, which is more representative of the fish's genomic evolution. A Github repository (Ram 2013; Wilson et al. 2014) is available with the algorithm's source code ([https://github.com/AnnaGrBio/ortho\\_fishing](https://github.com/AnnaGrBio/ortho_fishing)).

Some of the trees modified contain branch mismatches. These trees were treated by the algorithm. In the case where the clupeocephala branch clustered with the lamprey branch for example, the "Clupeocephala group" was not apparent in the branch. In any case, the algorithm stopped at the subgroup below the fish and lamprey, and the vertebrate tree was analyzed. If it contained clupeocephala fishes, then the subtree was built, and the orthologs were recovered. The cases where clupeocephala fishes were not monophyletic were also treated.

However, trees containing the artifacts mentioned above ran the risk of being unreliable, and our results may have been biased by the presence of genes that did not correspond to the actual human orthologs. To avoid this bias, we used a reciprocal protein BLAST to test whether the receptor ligands found in the trees truly corresponded to the orthologs of human genes.

Sequences of all proteins of the human genome and of each fish species' were downloaded from Ensembl 82. To



**FIG. 2.**—Percentage of reciprocal best-hits (protein) BLAST searches that did not work for all 10 fish. It was done only on the ligands and receptors genes.

ensure the stringency of the Ensembl trees, a reciprocal best-hits protein BLAST (Wall et al. 2003) was performed for each human genes encoding a ligand or receptor on the entire fish genome. The BLAST analysis was done with the default settings, because we wanted to be able to detect the duplicate copies that were modified by evolution. Each human gene was BLASTed against all fish genomes. If the ortholog (or orthologs in case of duplicate copies) found in the tree was present in the outputs of the BLAST, the gene was recovered and BLASTed against the human genome. If the good human gene was detected in the outputs of the BLAST, we considered that the BLAST was in keeping with the results of the modified trees. We restricted our focus to the ligands and receptors for which almost all of the proteins were present. An orthologous relationship was validated only if the blast of human gene was the best match for the orthologs in fish, and reciprocally. This methodology made it possible to validate all human/fish orthologous gene relationships on the ligand/receptor list (fig. 2).

#### 4: Analysis

For each species, we studied the evolution (retention vs. singleton) of each human gene that had an ortholog in fishes. To this end, we retrieved the information of the evolution of each of the 19,888 human genes that have an ortholog in fish species. The orthologs of each fish are listed in [Supplementary Data Sheets 1–10, Supplementary Material](#) online. We obtained an average of 12,895 human genes presenting an ortholog in fish species. This does not represent the entire genome of each fish but allowed us to make strong statistical predictions. For each fish, the global evolution the whole human orthologs was compared with the specific evolution of ligands and receptors. We studied whether the ligands and the receptors remained as a duplicate copy or had returned to singleton in the same proportion as human ortholog genes that are not ligands and receptors.

The evolution of the genes coding for the ligands and for their receptors among the ten species of fish was also compared. However, comparing fishes is more awkward than analyzing each fish species independently. In fact, the coverage of the sequencing is very unequal for all ten species ([supplementary data sheet 11, Supplementary Material](#) online). The zebrafish is the best sequenced, whereas fishes like cod and fugu have a low coverage. For example, the zebrafish and cave fish have the longest golden patches, whereas other fishes like the stickleback have very small ones. Similarly, the size of assemblies, L50 scaffolds and L50 contigs, is very uneven, the values corresponding to these variables being very low for Amazon Molly, platyfish, and Tilapia for example. Finally, some data such as std value or c value are missing for some species. If a gene is present in a duplicate copy in most of the fish species and present in a singleton copy—or even not present at all—in the few remaining fish types, it could mean that the gene specifically returned to singleton in these two species, or more likely, that the second copy of the gene was not sequenced or annotated.

To overcome this possible bias, we arbitrarily defined an 80% threshold for species sharing the same information (deletion, singleton, or duplicate), for each gene. If a gene coding for a ligand or receptor showed the same evolution in 80% of the fishes or more, we considered that the gene showed the same evolution in all of the fishes. Setting a threshold at 80% allows us to take into account genes that are absent in certain species, because this absence is possibly due to a lack of sequencing coverage (the data are available in [supplementary data sheet 12, Supplementary Material](#) online). Otherwise, we considered that the evolutionary history of this gene was questionable, due to the possible absence of sequence or annotation of gene copies in certain species. These data allowed us to compare the evolution of the combinations of pairs of ligands and receptors.

A matrix was developed in order to assess the percentage of genes coding for ligands and receptors whose evolution was the same among the fishes and to assess whether the

**Table 1**

Common versus Uncommon Evolution of Ligands and Receptors

	Amazon Molly	Cave Fish	Cod	Fugu	Medaka	Platyfish	Stickleback	Tetraodon	Tilapia	Zebrafish	Average by Fish	Comon Fate
Rd	233	213	175	187	180	210	184	187	210	215	199	161
Ld	173	177	140	142	129	233	140	144	163	195	157	121
Rs	279	296	293	286	288	233	299	289	292	301	291	236
Ls	245	236	244	259	251	233	247	255	253	237	248	206
Rnp	—	—	—	—	—	—	—	—	—	—	—	19
Lnp	—	—	—	—	—	—	—	—	—	—	—	14
R mixed	—	—	—	—	—	—	—	—	—	—	—	71
L mixed	—	—	—	—	—	—	—	—	—	—	—	61

NOTE.—The first ten columns represent the ten species of fish. The 11th column represents the average number of genes of each category between each species of fish. The 12th column represents the number of genes that have undergone the same evolution in all the species of fish. The line labeled Rd represents the receptors that are duplicated. The number of duplicated genes is given for each species of fish, as well as the number of genes duplicated in all the fish species. The same is shown for the duplicated ligand (Ld), the singleton receptors (Rs), and the singleton ligands (Ls). Rnp and Lnp represent the number of genes that are not present in only one or two species of fish. The R mixed is the number of receptors that have a distinct evolution in the ten species of fish. The L mixed is the number of ligands that have a distinct evolution in the ten species of fish.

evolution of ligands receptors—as duplicate or singleton—was congruent with the tree of life of the teleosts. For each gene present in two species, we assigned one point to both species if the gene had the same evolution and 0 if the genes had evolved in different ways. A dendrogram of distance was calculated, based on the scores between species, using the heatmap function of R (ggplot2 and ggcorrplot libraries).

A more thorough study was conducted on the ligands and receptors of the zebrafish, in order to understand if some of their characteristics were correlated with their evolution. The list of ligands and receptors of Ramilowski et al. (2015) is available on [supplementary data sheet 13, Supplementary Material](#) online, as well as the Ensembl accession numbers of genes on [supplementary data sheet 14, Supplementary Material](#) online.

## Results

### Recovery of Human Gene Orthologs in the Ten Fish Species

A program was built in order to recover the fish orthologs of human genes in the ten studied species of fish. A reciprocal best-hits protein BLAST (Wall et al. 2003) was performed for each human gene encoding a ligand or receptor, in order to confirm the accuracy of the Ensembl trees. We observed a very low rate of invalidation of the reciprocal best-hits BLAST searches. We get between 0% and 0.8% of invalidated genes, with the exception of the zebrafish for which 2.5% of the genes were not confirmed by the reciprocal best-hits BLAST (fig. 2). We considered that the algorithm we set working on the modified tree did not retrieve erroneous information.

An average of 12,895 fish orthologs were detected per species, with 10,751 showing common evolution in 80% of the fish species: 2,476 found in a duplicate copy and 8,275 in singleton. We found an average of 413 orthologous ligands in fishes, with 121 ligands duplicated in 80% of fish species, and 206 returned in singleton (table 1). We found an average of 521 ortholog receptors in fishes, with

161 duplicated in 80% of the species of fish, and 236 receptors returned in singleton.

### Comparison of Genes Encoding Pairs of Ligand/Receptor with Other Genes

Using the same threshold of 80% of species sharing the same information, that we defined before, we determined what proportion of ligands and receptors was maintained in a duplicate copy in comparison to the whole human gene orthologs.

Surprisingly, we discovered that the genes coding for the ligands, as well as for the receptors, were more likely to be retained in a duplicate copy than the other genes. In fact, we found that 36.89% of the ligands and 40.5% of the receptors were present in duplicate in the fish species (fig. 3). When considering the whole human ortholog, excluding ligands and receptors, we found that 23.03% of the genes were maintained in duplicate in the fish species, 22.59% excluding ligands and 22.65 excluding receptors. We tested whether the number of ligands and receptors in a duplicate copy was higher than the expected number considering the whole human ortholog. We found that the ligands were more likely than expected to be conserved in duplicate copies ( $\chi^2$  test,  $P$  value 0.02732), as were the receptors ( $\chi^2$  test,  $P$  value 0.002367).

### Comparison within Species

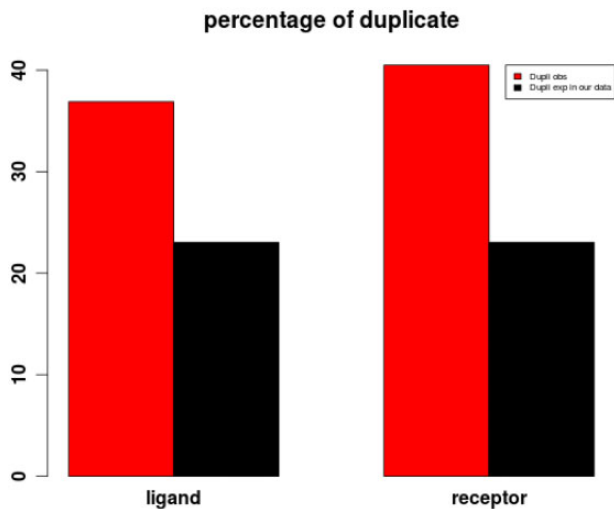
The same results were found when we looked for the genes evolution in each fish species (fig. 4). In fact, in each of the ten fish species, a mean of 40% of the ligands and receptors were retained in duplicate copies, against only 30% of the entire set of human orthologs that were still in a duplicate copy. The  $\chi^2$  test was significant or tended to be significant, except for the ligands in tetraodon (table 2). However, as mentioned above, these data have to be discussed carefully, given that the coverage of sequencing is very different from one fish species to another.



### Dendrogram of Similarity

A matrix was built in order to assess the similarity of evolution of the ligands and receptors between the ten fish species. For each gene present in two species, we attributed one point to these two species if the gene had the same evolution and 0 if the genes had evolved in different ways. A dendrogram was

calculated based on the matrix of similarity between species, using R packages ggplot2 and heatmap (supplementary data sheet 15, Supplementary Material online). The dendrogram showed a similarity with fish phylogeny (fig. 5). In fact, the differences in singleton returns and retention in duplicates of ligands and receptors seem to be species specific and appear to have occurred gradually during evolution, as expected.

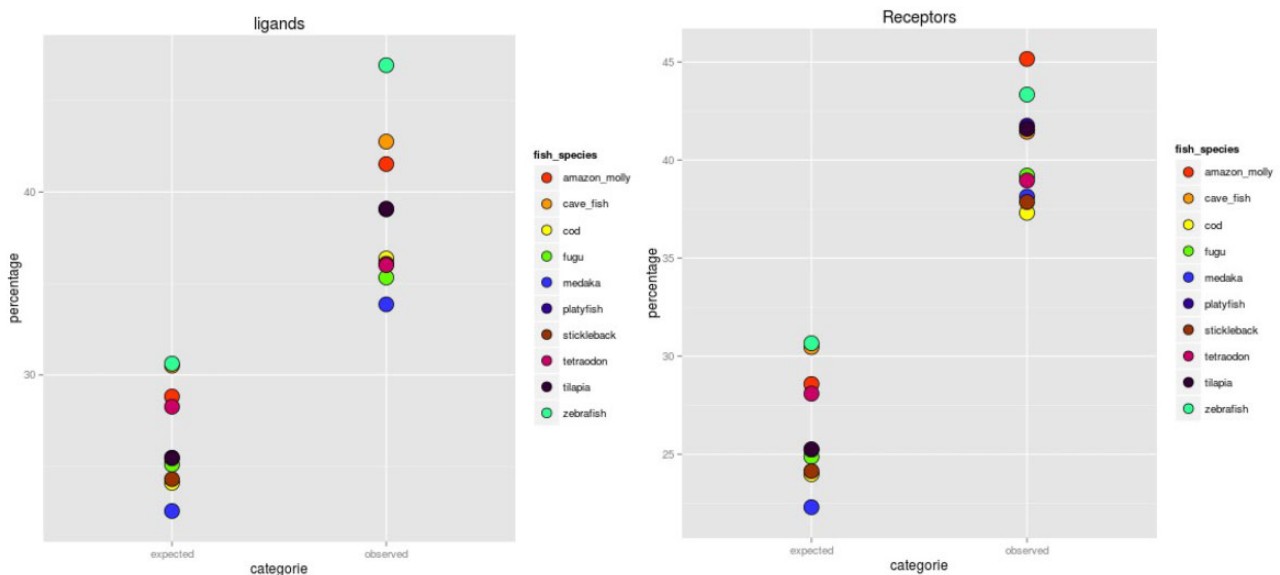


**Fig. 3.**—Percentages of genes found in duplicate, common to the ten fish species. The graph on the left represents the percentage of receptors. The red bar represents the percentage of ligands and receptors that were retained in a duplicate copy in 80% of the fish species. The black bar represents the percentage of human orthologs that were retained in a duplicate copy in 80% of the fish species.

### Gene Network

We tried to determine whether the evolution of one of the members of the ligand/receptor pairs was correlated with the evolution of the second member. For this analysis, we took into account all of the interactions present in our list. For example, if a receptor binds several ligands, its interaction with each ligand was assessed. Four possible situations were defined for each of the interactions: RSLs, the receptor is in singleton and its ligand is in singleton; RdLd, the receptor is in duplicate and its ligand is in duplicate; RSLd, the receptor is in singleton and its ligand is in duplicate; and RdLs, the receptor is in duplicate and its ligand is in singleton (table 3).

For each category, we compared the observed proportions within the ligand/receptor pairs with the expected proportions in the whole genome, using a  $\chi^2$  test. This analysis did not detect any particular distribution, meaning that the interaction of both partners should not have influenced their evolution. The number of interactions in which both members returned to singletons was the highest, which was expected, because the number of ligands and receptors that returned in singleton was higher than the number retained in duplicate.

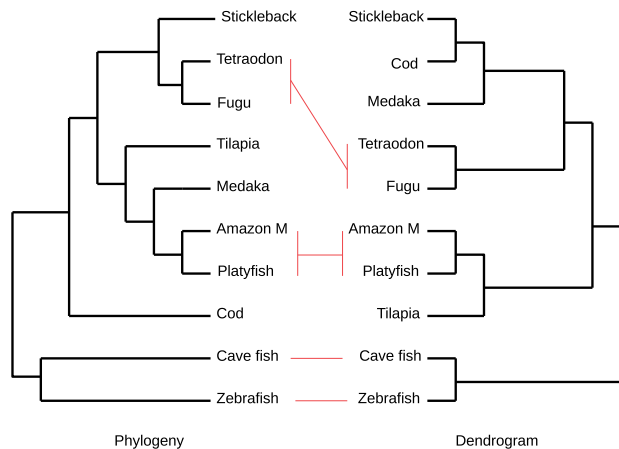


**Fig. 4.**—Percentages of genes found in duplicate according to the fish species. The graph on the left represents the percentage of receptors. The graph on the right represents the percentage of ligands. In each graph, all of the species of fish are represented by a color. The category “expected” represents the percentage of all of the human genes with an ortholog in the considered fish species (between 15 and 16,000 according to the fish species) which were maintained in duplicate. The category observed represents the percentage of ligands (graph on the left) and receptors (graph on the right) that were maintained in duplicate.

**Table 2**  
 $\chi^2$  Tests of the Proportion of Duplicate Ligands and Receptors

	Amazon Molly	Cave Fish	Cod	Fugu	Medaka	Platyfish	Stickleback	Tetraodon	Tilapia	Zebrafish
R P value	0.002309	0.001013	0.0111	0.00695	0.002126	0.001656	0.01656	0.04726	0.001801	0.0165
L P value	0.003407	0.04728	0.03517	0.0750	0.04663	0.001756	0.05874	0.2007	0.01729	0.04876

NOTE.—The columns represent the ten species of fish. The R P value line represents the P value of the  $\chi^2$  test on the receptors. The  $\chi^2$  H0 is the hypothesis that the number of receptors retained in duplicate copy is the same as the expected number taking into account all human orthologs. The L P value line represents the P value of the  $\chi^2$  test on the ligands.



**Fig. 5.**—Comparison of the phylogenetic tree of fish species (left) and the dendrogram resulting from the similarity of the evolution of the genes coding for pairs of ligand/receptor between each fish. The similarity between the two trees suggests that the differences in singleton returns and retention in duplicates of ligands and receptors seem to be species specific and appear to have occurred gradually during evolution.

### Ligand/Receptors Characteristics

Because the receptors could bind several ligands (e.g., FZD1 binds 6 ligands) and the ligands could bind multiple receptors (e.g., WNT5A binds 12 receptors in human), we also investigated whether all partners of a given molecule evolved in the same way.

We found no differences between all of the possible situations. For example, the receptors that had only one ligand may have evolved in the same way as their ligand, or in contradiction to it. Similarly, we studied whether the number of partners had an influence on the way these partners evolved. Here, again, we reviewed all of the possible evolutionary scenarios. For example, for a receptor with several ligands, there were cases in which all of the ligands evolved in the same way as their receptor, others in which the ligands evolved in an opposite way, and we found a range of intermediate possibilities without significant differences. The seemingly random categories are depicted in figure 6.

### Deepening of Receptor Ligand Couples

We tried to establish whether or not the family of the ligands and receptors had an impact on the fact that they returned in

a singleton copy or were maintained in duplicate. A study was conducted in zebrafish because it has the best annotated genome. Data are in [supplementary data sheet 16–19, Supplementary Material](#) online.

We found 226 families of receptors, 72 of which had several members present in zebrafish. We observed that in 20 small families (2–4 members), all of the genes evolved in the same way. For example, all four members of the Vascular Endothelial Growth Factor (VEGF) receptor family present in zebrafish returned in singleton. We found 214 families of ligands, 55 of which had several members present in zebrafish. We also found 20 small families (2–5 members) of ligands in which all of the molecules evolved in the same way. For example, the five ligands of the R-spondin family present in zebrafish returned in singleton. However, concerning or ligands belonging to any of the other families with several members (52 families of receptors and 35 families of ligands), which include the largest ones (e.g., 20 molecules in the integrin family), our results did not show any specific rule. Some of the members of these multigene families returned in singleton, and the others remained in duplicate. We did not observe any rule concerning the superfamilies, in which, for the ligands as well as for the receptors, one part of the genes remained in singleton, and the other remained in duplicate. Similarly, we could not establish any rule by looking at the nature of the ligands (hormones, neuropeptides, or anchored in the membrane), which had no impact on whether they returned to singleton or remained in duplicate.

## Discussion

### Trends Observed in Fishes

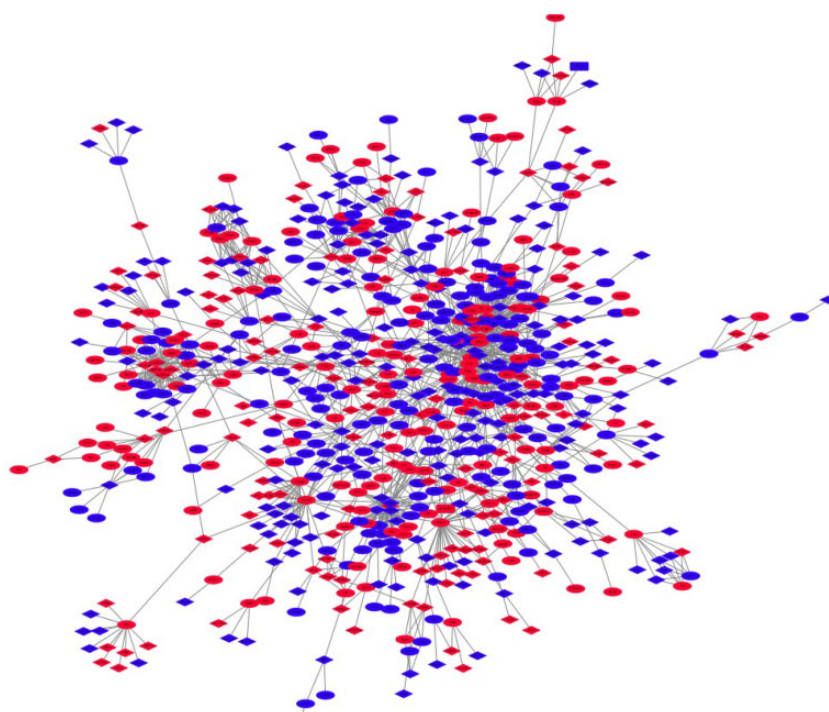
We were able to observe, for the first time, that the genes encoding teleost fish ligands and receptors remained in duplicate more frequently than other genes of the genome. In fact, using a threshold of 80% of studied fish species to validate a duplicate in the ten species studied, we observed that the number of receptors and ligands that were retained in duplicate in fish species was higher than expected, considering the entire set of human genome orthologs. Using our threshold, we found that 23% of the 10,751 human genes that had an ortholog in fish were retained in a duplicate copy, against 40.5% of the receptors and 36% of the ligands. Our resulting percentage of human orthologs maintained in

**Table 3**

Comparison of the Distribution of Interactions

O/E	Amazon Molly		Cave Fish		Cod		Fugu		Medaka		Platyfish		Stickleback		Tetraodon		Tilapia		Zebrafish	
	O	E	O	E	O	E	O	E	O	E	O	E	O	E	O	E	O	E	O	E
RsLS	436	725	457	659	422	684	460	717	446	820	449	759	449	727	479	676	432	758	462	799
RdLd	334	124	353	132	206	72	230	83	171	72	287	93	231	78	216	108	277	93	382	163
RsLd	312	299	380	295	311	222	288	244	251	244	294	265	306	238	273	270	310	265	499	361
RdLs	365	299	290	295	260	222	311	244	324	444	353	265	295	238	355	270	361	265	341	361
RsLS <i>P</i> value	0.5399		0.9307		0.2		0.307		0.009266		0.359		0.1863		0.4208		0.1834		0.3355	
RdLd <i>P</i> value	0.07909		0.8346		0.1272		0.08249		0.8465		0.06326		0.05741		0.3144		0.1084		0.3246	
RsLd <i>P</i> value	0.6752		0.3939		0.2961		0.7844		0.2729		0.6		0.6257		0.6386		0.931		0.2579	
RdLs <i>P</i> value	0.6979		0.3146		0.4917		0.6988		0.547		1		0.8616		0.5306		0.8023		0.2452	

NOTE.—The columns represent the ten species of fish. The O/E lines represents the observed versus expected number of interactions. RsLS represents the number of interactions per fish species, in which the ligands and the receptors both returned to singleton. The expected proportion is calculated according to the proportion of ligands and receptors that were retained in duplicate copies or went back to singleton. Then, for each category, the expected proportion was calculated in this way: (total number of genes)  $\times$  (proportion of genes expected in the receptor category)  $\times$  (proportion of genes expected in the receptor category). Also, for the RdLd category, the calculation was total: (number of genes)  $\times$  (% of duplicated genes)  $\times$  (% of genes in singleton). RdLd represents the number of interactions per fish species, in which both ligands and receptors were maintained in a duplicate copy. RsLd is the number of interactions for which the receptors are in singleton and the ligands in duplicate. RdLs is the opposite case. The global *P* value is the *P* value of the  $\chi^2$  test that compared the output observed and the one expected for each category. All four categories are then compared using a  $\chi^2$  test.



**FIG. 6.**—Representation of a part of the zebrafish gene network of the ligands and receptors, for proteins that are connected with more than one partner. The ovals represent the ligands. The rhombi represent the receptors. The objects in red are the molecules present in duplicate. The objects in blue are the molecules present in singleton. We observed all of the cases: several ligands for one receptor and one ligand for several receptors, regardless of whether they were in singleton or duplicate, and regardless of the gene biological network.

duplicate is similar to other percentages observed in literature, that predicted between 15% and 25% of genes retained in duplicate (Ravi and Venkatesh 2008). Moreover, these high proportions of ligands and receptors kept in duplicate were also observed in every fish species. For each of the ten species of fish, the percentage of genes coding for ligands and receptors that remained in duplicate was close to 40%, whereas

the average number of genes that were retained in a duplicate copy, in the entire genome, was about 30%. Here, the percentage of human orthologs maintained in duplicate (30%) was higher than the percentage documented in literature. However, this can be explained by the fact that each fish was studied independently, and specific duplications of each species were also taken into account. We compared our



data with those of Pasquier et al. (2017), who estimated that about 15% of the spotted gar genes remained in duplicate in zebrafish. However, these authors based their study on the genes of the spotted gar genome, a fish genome that has not duplicated, compared with *Clupeocephala*, whereas we have based our study on the human genome. Moreover, the authors stated that their methodology undoubtedly underestimated the number of genes remaining in duplicate.

It was not the first time that the retention of genes coding for receptors was observed in high frequency after a duplication event. For the endothelin receptors, it was shown that ligands and receptors were duplicated in teleost fishes, and a coevolution between ligands and receptors was observed (Braasch et al. 2009). That was also the case of secretin receptors, where the second receptor is in retention in all fish species except the *Tilapia* (Cardoso et al. 2006). More globally, in 2006, a study was conducted on 9,461 gene families of 7 vertebrate species, in order to investigate the retention of the vertebrate genes. The authors highlighted higher gene retention of the genes coding for receptors involved in signal transduction (Blomme et al. 2006).

One hypothesis to explain such a retention concerns the dosing balance of the corresponding proteins. In addition to their functional importance in all living organisms, the duplication of each of the partners makes it possible to double the interactions without modifying the dosage of one of the partners.

However, our results do not support the theory of the balance of dosage in the case of our ligand/receptor pairs. Indeed, if such was the case, we would have expected that the ligands whose receptors returned in singleton would also more frequently return in singleton, and that the duplicated receptors would have their ligands duplicated as well. In contrast, our results showed that when a receptor returned to a singleton, its ligands returned to a singleton as often as they remained in duplicate. Moreover, in the case in which one of the molecules, ligand or receptor, had several partners, it would either return to singleton or remain in duplicate. As a consequence, the network of interactions was very heterogeneous (fig. 6). Because we did not observe any correlation between the evolution of the receptors and their ligands, it could mean that the dosage balance (Veitia 2004) was not the only explanation for the retention of these molecules.

Another hypothesis could be that ligands and receptors evolve more slowly than other genes, leading to the fact that they are more likely to remain in duplicate than the rest of the genome. However, several arguments seem to suggest that this is not the case. For example, as depicted in [supplementary data, Supplementary Material](#) online, the genes encoding ligands and receptors involved in immunity, that are known to evolve quickly (Schlesinger et al. 2014), were maintained in duplicate at the same rates as the other ones. More precisely, all of the genes coding for chemokine receptors were maintained in duplicate, as well as 40% and

33% of the genes respectively encoding tumor necrosis factors and interleukin receptors, ([supplementary data](#) p. 16 and 17, [Supplementary Material](#) online). Indeed, the retention rate for these ligands and receptors that evolve quickly is higher than for the rest of the genome (average of 23.03% retention).

The genes showing the slowest evolution are genes with high levels of expression (Pál et al. 2001) and involved in the most vital functions (Hurst and Smith 1999). In our study, the genes coding for ligands and receptors with the highest levels of expression—like the neuropeptides—or involved in the most vital functions—like the genes of the metabolism (e.g., Serpin family, all returned in singleton), the growth factors (e.g., VEGF receptor family and Rspodin family, all returned in singleton) or even the genes involved in embryogenesis (WNT family, 40% duplicate)—are equally as likely to be retained than the other ligands and receptors, and sometimes even less.

Finally, the size of the genes also plays a role in the speed of their evolution. More particularly, small genes have a higher probability of evolving more rapidly—which also affects their phylogenetic signals (Grandchamp and Monget 2018)—and ligands are often smaller than their receptors. However, we show here that both ligands and receptors show a higher tendency to remain in duplicate than the rest of the genome, thus invalidating the impact of molecular weight.

These arguments suggest that ligands and receptors, following complete genome duplication, are subject to the same evolutionary constraints as all other genes.

If dosage balance was not a main driver in explaining the evolution of genes encoding ligands and receptors after genome duplication in teleosts, and barring a bias due to slow evolution, other hypotheses can be proposed. For example, the fact that the number of ligands is doubled for a receptor returned in singleton, and vice versa, could be advantageous. Such cases are common in multigene families. The fibroblast growth factor family and their Fibroblast Growth Factor (FGF) receptors have been amplified by duplication events (Itoh and Ornitz 2004). In the human genome, in which 22 genes encode Fgf ligands and 4 genes encode receptors, ligands were first amplified in the first metazoans. Subsequently, acquisition of additional receptors from one original was made later in vertebrates (Itoh and Ornitz 2004). Moreover, it has already been shown that an increase in the number of only one of the partners could be beneficial. This was the case, for example, for the NKG2D receptor involved in innate immunity, which has many more ligands than other members of its family. Several reasons have been cited for such a variety of ligands, including evolution driven by disease-induced selection pressure (Eagle and Trowsdale 2007).

One hypothesis to explain why one of the partners remained in duplicate is that the duplicated molecules acquired a new function. It has been shown that the rate of evolution was higher in teleost fishes (Ravi and Venkatesh

2008). Such a quick evolution allows the phenomena of neofunctionalization. It was the case, for example, of the genes coding for pigmentation in fishes. It has been shown that pigmentation genes were most likely to be retained in duplicate in fishes, allowing new pigments in the different species (Braasch et al. 2009). In particular, the acquisition of a new function can be correlated with a change of partner, or a change of affinity with the partner, and it may take some time after the duplication event for the new functions to be acquired. This is the case, for example, of RAR receptors, whose common ancestor in chordates is similar to the RARbeta mammalian isoform. Other alpha and gamma isoforms have evolved in vertebrate species. All RAR still bind their common ligand, ATRA. However, the differences between their binding pocket give rise to differential sensitivity to pharmacological agonists (Escriva et al. 2006). Similarly, the acquisition of new ligands has been reported for many receptors in multigene families, for example, LXR mammalian receptors. LXR mammalian receptors bind to several specific oxysterol ligands, whereas the ancestral receptor only bound to few chemical compounds (androstane and pregnane steroids oxysterols) (Reschly et al. 2008). In a recent study on nuclear receptors, it has been proposed that the first nuclear receptors to appear bound to chemical compounds, and that the acquisition of specific oxysterol ligands took place during evolution, after several events of duplication and neofunctionalization of the receptors (Markov and Laudet 2011).

Surprisingly, some receptors have been shown to have lost their ligand binding and play a role without a ligand. This is the case for example of RXR-USP (Iwema et al. 2007). Such evolution would explain why different species of fish did not retain the same receptors in duplicate copies.

### Differences between Fishes

Our results showed that most of the ligands and receptors in our set evolved in the same way in the different species of fish. In fact, we found that a mean of 236 receptors and 206 ligands have returned to singleton in all of the species of fish, each species having an average of 291 receptors and 248 ligands in singleton. In the same way, we found that 161 receptors were in duplicate in each fish, with an average of 199 receptors in duplicate per species of fish, and that 121 ligands were in duplicate, with an average of 156 by species of fish.

We found that most of the genes that returned to singleton status are shared by all the fishes. Such a phenomenon could be explained by the fact that the complete genome duplication in the common ancestor of *clupeocephala* was followed by a first return to singleton before the divergence of the fish species occurred (Ravi and Venkatesh 2008).

We also observed that 71 receptors and 61 ligands seem to have evolved independently in the 10 species of fish. These specificities in species could be due to different ecological

constraints (Van de Peer et al. 2017). For example, it has been demonstrated that the genetic variability of plants could increase their tolerance to the largest ecological ranks (Hahn et al. 2012; Te Beest et al. 2012). It was suggested that the polyploidization in animals was correlated with periods of climatic changes and instability in the environment (Mable et al. 2011). Nonetheless, the differences between species should be considered carefully. All ten species of fish studied were not sequenced with the same coverage (see [supplementary data sheet 11, Supplementary Material](#) online). The genes in singleton copies, regardless of whether or not they can be found in a duplicate copy in other species, could be missing from the annotated databases but actually be present in duplicate in all the species.

Surprisingly, the dendrogram that we built based on the similarity of evolution between fish species showed some interesting similarities with the phylogeny. In fact, we observed that cave fish and zebrafish were paired, and mostly differed from the other species. The zebrafish diverged from the other branches 255 Ma, and the cave fish, 150 Ma. Both branches were the first to diverge from the other species of fish. The Amazon molly and the platy fish were also grouped together in the dendrogram, as well as the tetraodon and fugu. These two groups of species are the last branches of species to have diverged, out of the ten species. In fact, tetraodon and fugu diverged about 70–50 Ma, and Amazon molly and platyfish diverged about 50 Ma. The convergence between the phylogeny and the evolution of ligands and receptors can attest to a most recent evolution of the duplicated ligands and receptors. However, this convergence could also have arisen from the difference in sequencing techniques between the ten species of fish. In fact, the zebrafish and the cave fish are the best sequenced species. Their similarity of evolution could simply result from a better identification of the evolution of the genes. Following a complete duplication of the genome, several models predicted a fast return to the state of singleton of some of the genes. During evolution, other genes were, however, pseudogenized in a more progressive way. The fact that the dendrogram was partly similar to phylogeny may indicate that ligands and receptors have been delayed in their return to singleton state, in specific species. Nevertheless, the dendrogram did not completely follow the phylogeny of fishes. In the dendrogram, there was a grouping between stickleback, cod, and medaka, which was not congruent with the phylogenetic tree. Similarly, Amazon molly and platy fish were found to be close to tilapia, which is a distant species in phylogeny. In a very interesting way, these groupings are coherent from a point of view of the ecology of these species. Indeed, the three stickleback, cod, and medaka species correspond to species that live in the northern hemisphere. All can live in the marine environment, although stickleback and medaka species can also live in freshwater. In contrast, the Amazon molly platy fish and tilapia are tropical ecosystem fish, which live in fresh water (tilapia could live in salt water;

nevertheless the sequenced tilapia is from the Nile River). It could also be, if confirmed by further studies at a more detailed phylogenetic scale, that the evolution of their ligands and receptors is, in this case, correlated with their way of life rather than their evolutionary history.

It is important to note, however, that species sequencing coverage is very uneven. As a matter of fact, all the results observed species by species have to be carefully analyzed. Any discussion on the matter should be seen as suggestions and not be taken as a conclusion or a result.

## Conclusion

In our study, we demonstrated that the genes coding for ligands and receptors were retained in duplicate more systematically than the rest of the genome in fishes, and independently among different species of fish. Some specific studies *in vivo/in vitro* would now be necessary to understand the strength that operated on these genes, and the fate of the duplicated copies.

## Supplementary Material

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

## Acknowledgments

The authors are grateful to Dr Julien Bobe and Yann Guiguen for helpful discussions. Furthermore, the authors are very grateful to Hugues Roest Crollius and Alexandra Louis for their phylogenetic trees. This work was supported by the French Ministry of Research. The present study was supported by a fellowship from the French Ministry of Research and by the Institut National de la Recherche Agronomique. We are funded by INRA (Institut national de la recherche agronomique) and French ministry of research. So we also should thank INRA and French ministry of research.

## Literature Cited

- Adams KL, Wendel JF. 2005. Polyploidy and genome evolution in plants. *Curr Opin Plant Biol.* 8(2):135–141.
- Albalat R, Cañestro C. 2016. Evolution by gene loss. *Nat Rev Genet.* 17(7):379.
- Berthelot C. 2014. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat Commun.* 5:3657.
- Betancur-R R, et al. 2017. Phylogenetic classification of bony fishes. *BMC Evol Biol.* 17(1):162.
- Blomme T, et al. 2006. The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol.* 7(5):R43.
- Bongrand P. 1999. Ligand–receptor interactions. *Rep Prog Phys.* 62(6):921.
- Braasch I, et al. 2009. Pigmentation pathway evolution after whole-genome duplication in fish. *Genome Biol Evol.* 1:479–493.
- Brunet FG, et al. 2006. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol Biol Evol.* 23(9):1808–1816.
- Butruille DV, Boiteux LS. 2000. Selection–mutation balance in polysomic tetraploids: impact of double reduction and gametophytic selection on the frequency and subchromosomal localization of deleterious mutations. *Proc Natl Acad Sci U S A.* 97(12):6608–6613.
- Byrne KP, Wolfe KH. 2007. Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication. *Genetics* 175(3):1341–1350.
- Cardoso JCR, Pinto VC, Vieira FA, Clark MS, Power DM. 2006. Evolution of secretin family GPCR members in the metazoa. *BMC Evol Biol.* 6(1):108.
- Carroll SB. 2001. Chance and necessity: the evolution of morphological complexity and diversity. *Nature* 409(6823):1102.
- Christoffels A, et al. 2004. Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. *Mol Biol Evol.* 21(6):1146–1151.
- Clark AG. 1994. Invasion and maintenance of a gene duplication. *Proc Natl Acad Sci U S A.* 91(8):2950–2954.
- Comai L. 2005. The advantages and disadvantages of being polyploid. *Nat Rev Genet.* 6(11):836.
- Conant GC, Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet.* 9(12):938.
- Cuatrecasas P, Hollenberg MD. 1976. Membrane receptors and hormone action. *Adv Protein Chem.* 30:251–451.
- Eagle RA, Trowsdale J. 2007. Promiscuity and the single receptor: NKG2D. *Nat Rev Immunol.* 7(9):737.
- Escriva H, et al. 2006. Neofunctionalization in vertebrates: the example of retinoic acid receptors. *PLoS Genet.* 2(7):e102.
- Glasauer SM, Neuhauss SC. 2014. Whole-genome duplication in teleost fishes and its evolutionary consequences. *Mol Genet Genomics.* 289(6):1045–1060.
- Grandchamp A, Monget P. 2018. Synchronous birth is a dominant pattern in receptor–ligand evolution. *BMC Genomics.* 19(1): 611.
- Hahn MA, van Kleunen M, Müller-Schärer H. 2012. Increased phenotypic plasticity to climate may have boosted the invasion success of polyploid *Centaurea stoebe*. *PLoS One* 7(11):e50284.
- Huffon AL, Panopoulou G. 2009. Polyploidy and genome restructuring: a variety of outcomes. *Curr Opin Genet Dev.* 19(6):600–606.
- Hurst LD, Smith NG. 1999. Do essential genes evolve slowly? *Curr Biol.* 9(14):747–750.
- Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet.* 11(2):97.
- Itoh N, Ornitz DM. 2004. Evolution of the Fgf and Fgfr gene families. *Trends Genet.* 20(11):563–569.
- Iwema T, et al. 2007. Structural and functional characterization of a novel type of ligand-independent RXR-USP receptor. *EMBO J.* 26(16):3770–3782.
- Jaillon O, et al. 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 431(7011):946–957.
- Kasahara M, et al. 2007. The Medaka draft genome and insights into vertebrate genome evolution. *Nature* 447(7145):714.
- Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 428(6983):617.
- Khosla A, Lucas SG. 2016. Cretaceous period: biotic diversity and biogeography: bulletin 71. Albuquerque:New Mexico Museum of Natural History and Science.
- Kohn M, et al. 2006. Reconstruction of a 450-My-old ancestral vertebrate protokaryotype. *Trends Genet.* 22(4):203–210.
- Kuroda M, et al. 2001. Whole genome sequencing of methicillin-resistant *Staphylococcus aureus*. *Lancet* 357(9264):1225–1240.

- Long M, et al. 2003. The origin of new genes: glimpses from the young and old. *Nat Rev Genet.* 4(11):865.
- Louis A, Nguyen NTT, Muffato M, Crollius HR. 2015. Genomic update 2015: KaryoView and MatrixView provide a genome-wide perspective to multispecies comparative genomics. *Nucleic Acids Res.* 43(D1):D682–D689.
- Lynch M, O’Hely M, Walsh B, Force A. 2001. The probability of preservation of a newly arisen gene duplicate. *Genetics* 159(4):1789–1804.
- Mable BK, Alexandrou MA, Taylor MI. 2011. Genome duplication in amphibians and fish: an extended synthesis. *J Zool.* 284(3):151–182.
- Maere S, et al. 2005. Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci U S A.* 102(15):5454–5459.
- 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.
- Markov GV, Laudet V. 2011. Origin and evolution of the ligand-binding ability of nuclear receptors. *Mol Cell Endocrinol.* 334(1-2):21–30.
- Martín-Abad H, Poyato-Ariza FJ. 2013. Historical patterns of distribution in pycnodontiform and amiiform fishes in the context of moving plates. *Geol Belgica* 16: number 4.
- Meyer A, Van de Peer Y. 2005. From 2R to 3R: evidence for a Fish-Specific Genome Duplication (FSGD). *BioEssays* 27(9):937–945.
- Near TJ, et al. 2012. Resolution of ray-finned fish phylogeny and timing of diversification. *Proc Natl Acad Sci U S A.* 109(34):13698–13703.
- O’Hely M. 2006. A diffusion approach to approximating preservation probabilities for gene duplicates. *J Math Biol.* 53(2):215–230.
- Ohno S, Wolf U, Atkin NB. 1968. Evolution from fish to mammals by gene duplication. *Hereditas* 59(1):169–187.
- Olmo E. 1983. Nucleotype and cell size in vertebrates: a review. *Basic Appl Histochem.* 27(4):227–256.
- Pál C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. *Genetics* 158(2):927–931.
- Papp B, Pal C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature* 424(6945):194.
- Pasquier J, 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.
- Ram K. 2013. Git can facilitate greater reproducibility and increased transparency in science. *Source Code Biol Med.* 8(1):7.
- Ramilowski JA, et al. 2015. A draft network of ligand–receptor-mediated multicellular signalling in human. *Nat Commun.* 6:7866.
- Ramsey J, Schemske DW. 2002. Neopolyploidy in flowering plants. *Annu Rev Ecol Syst.* 33(1):589–639.
- Ravi V, Venkatesh B. 2008. Rapidly evolving fish genomes and teleost diversity. *Curr Opin Genet Dev.* 18(6):544–550.
- Reschly EJ, et al. 2008. Ligand specificity and evolution of liver X receptors. *J Steroid Biochem Mol Biol.* 110(1-2):83–94.
- Riechmann JL, et al. 2000. *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes. *Science* 290(5499):2105–2110.
- Sankoff D, Zheng C, Zhu Q. 2010. The collapse of gene complement following whole genome duplication. *BMC Genomics.* 11(1):313.
- Schlesinger K, et al. 2014. Coevolutionary immune system dynamics driving pathogen speciation. *PLoS One* 9(7): e102821.
- Taylor JS, Raes J. 2004. Duplication and divergence: the evolution of new genes and old ideas. *Annu Rev Genet.* 38:615–643.
- Te Beest M, et al. 2012. The more the better? The role of polyploidy in facilitating plant invasions. *Ann Bot.* 109(1):19–45.
- Van de Peer Y, Maere S, Meyer A. 2009. The evolutionary significance of ancient genome duplications. *Nat Rev Genet.* 10(10):725.
- Van de Peer Y, Mizrachi E, Marchal K. 2017. The evolutionary significance of polyploidy. *Nat Rev Genet.* 18(7):411.
- Veitia RA. 2004. Gene dosage balance in cellular pathways. *Genetics* 168(1):569–574.
- Vilella AJ, et al. 2009. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* 19(2):327–335.
- Wall DP, Fraser HB, Hirsh AE. 2003. Detecting putative orthologs. *Bioinformatics* 19(13):1710–1711.
- Wilson G, et al. 2014. Best practices for scientific computing. *PLoS Biol.* 12(1):e1001745.

Associate editor: Adam Eyre-Walker