Heliyon 9 (2023) e20217

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

# Genes encoding teleost orthologues of human signal transduction proteins remain duplicated or triplicated more frequently than the whole genome

Floriane Picolo [a], Benoît Piégu [a], Philippe Monget [a,*]

[a] PRC, UMR85, INRAE, CNRS, IFCE, Université de Tours, F-37380 Nouzilly, France

## ABSTRACT

Cell signalling involves a myriad of proteins, many of which belong to families of related proteins, and these proteins display a huge number of interactions. One of the events that has led to the creation of new genes is whole genome duplication (WGD), a phenomenon that has made some major innovations possible. In addition to the two WGDs that happened before gnathostome radiation, teleost genomes underwent one (the 3WGD group) or two (the 4WGD group) extra WGD after separation from the lineage leading to holostei. In the present work, we studied in 63 teleost species whether the orthologues of human genes involved in 47 signalling pathways (HGSP) remain more frequently duplicated, triplicated or in the singleton state compared with the whole genome. We found that these genes have remained duplicated and triplicated more frequently in teleost of the 3WGD and 4WGD groups, respectively. Moreover, by examining pairs of interacting gene products in terms of conserved copy numbers, we found a majority of the 1:1 and 1:2 proportions in the 3WGD group (between 54% and 60%) and of the 2:2 and 2:4 proportions in the 4WGD group (30%). In both groups, we observed the 0:n proportion at a mean of approximately 10%, and we found some pseudogenes in the concerned genomes. Finally, the proportions were very different between the studied pathways. The n:n (i.e. same) proportion concerned 20%–65% of the interactions, depending on the pathways, and the n:m (i.e. different) proportion concerned 34%–70% of the interactions. Among the n:n proportion, the 1:1 ratio is most represented (25.8%) and among the n:m ratios, the 1:2 is most represented (25.0%). We noted the absence of gene loss for the JAK-STAT, FoxO and glucagon pathways. Overall, these results show that the teleost gene orthologues of HGSP remain duplicated (3WGD) and triplicated (4WGD) more frequently than the whole genome, although some genes have been lost, and the proportions have not always been maintained.

## 1. Introduction

Cell signaling involves a large number of proteins, many of which belong to families of more or less related proteins, and these proteins together display a huge number of interactions. One of the events that led to the creation of new genes was whole genome duplication (WGD), which made some major innovations possible. In additional to the two WGDs that happened in vertebrate genomes, the common ancestor of extant teleost fish experienced a third WGD ∼320 Million Years Ago (MYA) after separation from the lineage leading to holostei. Moreover, the salmonid and carp lineages have experienced additional independent fourth WGD events 100 and 10 MYA respectively. One of the major assumptions of the preservation of duplicated genes is a dose effect [1]. Indeed, this postulate is based on the idea that the number of copies of a gene is influenced by the dosage of the products with which it interacts (as for example in a signaling pathway) [2]. The products of a signaling pathway should interact in proportion amounts and an imbalance in dosage could lead to a physiological disorder [3]. The functional capacity of a gene lies partly in its dosage, as for example for haplo-insufficient genes which do not produce a functional wild type phenotype if they are in a single copy [4]. We have shown recently that these haplo-insufficient genes are more often maintained in duplicate than the rest of the genes of the genome in teleost

[*] Corresponding author.
*E-mail address:* philippe.monget@inrae.fr (P. Monget).

https://doi.org/10.1016/j.heliyon.2023.e20217
Received 23 February 2023; Received in revised form 18 August 2023; Accepted 13 September 2023
Available online 16 September 2023

[5]. Most of the interactions between signaling proteins concern enzymatic chain reactions, for which the stoichiometry of the interactions is supposed to be important. From this observation, it would be natural to think that the proportion of proteins involved in these interactions should be of the order of 1:1 (or more generally n:n) for pairs of genes that interact together as may be the case in an intracellular signaling pathway.

The rapid evolution of vertebrates, in terms of diversification and the emergence of species, is notably due to this double WGD, but also due to deletions and more recently species-specific duplications of genes. These processes have rapidly increased the diversity of species [6]. The teleost group, which represents for half of the vertebrates [7], is an interesting group to study from an evolutionary point of view because of its huge phenotypic, environmental and genome size diversity.

The teleost clade (63 teleost species examined in this study) comprises sub-clades that we have divided into two groups in this work. The first sub-clade contains 54 species that have undergone a third WGD (the 3WGD group) after separation from the lineage leading to gars and bowfins (the holostei clade); the second sub-clade includes the salmonid clade (Huchen, Atlantic salmon, Rainbow trout, Coho salmon, Brown trout and Chinook salmon) and the carp clade (Goldfish, Common carp and Golden-line barbel), which have undergone a fourth WGD each (the 4WGD group) [8,9]. Of note, the fourth genome duplication did not occur in the same way in the salmonid and cyprinid groups. The salmonid 4WGD occurred 100 MYA and has been proposed to be an auto-tetraploidization, whereas the cyprinid 4WGD occurred more recently (10 MYA) and has been proposed to be an allo-tetraploidisation [10].

The intracellular signalling pathways are very well referenced in humans, rodents and yeast, but they have been poorly studied in teleosts (there are a little more than 1000 PubMed results with the keywords 'teleost signalling pathway' against more than 600,000 PubMed results with the keywords 'mammals signalling pathway'; https://pubmed.ncbi.nlm.nih.gov). Some pathways are also present and referenced in yeast, notably the sphingolipid [11], MAPK [12] or cAMP-PKA [13] signalling pathway. If a pathway is present in yeast, we hypothesise that it is also present in teleosts, even if some orthologues of human genes that encode proteins involved in a signalling pathway (HGSP) may have been lost during evolution.

In the present work, we investigated whether the teleost orthologues of HGSP have remained duplicated (3WGD) or triplicated (4WGD) or returned to the singleton state or in the duplicate state relative to the whole genome.

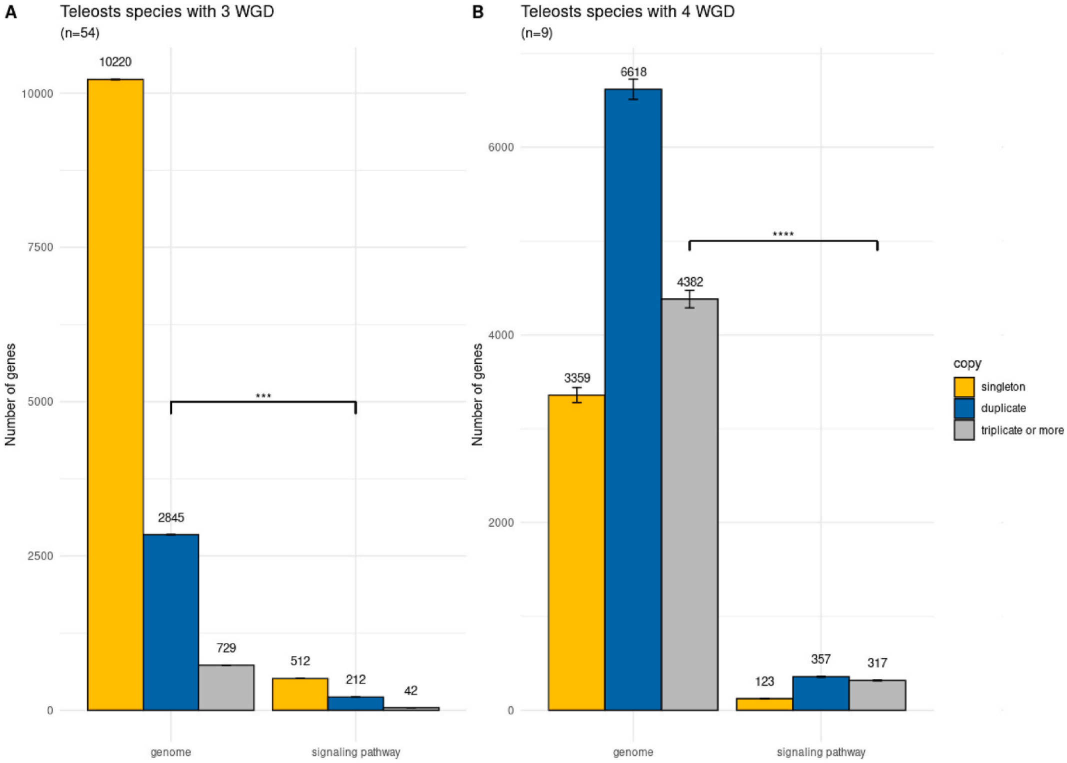


**Fig. 1.** Bar plot of the global distribution of the genes. (A) The means of teleost 3WGD orthologues, with human genes of the whole genome on the left and HGSP on the right. (B) The means of teleost 4WGD orthologues with human genes of the whole genome on the left and HGSP on the right. The yellow bars correspond to the genes that returned to the singleton state (1 copy); the blue bars correspond to the genes that have been retained (for the 3WGD teleost) or returned (for the 4WGD teleost) to duplicate (two copies). The grey bars correspond to the genes present in three or more copies. The results are presented as the mean ± standard error of the mean. * indicates a significant difference for duplicate orthologues of HGSP (***$p < 0.001$) and triplicate orthologues of HGSP (****$p < 0.0001$) compared with the whole genome of the 3WGD and 4WGD groups, respectively.

## 2. Results

### 2.1. Global distribution of HGSP in 63 teleost species

We found that out of 22,727 human protein-coding genes, on average, 13,875 genes (61.1%) have at least one teleost orthologue. Of these genes, an average of 9240 genes (ranging from 2328 to 10,695 depending on the studied species) have reverted to the singleton state, an average of 3384 genes (ranging from 2423 to 8263) have remained duplicated and an average of 1,251genes (ranging from 502 to 6374) have three or more copies.

Of the 2989 HGSP, an average of 770 (33.5%) have at least one teleost orthologue. Of these genes, an average of 457 (ranging from 78 to 542) have reverted to the singleton state, an average of 232 (ranging from 178 to 445) have been retained duplicated and an average of 81 genes (ranging from 27 to 422) have three or more copies (Fig. 1 and Suppl. Data 1).

For each of the 54 teleost species in the 3WGD group, more HGSP have been retained in duplicate compared with the whole genome (for 54 species, p-value from 1.00E-03 to 7.74E-15 by $\chi^2$ analysis after Benjamini–Hochberg [BH] correction, and p-value from 5.48E-04 to 5.49E-04 after hypergeometric test and BH correction, depending on the species considered; see Suppl. Data 1). For each of the nine teleost species in the 4WGD group, more HGSP have been retained in triplicate compared with the whole genome (for all nine species, p-value from 2.00E-05 to 1.33E-08 by $\chi^2$ analysis after BH correction, and p-value from 2.00E-05 to 9.25E-09 after hypergeometric test and BH correction, depending on the species considered; see Suppl. Data 1). For these 4WGD species, there are not more duplicated HGSP relative to the whole genome, except in the golden-line barrel (BH-corrected p = 1.00E-03).

### 2.2. Global proportion for gene interactions in cellular pathways of each teleost species

In this part of the study, we did not include the PPAR pathway because in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database, it is only represented by gene expression, instead of protein–protein interactions. The signalling pathways are characterised by a chain of protein–protein interactions, which led us to consider whether the proportion of gene–gene interactions (in our case) is respected. First, we determined whether the proportion was maintained in some teleost species and not in others, considering both the 3WGD and 4WGD groups.

For the 3WGD group (Fig. 2), the 1:1 proportion was the highest: a mean of 30.3% with a range from 17.7% (*Paramormyrops kingsleyae*) to 35.2% (Turbot). We observed the 2:2 proportion at a mean of 7.2%, ranging from 4.4% (Lumpfish) to 19.2% (*P. kingsleyae*); the 3:3 proportion had a mean of 0.1%. The 1:2 proportion had a mean of 27.3% with a range from 23.0% (Tetraodon) to 36.6% (Asian bonytongue). Finally, we observed the 0:1 proportion at a mean of 18.0% with a range from 11.2% (*P. kingsleyae*) to 22.1% (Javanese ricefish).

For the 4WGD group (Fig. 2), we found a lower 0:n proportion than in the 3WGD group. We found a mean of 2.0% for the 1:1 interaction. The 2:2 proportion was the highest in the 4WGD group: a mean of 15.8% with a range from 7.7% (Chinook salmon) to 24.9% (Common carp). The 2:4 proportion had a mean of 14.0% with a range from 6.9% (Chinook salmon) to 23.9% (Common carp). For the 0:2 proportion, we observed a mean of 11.6% with a range from 9.7% (Chinook salmon) to 15.3% (Common carp) (Fig. 2). This third proportion represents the loss of one of the elements of the interaction while the other partner has returned to being present in

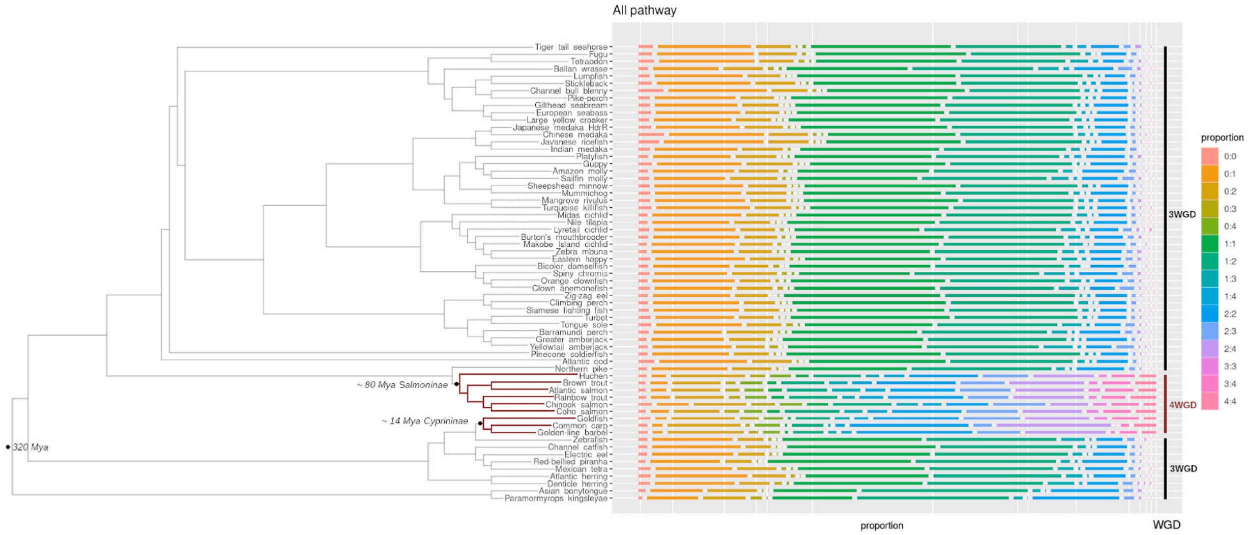


**Fig. 2.** Distribution of the gene–gene interaction proportions by teleost species for all genes involved in signalling pathways. The tree of life of teleosts is presented on the left side. The nodes with a bullet represent a WGD. The black branches are the species with 3WGD, and the red branches are the species with 4WGD. The tree is not to scale, and the genome duplication times have been added according to previous studies [9,22–24]. The right side shows the distribution of gene–gene interaction proportions. Each bar colour represents one proportion.

duplicate. In the 4WGD group, the Common carp showed the highest 2:4 (23%) and 2:2 (24%) proportions, while the Chinook salmon showed the lowest percentages (7%). Interestingly, despite these different 'histories' of fourth duplication between the salmonid and cyprinid groups, the proportions between partner proteins are very similar between both groups (Fig. 2).

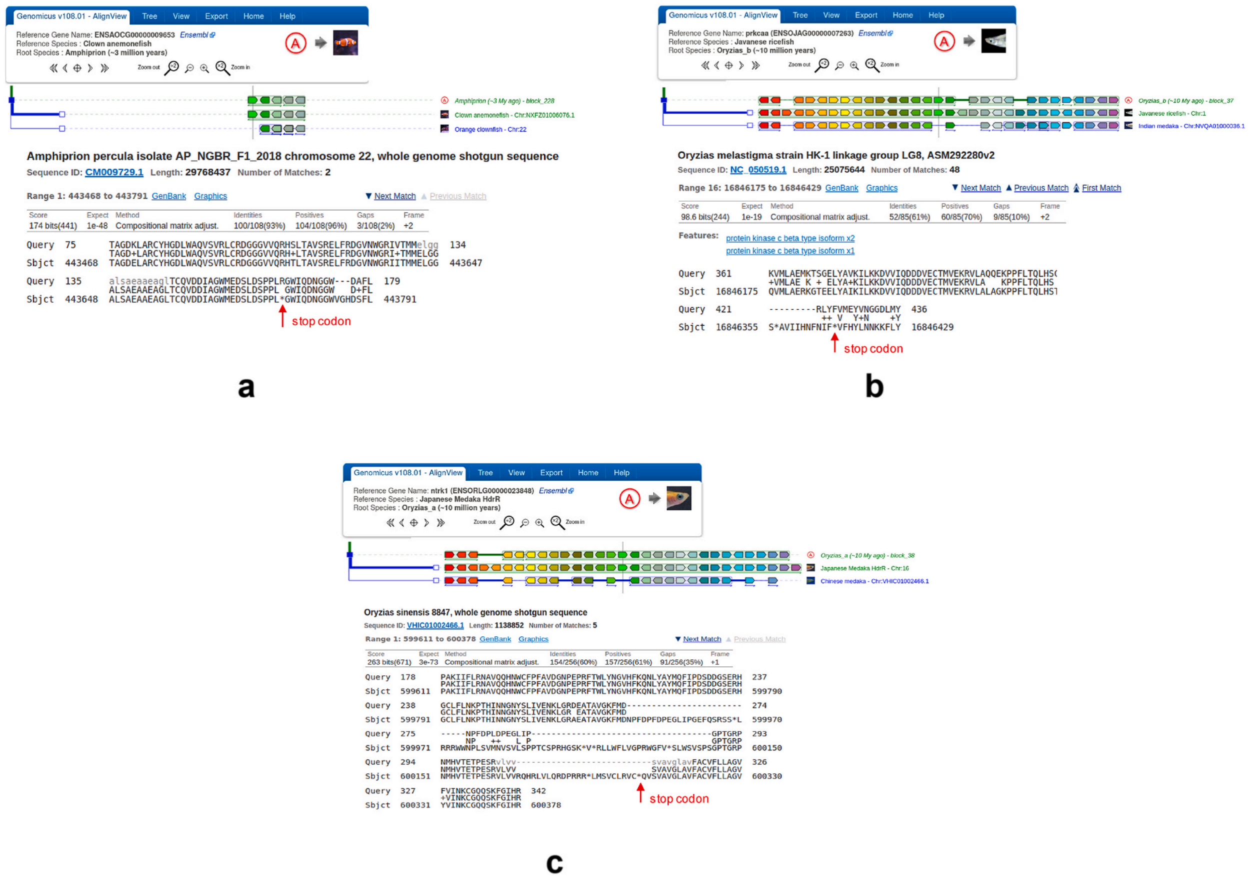
In both the 3WGD and 4WGD groups, a very large majority of genes almost never respect an n:n proportion (in more than 95% of the species studied here). These genes belong to all signaling pathways.

For both the 3WGD and 4WGD groups, the 0:0, 0:1, 0:2, 0:3 and 0:4 proportions occurred at a mean of 5.8% with a range from 0.1% to 22.1%. We observed the complete loss of interaction – with the loss of both partners, *PYCARD* and *CASP8* (corresponding to the 0:0 proportion) – for all the species except *P. kingsleyae*, which diverged early from the lineage that contains the majority of teleost species in the tree of life. Moreover, several genes are absent in a large majority of the 63 studied species except for less than three species (they are never the same species), and for which the proportion had not been maintained. Once again, the pycard gene is absent in a majority of species and the other concerned genes encode proteins involved in different pathways: *bad, ccl26, bid, map2k3, traf3ip2, casp8, nlrp1, nfkbia, csnk2a1, cd*40lg, *mefv, bbc3, prkd3, rap1a, tp53aip1, ticam1, traf3, ywhaq, tlr4, nlrp12, ly96, nlrp3, ifi16, aim2, icos, rgs14, cycs, cxcr6, nlrp6, pydc2* and *pydc5*, some of which interact with pycard (see Suppl. Data 2).

These 0:n proportions strongly suggest the loss by pseudogenisation of several teleost orthologues of HGSP. We failed to find the pycard pseudogenes in the 62 teleosts in which the gene is absent from the Ensembl phylogenetic trees in the teleost except in *P. kingsleyae*, probably because the evolutionary distance is too great. On the other hand, we found some other pseudogenes like that of *bcl2 like*, present in two copies in Clown anemonefish and only in one copy in Orange clownfish. We also found a pseudogene prkca in the Indian medaka species (Fig. 3).

### 2.3. Proportion of gene interactions in all teleost species for each cellular pathway

We evaluated the proportion of gene–gene interactions for each of the intracellular signalling pathways. We classified the different proportions into the following categories: 0:0; n:n, representing equal proportions (1:1, 2:2, 3:3 and 4:4); and n:m, representing



Reference Gene Name: ENSAOCG00000009653 Ensembl
Reference Species : Clown anemonefish
Root Species : Amphiprion (~3 million years)

Amphiprion (~3 My ago) - block_228
Clown anemonefish - Chr:NXFZ01006076.1
Orange clownfish - Chr:22

**Amphiprion percula isolate AP_NGBR_F1_2018 chromosome 22, whole genome shotgun sequence**

Sequence ID: CM009729.1 Length: 29768437 Number of Matches: 2

Range 1: 443468 to 443791 GenBank Graphics ▼ Next Match ▲ Previous Match

```
Score           Expect  Method                         Identities     Positives      Gaps        Frame
174 bits(441)   1e-48   Compositional matrix adjust.   100/108(93%)   104/108(96%)   3/108(2%)   +2

Query  75      TAGDKLARCYHGDLWAQVSVRLCRDGGGVVQRHSLTAVSRELFRDGVNWGRIVTMMelgg  134
               TAGD+LARCYHGDLWAQVSVRLCRDGGGVVQRH+LTAVSRELFRDGVNWGRI+TMMELGG
Sbjct  443468  TAGDELARCYHGDLWAQVSVRLCRDGGGVVQRHTLTAVSRELFRDGVNWGRIITMMELGG  443647

Query  135     alsaeaaeaglTCQVDDIAGWMEDSLDSPPLRGWIQDNGGW---DAFL  179
               ALSAEAAEAGLTCQVDDIAGWMEDSLDSPPL GWIQDNGGW   D+FL
Sbjct  443648  ALSAEAAEAGLTCQVDDIAGWMEDSLDSPPL*GWIQDNGGWVGHDSFL  443791
```

Reference Gene Name: prkcaa (ENSOJAG00000007263) Ensembl
Reference Species : Javanese ricefish
Root Species : Oryzias_b (~10 million years)

Oryzias_b (~10 My ago) - block_37
Javanese ricefish - Chr:1
Indian medaka - Chr:NVQA01000036.1

**Oryzias melastigma strain HK-1 linkage group LG8, ASM292280v2**

Sequence ID: NC_050519.1 Length: 25075644 Number of Matches: 48

Range 16: 16846175 to 16846429 GenBank Graphics ▼ Next Match ▲ Previous Match ▲ First Match

```
Score          Expect  Method                         Identities   Positives    Gaps        Frame
98.6 bits(244) 1e-19   Compositional matrix adjust.   52/85(61%)   60/85(70%)   9/85(10%)   +2
```

Features: protein kinase c beta type isoform x2
protein kinase c beta type isoform x1

```
Query  361       KVMLAEMKTSGELYAVKILKKDVVIQDDDVECTMVEKRVLAQQEKPPFLTQLHS(
                 +VMLAE K + ELYA+KILKKDVVIQDDDVECTMVEKRVLA   KPPFLTQLHS
Sbjct  16846175  QVMLAERKGTEELYAIKILKKDVVIQDDDVECTMVEKRVLALAGKPPFLTQLHST

Query  421       ---------RLYFVMEYVNGGDLMY  436
                      ++ V  Y+N    +Y
Sbjct  16846355  S*AVIIHNFNIF*VFHYLNNKKFLY  16846429
```

Reference Gene Name: ntrk1 (ENSORLG00000023848) Ensembl
Reference Species : Japanese Medaka HdrR
Root Species : Oryzias_a (~10 million years)

Oryzias_a (~10 My ago) - block_38
Japanese Medaka HdrR - Chr:16
Chinese medaka - Chr:VHIC01002466.1

**Oryzias sinensis 8847, whole genome shotgun sequence**

Sequence ID: VHIC01002466.1 Length: 1138852 Number of Matches: 5

Range 1: 599611 to 600378 GenBank Graphics ▼ Next Match ▲ Previous Match

```
Score           Expect  Method                         Identities     Positives      Gaps          Frame
263 bits(671)   3e-73   Compositional matrix adjust.   154/256(60%)   157/256(61%)   91/256(35%)   +1

Query  178     PAKIIFLRNAVQQHNWCFPFAVDGNPEPRFTWLYNGVHFKQNLYAYMQFIPDSDDGSERH  237
               PAKIIFLRNAVQQHNWCFPFAVDGNPEPRFTWLYNGVHFKQNLYAYMQFIPDSDDGSERH
Sbjct  599611  PAKIIFLRNAVQQHNWCFPFAVDGNPEPRFTWLYNGVHFKQNLYAYMQFIPDSDDGSERH  599790

Query  238     GCLFLNKPTHINNGNYSLIVENKLGRDEATAVGKFMD-----------------------  274
               GCLFLNKPTHINNGNYSLIVENKLGR EATAVGKFMD
Sbjct  599791  GCLFLNKPTHINNGNYSLIVENKLGRAEATAVGKFMDNPFDPFDPEGLIPGEFQSRSS*L  599970

Query  275     -----NPFDPLDPEGLIP---------------------------------GPTGRP  293
                    NP   ++    L P                                 GPTGRP
Sbjct  599971  RRRWWNPLSVMNVSVLSPPTCSPRHGSK*V*RLLWFLVGPRWGFV*SLWSVSPSGPTGRP  600150

Query  294     NMHVTETPESRvlvv-------------------------svavglavFACVFLLAGV  326
               NMHVTETPESRVLVV                         SVAVGLAVFACVFLLAGV
Sbjct  600151  NMHVTETPESRVLVVRQHRLVLQRDPRRR*LMSVCLRVC*QVSVAVGLAVFACVFLLAGV  600330

Query  327     FVINKCGQQSKFGIHR  342
               +VINKCGQQSKFGIHR
Sbjct  600331  YVINKCGQQSKFGIHR  600378
```

**Fig. 3.** Pseudogenes found by tblastn. (A) A screen capture of Genomicus (https://www.genomicus.bio.ens.psl.eu/genomicus-100.01/cgi-bin/search.pl), which we used to visualise a gene present in one species and absent in another. The gene of interest is in light green. We chose species that are close to each other to maximise our chances to find a trace of our genes. (B) Protein sequence alignment of the gene of the closest species against our species of interest. This is a tblastn alignment against the whole genome of the target species. The stars represent the stop codons in the sequence.

proportions that have not been maintained (0:1, 0:2, 1:2, …, 3:4). We found that the proportion of gene–gene interactions is very different between the pathways, with the n:n proportion being more represented in the Hedgehog and oestrogen pathways than the *Rap1, IL-17* or *RIG-I-like* receptor pathways. The percentage of this type of interaction decreased among all pathways, from 65% to 20%. Among the n:n proportions, the 1:1 proportion had a mean of 71.4% of interactions, the 2:2 had a mean of 26.0% and the 3:3 and 4:4 proportions had a mean of 0.9% and 1.7%, respectively (Fig. 4).

The n:m proportion is widely represented in all signalling pathways, ranging from 34% (Hedgehog) to about 70% (*Rap1, IL-17, RIG-I-like* receptor and adipocytokine). The proportions most represented in the n:m category are 1:2 (mean 44.3%), 0:1 (mean 24.2%), 0:2 (mean 13.2%) and then the other proportions (mean 18.3%).

We observed no gene loss for the *JAK-STAT, FoxO* and glucagon pathways. Most gene loss (>7.7%) occurred for the C-type lectin receptor, *RIG-I-like* receptor, *NOD-like* receptor and *NF-kappa B* pathways (see Suppl. Data 3).

## 3. Discussion

We found that the teleost orthologues of HGSP remained duplicated for the 3WGD species and triplicated for the 4WGD species

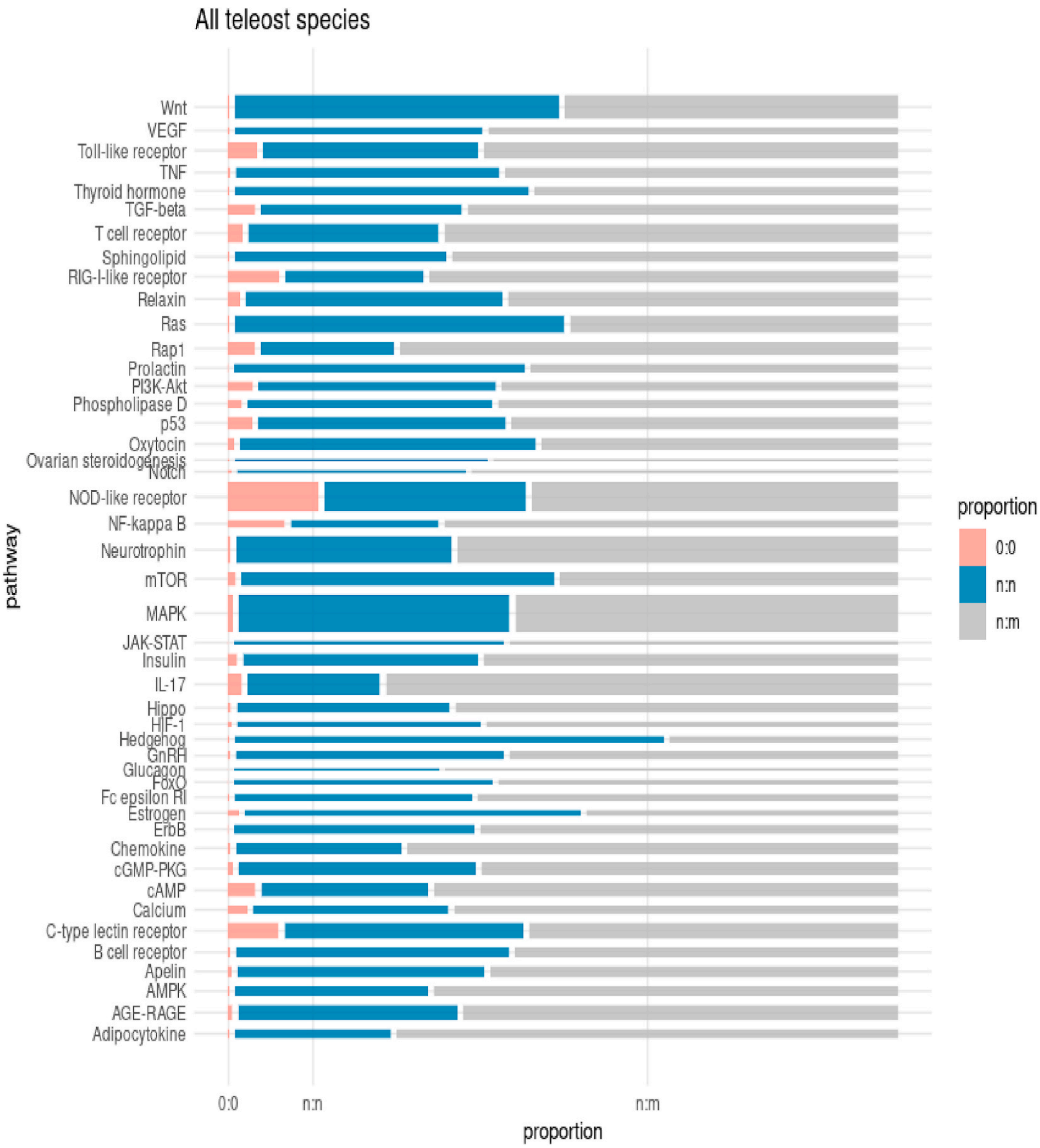


**Fig. 4.** Distribution of orthologues of the human gene–gene interaction proportions by signalling pathway species for all genes with an HGSP orthologue in teleost species. Each bar colour represents one type of proportion: 0:0 (pink), n:n (1:1, 2:2, 3:3 and 4:4, blue) and n:m (0:1, 0:2, …,3:4, grey). The thickness of each bar is proportional to the number of unique interactions in each pathway.

more frequently than the whole genome. We obtained similar results for teleost orthologues of genes encoding ligand/membrane receptor pairs [14]. This previous work had only been carried out with 10 teleost species. Here we have studied 63 species, which strengthens our conclusions. The fact that these genes remained in duplicate or more instead of returning to singleton suggest that there is selection pressure that keeps them in duplicate, triplicate or more. The hypothesis is that evolution would favor the maintenance of these HGSP genes in large quantities because they would thus be more favorable to the survival of the species.

Concerning signalling proteins, as for ligand–receptor pairs, one of the questions raised here concerns whether the proportion of the interactions studied has been respected. In the case of ligand–receptor interactions, partners have not necessarily evolved in the same way, and there have been situations in which one of the partners has returned to the singleton state while the other one has been maintained in duplicate [14]. This suggests that changes in ligand–receptor interactions may have taken place during the evolution of teleosts. In the present study, the n:n proportion was more represented in the 3WGD group (>37.5%) than in the 4WGD group (>23.1%). Interestingly, among the genes that remained duplicated or triplicated depending on the group of teleost concerned, we found 74 teleost gene orthologues of HGSP that are haplo-insufficient (*APC*, a regulator of the *WNT* signalling pathway; *AMT* [serine/threonine kinase]; *JAG1* [jagged canonical Notch ligand 1]; and 107 orthologues of HGSP that are mono-allelically expressed [*cd38, cd86, cd72*, etc.]). We have previously studied these genes [5], a fact that strengthens the conclusions of the present work. Interestingly, despite these different 4WGD 'histories' between the salmonid and cyprinid groups, the proportions between partner proteins are very similar in both groups (Fig. 2). This suggests that the process of polyploidisation does not influences whether the genes encoding teleost HGSP return to triplicates or remain in quadruplicates.

Only one interaction, *CASP1:CARD16*, respects the strict proportion (from 2:2 in Tetraodon to 14:14 in Zebra mbuna) among the 63 teleost species we studied. This finding strongly suggests evolutionary pressure to maintain a strict equilibrium in term of the concentration of both partners. Several genes respect this n:n proportion in at least 95% of teleost species. The concerned genes encode proteins involved in different pathways: *CASP8, NOS1, MAPK1, PRKCA, TP53, BAX, FZD10, DVL1, FOS, GNA12, CALM6, SRC, NFkB* and *MTOR*, among others. This suggests that the n:n proportion of these partners has been maintained during the evolution of certain species, but it has been relaxed in others. However, we do not know whether these paralogues are expressed in the same cells or if some have not become sub-functionalised or neo-functionalised.

The presence of several cases of 0:n interactions strongly suggests that several orthologues of HGSP have been lost during evolution. Although we found some pseudogenes (bcl2-like, nerve growth factor receptor), we failed to find many others. It is possible that the evolutionary distance is too great to find any stop codon or other frameshift mutations, insertions or deletions. It is also possible that some of these predicted absent genes are truly present but have not been sequenced or annotated in these recently sequenced genomes. Moreover, there are about 74,962 protein-coding genes in the genome of 4WGD species (range from 55,255 to 99,592), and about 33,640 protein-coding genes in the genome of 3WGD species range from 23,886 to 100,231). Regarding the HGSP orthologues, there are about 14,359 genes in the 4WGD species, and 13,794 in the 3WGD species. This strongly suggests that the genomes of the 4WGD group have lost many more HGSP orthologues than the genomes of the 3WGD group [10], likely due to a more intense process of pseudogenization.

Regarding these predicted 0:n interactions, it is unlikely that the signalling pathway functions without the presence of one of its members. On the other hand, it is very probable that the lost gene has been replaced by a paralogue of the same family which could replace the member described in KEGG. In particular, the majority if not all of these genes belong to families (bad, mapk, casp, etc.) for which functional redundancy is well documented [14,15].

Some genes have unusual characteristics. For example, clec4d, clec4m and clec6a are present in 3–33 copies in the 63 teleosts we studied. These genes encode membrane receptors with an extracellular C-type lectin-like domain that recognises several pathogens, and are involved in inflammation and immune response, suggesting a particularly important role for these biological functions in teleost species. The stat1 gene is also present in 15 copies in Ballan wrasse; the casp1 gene is present in 10, 14 and 12 copies in Eastern happy, Zebra mbuna and Golden-line barbel; and the *IFNG:IFNGR1* interaction in present as a 4:13, 4:16 and 8:17 ratio in Golden-line barbel, Common carp and Goldfish, respectively. Additional investigations are needed to understand the biological significance of these duplications.

The p53 gene also caught our attention: there are 25 copies in the Siamese fighting fish, reminiscent of the massive duplications of the same gene in the elephant [15,16]. In this latter species, it has been suggested that this duplication is partly responsible for the low incidence of cancer and its long life expectancy (Peto's paradox). The Siamese fighting fish is an aquarium fish with a size of 6–8 cm and a life expectancy that is not particularly long (3–5 years in captivity). It is possible that this teleost develops cancer less frequently than other aquarium fish species. However, it is probably difficult to compare the elephant and an aquarium fish in terms of life expectancy and cancer incidence.

Overall, our work shows that teleost genes orthologues of HGSP remain more duplicated in the 3WGD group and more triplicated in the 4WGD group compared with the whole genome. Moreover, some genes involved in almost all pathways studied have been lost, and there are many protein–protein interactions for which the proportion has not been maintained.

## 4. Material and methods

### 4.1. Implementation of the database

We focused on human genes coding for protein involved in 47 different signalling pathways. We obtained the human genes of these pathways with KEGG database V104.0 (https://www.genome.jp/kegg), which is the most popular metabolic database [17], by using the keywords 'signalling pathway' or 'ovarian' and 'human'. These signalling pathways are named as follows in KEGG: *PPAR, MAPK,* *ErbB, Ras, Rap1, Calcium, cGMP-PKG, cAMP*, Chemokine, *NF-kappa B, HIF- 1, FoxO,* Sphingolipid, Phospholipase D, *p53, mTOR, PI3K-Akt, AMPK, Wnt, Notch*, Hedgehog, *TGF-beta, VEGF,* Apelin, Hippo, Toll-like receptor, *NOD-like* receptor, *RIG-I-like* receptor, C-type lectin receptor, *JAK-STAT, IL-17*, T cell receptor, B cell receptor, Fc epsilon RI, *TNF*, Neurotrophin, Insulin, *GnRH*, Ovarian steroidogenesis, Estrogen, Prolactin, Thyroid hormone, Adipocytokine, Oxytocin, Glucagon, Relaxin and *AGE-RAGE*.

The number of genes encoding proteins involved in these signalling pathways varies from 51 to 353, and several proteins are involved in different pathways. We investigated a total of 2295 genes (see Suppl. Data 2).

We studied 63 species of teleosts available on Ensembl database: Amazon molly (*Poecilia formosa*), Asian bonytongue (*Scleropages formosus*), Atlantic cod (*Gadus morhua*), Atlantic herring (*Clupea harengus*), Atlantic salmon (*Salmo salar*), Ballan wrasse (*Labrus bergylta*), Barramundi perch (*Lates calcarifer*), Bicolour damselfish (*Stegastes partitus*), Brown trout (*Salmo trutta*), Burton's mouthbrooder (*Haplochromis burtoni*), Channel bull blenny (*Cottoperca gobio*), Channel catfish (*Ictalurus punctatus*), Chinese medaka (*Oryzias sinensis*), Chinook salmon (*Oncorhynchus tshawytscha*), Climbing perch (*Anabas testudineus*), Clown anemonefish (*Amphiprion ocellaris*), Coho salmon (*Oncorhynchus kisutch*), Common carp (*Cyprinus carpio*), Denticle herring (*Denticeps clupeoides*), Eastern happy (*Astatotilapia calliptera*), Electric eel (*Electrophorus electricus*), European seabass (*Dicentrarchus labrax*), Fugu (*Takifugu rubripes*), Gilthead seabream (*Sparus aurata*), Golden-line barbel (*Sinocyclocheilus grahami*), Goldfish (*Carassius auratus*), Greater amberjack (*Seriola dumerili*), Guppy (*Poecilia reticulata*), Huchen (*Hucho hucho*), Indian medaka (*Oryzias melastigma*), Japanese medaka HdrR (*Oryzias latipes*), Javanese ricefish (*Oryzias javanicus*), Large yellow croaker (*Larimichthys crocea*), Lumpfish (*Cyclopterus lumpus*), Lyretail cichlid (*Neolamprologus brichardi*), Makobe Island cichlid (*Pundamilia nyererei*), Mangrove rivulus (*Kryptolebias marmoratus*), Mexican tetra (*Astyanax mexicanus*), Midas cichlid (*Amphilophus citrinellus*), Mummichog (*Fundulus heteroclitus*), Nile tilapia (*Oreochromis niloticus*), Northern pike (*Esox lucius*), Orange clownfish (*Amphiprion percula*), *P. kingsleyae*, Pike-perch (*Sander lucioperca*), Pinecone soldierfish (*Myripristis murdjan*), Platyfish (*Xiphophorus maculatus*), Rainbow trout (*Oncorhynchus mykiss*), Red-bellied piranha (*Pygocentrus nattereri*), Sailfin molly (*Poecilia latipinna*), Sheepshead minnow (*Cyprinodon variegatus*), Siamese fighting fish (*Betta splendens*), Spiny chromis (*Acanthochromis polyacanthus*), Stickleback (*Gasterosteus aculeatus*), Tetraodon (*Tetraodon nigroviridis*), Tiger tail seahorse (*Hippocampus comes*), Tongue sole (*Cynoglossus semilaevis*), Turbot (*Scophthalmus maximus*), Turquoise killifish (*Nothobranchius furzeri*), Yellowtail amberjack (*Seriola lalandi dorsalis*), Zebra mbuna (*Maylandia zebra*), Zebrafish (*Danio rerio*) and Zig-zag eel (*Mastacembelus armatus*).

The genomes of these teleost species have been subjected to a third duplication after divergence with tetrapod (the 3WGD group) or to a fourth duplication (the 4WGD group). The human genes were extracted from Ensembl (http://www.ensembl.org) [18].

## 5. Analysis

We generated a list of human genes (GRCh38.p13) by using BioMart from Ensembl Genes 107. We selected the set of human genes coding for a protein (protein coding) in the gene type filter. The attributes selected in the homologous category were the different teleost species listed in Ensembl. We only selected stable gene identifiers. We generated a list of 22,881 human protein-coding genes. We established the orthologue of each gene in each of the 63 fish species. Then, in each species, we studied the fate (loss of gene, singleton, duplicate, triplicate or more) of all HGSP orthologues. We obtained between 12,863 (Tetraodon) and 14,542 (Brown trout) orthologous genes per fish species (mean 13,878). This does not represent the entire genome of each fish, but it allowed for solid statistical analysis. We investigated whether these fish orthologues of signalling transduction genes remained in triplicate, in duplicate or returned to the singleton state relative to the whole genome. For the statistical analysis, we used the $\chi^2$ test and a hypergeometric analysis with BH correction to test these hypotheses. Regarding genes encoding proteins of signalling pathways that interact in pairs, we recovered a total of 2317 single gene–gene interactions among the 47 pathways of the KEGG database. We used R for all statistical analysis.

## 6. Identification of pseudogenes

We also performed a systematic search for pseudogenes by using tblastn in the studied genomes for genes with no orthologue identified in at least one of the species of interest. This allowed us to test the hypothesis that evolution of teleost orthologues of HGSP is partly characterised by a gene loss pattern, as previously described for seminal plasma genes and for genes encoding proteins of oviductal fluids in mammals [19–21]. We inferred the pseudogene status in a genome if we found a stop codon or an indel in the sequence identified by the similarity search in the syntenic locus in comparison with the other species of interest. For genes that we did not find, and for which there was no pseudogene in the syntenic locus, we can only hypothesise that the gene has been lost.

## Author contribution statement

Floriane Picolo: Performed the experiments; Analyzed and interpreted the data; Wrote the paper. </p>
Benoît Piégu:Analyzed and interpreted the data. </p>
Philippe Monget: Conceived and designed the experiments; Analyzed and interpreted the data; Wrote the paper. </p>

## Data availability statement

The authors are unable or have chosen not to specify which data has been used.

## Declaration of competing interest

I hereby declare that the disclosed information is correct and that no other situation of real, potential or apparent conflict of interest is known to me. I undertake to inform you of any change in these circumstances, including if an issue arises during the course of the meeting or work itself.

## Acknowledgements

The authors are grateful to Dr Julien Bobe and Yann Guiguen for helpful discussions. Furthermore, the authors are very grateful to Alexandra Louis for discussion and their phylogenetic trees.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2023.e20217.

## References

[1] B. Papp, C. Pál, L.D. Hurst, Dosage sensitivity and the evolution of gene families in yeast, Nature 424 (6945) (2003), https://doi.org/10.1038/nature01771. Art. n° 6945, juill.

[2] M. Freeling, B.C. Thomas, Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity, Genome Res. 16 (7) (2006) 805–814, https://doi.org/10.1101/gr.3681406, juill.

[3] T. Makino, A. McLysaght, Ohnologs in the human genome are dosage balanced and frequently associated with disease, Proc. Natl. Acad. Sci. 107 (20) (2010) 9270–9274, https://doi.org/10.1073/pnas.0914697107, mai.

[4] V.T. Dang, K.S. Kassahn, A.E. Marcos, M.A. Ragan, Identification of human haploinsufficient genes and their genomic proximity to segmental duplications, Eur. J. Hum. Genet. 16 (11) (2008) 1350–1357, https://doi.org/10.1038/ejhg.2008.111, juin.

[5] F. Picolo, A. Grandchamp, B. Piégu, A.D. Rolland, R.A. Veitia, P. Monget, Genes encoding teleost Orthologs of human haploinsufficient and monoallelically expressed genes remain in duplicate more frequently than the whole genome, Int. J. Genomics 2021 (2021), 9028667, https://doi.org/10.1155/2021/9028667.

[6] J.S. Taylor, J. Raes, Duplication and divergence: the evolution of new genes and old ideas, Annu. Rev. Genet. 38 (2004) 615–643, https://doi.org/10.1146/annurev.genet.38.072902.092831.

[7] J. Nelson, T. Grande, M. Wilson, Fishes of the World, fifth ed., 2016, https://doi.org/10.1002/9781119174844.

[8] I. Braasch, J.H. Postlethwait, Polyploidy in fish and the teleost genome duplication, in: P.S. Soltis, D.E. Soltis (Eds.), Polyploidy and Genome Evolution, Springer, Berlin, Heidelberg, 2012, pp. 341–383, https://doi.org/10.1007/978-3-642-31442-1_17.

[9] S. Lien, et al., The Atlantic salmon genome provides insights into rediploidization, Nature 533 (7602) (2016) 200–205, https://doi.org/10.1038/nature17164, mai.

[10] E. Parey, A. Louis, J. Montfort, Y. Guiguen, H.R. Crollius, C. Berthelot, An atlas of fish genome evolution reveals delayed rediploidization following the teleost whole-genome duplication, Genome Res. 32 (9) (2022) 1685–1697, https://doi.org/10.1101/gr.276953.122, sept.

[11] D.J. Montefusco, N. Matmati, Y.A. Hannun, The yeast sphingolipid signaling landscape, Chem. Phys. Lipids 177 (2014) 26–40, https://doi.org/10.1016/j.chemphyslip.2013.10.006, janv.

[12] H. Saito, Regulation of cross-talk in yeast MAPK signaling pathways, Curr. Opin. Microbiol. 13 (6) (2010), https://doi.org/10.1016/j.mib.2010.09.001 déc.

[13] P. Portela, S. Rossi, cAMP-PKA signal transduction specificity in Saccharomyces cerevisiae, Curr. Genet. 66 (6) (2020) 1093–1099, https://doi.org/10.1007/s00294-020-01107-6, déc.

[14] A. Grandchamp, B. Piégu, P. Monget, Genes encoding teleost fish ligands and associated receptors remained in duplicate more frequently than the rest of the genome, Genome Biol. Evol., avr (2019), https://doi.org/10.1093/gbe/evz078.

[15] R. Peto, Quantitative implications of the approximate irrelevance of mammalian body size and lifespan to lifelong cancer risk, Philos. Trans. R. Soc. Lond. B Biol. Sci. 370 (1673) (2015), 20150198, https://doi.org/10.1098/rstb.2015.0198 juill.

[16] M. Sulak, et al., TP53 copy number expansion is associated with the evolution of increased body size and an enhanced DNA damage response in elephants, Elife 5 (2016), e11994, https://doi.org/10.7554/eLife.11994 sept.

[17] G.D. Bader, M.P. Cary, C. Sander, Pathguide: a pathway resource list, no Database issue, Nucleic Acids Res. 34 (2006) D504, https://doi.org/10.1093/nar/gkj126, 506, janv.

[18] F. Cunningham, et al., Ensembl 2022, Nucleic Acids Res. 50 (D1) (2022) D988–D995, https://doi.org/10.1093/nar/gkab1049, janv.

[19] C. Meslin, F. Brimau, P. Nagnan-Le Meillour, I. Callebaut, G. Pascal, P. Monget, The evolutionary history of the SAL1 gene family in eutherian mammals, BMC Evol. Biol. 11 (2011) 148, https://doi.org/10.1186/1471-2148-11-148, mai.

[20] C. Meslin, et al., Evolution of genes involved in gamete interaction: evidence for positive selection, duplications and losses in vertebrates, PLoS One 7 (9) (2012), e44548, https://doi.org/10.1371/journal.pone.0044548.

[21] C. Moros-Nicolás, S. Fouchécourt, G. Goudet, P. Monget, Genes encoding mammalian oviductal proteins involved in fertilization are subjected to gene death and positive selection, J. Mol. Evol. 86 (9) (2018) 655–667, https://doi.org/10.1007/s00239-018-9878-0, déc.

[22] O. Jaillon, et al., Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype, Nature 431 (7011) (2004) 946–957, https://doi.org/10.1038/nature03025, oct.

[23] T. Kon, et al., Single-cell transcriptomics of the goldfish retina reveals genetic divergence in the asymmetrically evolved subgenomes after allotetraploidization, Commun. Biol. 5 (1) (2022) 1404, https://doi.org/10.1038/s42003-022-04351-3, déc.

[24] P. Xu, et al., The allotetraploid origin and asymmetrical genome evolution of the common carp Cyprinus carpio, Nat. Commun. 10 (1) (2019) 4625, https://doi.org/10.1038/s41467-019-12644-1, oct.