Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/5/6/eaav0547/DC1

Supplementary Materials for

De novo assembly of the goldfish (*Carassius auratus*) genome and the evolution of genes after whole-genome duplication

Zelin Chen, Yoshihiro Omori, Sergey Koren, Takuya Shirokiya, Takuo Kuroda, Atsushi Miyamoto, Hironori Wada, Asao Fujiyama, Atsushi Toyoda, Suiyuan Zhang, Tyra G. Wolfsberg, Koichi Kawakami, Adam M. Phillippy, NISC Comparative Sequencing Program, James C. Mullikin, Shawn M. Burgess*

*Corresponding author. Email: burgess@mail.nih.gov

Published 26 June 2019, *Sci. Adv.* **5**, eaav0547 (2019)
DOI: 10.1126/sciadv.aav0547

The PDF file includes:

Other Supplementary Material for this manuscript includes the following:

(available at advances.sciencemag.org/cgi/content/full/5/6/eaav0547/DC1)

NISC consortium members (Microsoft Excel format). List of members of the NISC Comparative Sequencing Program.

Table S1. PacBio read statistics.

	Raw Reads	Corrected Reads
Counts	16,671,136	11,884,085
Mean length (bp)	7,800	6,810
Coverage	~71	~45
Peak length (kbp)	~9.8	~8.0

Table S2. Assembly statistics for different coverage groups.

Read Depth	Contig Counts	bp	N50 (bp)
0-0.6	6,937	497,816,144	114,500
0.6-1.8	2,393	1,347,156,259	1,372,944
>1.8	85	4,078,364	-

Table S3. Repeated DNA statistics.

	Goldfish	Common Carp (*23*)	Zebrafish (*26*)
Total base pairs	721,087,053 (39.6%)	672,246,354 (39.2%)	715,370,858 (52.24%)
DNA transposon	16.38%	17.53%	34.3%
LTR	4.89%	4.35%	5.07%
LINE	4.50%	4.90%	2.83%
SINE	0.47%	0.47%	2.34%
Satellite	1.27%	-	1.78%
RC	1.89%	-	0.94%
Simple	3.27%	-	4.12%
Unknown	6.88%	-	0.34%

The breakdown of the various repeat elements presented in goldfish, common carp, and zebrafish. The percentage of the total genome is indicated in parentheses. The larger fraction of DNA transposons in zebrafish is responsible for its significantly larger size compared to the pre-duplication carp or goldfish genomes.

Table S4. Core eukaryotic genes using BUSCOs.

	Goldfish	Common carp	Zebrafish
Complete BUSCOs	4,204	3,828	4,384
Complete and single-copy BUSCOs	1,990	1,695	4,145
Complete and duplicated BUSCOs	2,214	2,133	239
Fragmented BUSCOs	257	436	113
Missing BUSCOs	123	320	87
Total BUSCO groups searched	4,584	4,584	4,584

Using the “Benchmarking of Universal Single-Copy Orthologs” *Actinopterygii* gene set, we determined the goldfish genome assembly has 97.3% of the BUSCO in at least one copy (91.7% complete BUSCO genes, 5.6% fragmented, and 2.7% missing) with 48.3% complete in both copies, compared to the common carp assembly which has 83.5% complete BUSCO, 9.5% fragmented, 7% missing and 46.5% complete with both gene pairs represented.

Table S5. Statistics of exon gain/loss. ZF: zebrafish ortholog, GF1/2: the two goldfish ohnologs. There are 547 triplets with reciprocal GF1,GF2 exon gain/loss (i.e. both (ZF,GF1)+(GF1)>0 and (ZF,GF2)+(GF2)>0.

	Count	Length	Percent
Total	111,564	27,192,336	100.00
(ZF,GF1,GF2)*	92,484	24,228,374	89.10
(ZF,GF1)	4,563	799,302	2.94
(ZF,GF2)	4,208	735,520	2.70
(ZF)	2,152	305,385	1.12
(GF1,GF2)	7,104	993,528	3.65
(GF1)	527	60,815	0.22
(GF2)	526	69,412	0.26
ZF gain GF1 loss	6,360	1,040,905	3.83
ZF gain GF2 loss	6,715	1,104,687	4.06
ZF gain GF loss	13,075	2,145,592	3.95[#]
ZF gain GF (singleton gain)[$]	8,771	1,534,822	2.82[#]
ZF loss GF gain	15,261	2,117,283	3.89[#]
ZF loss GF gain (singleton gain)[$]	1,053	130,227	0.24[#]

* means present in the species ortholog or ohnologs.

$ singleton gain means gain in either GF1 or GF2, but not both

# percentage of 2*total number of CNEs

Table S6. Statistics of CNE gain/loss. ZF: zebrafish ortholog, GF1/2: the two goldfish ohnologs. There are 1159 triplets with reciprocal GF1,GF2 CNE gain/loss (i.e. both (ZF,GF1)+(GF1)>0 and (ZF,GF2)+(GF2)>0.

	Count	Length	Percent
Total	122,184	20,322,650	100.00
(ZF,GF1,GF2)*	95,413	15,873,505	78.11
(ZF,GF1)	5,003	680,524	3.35
(ZF,GF2)	4,089	518,064	2.55
(ZF)	6,628	1,020,348	5.02
(GF1,GF2)	7,127	1,233,018	6.07
(GF1)	1,831	470,719	2.32
(GF2)	2,093	526,472	2.59
ZF gain GF1 loss	10,717	1,538,412	7.57
ZF gain GF2 loss	11,631	1,700,872	8.37
ZF gain GF loss	22,348	3,239,284	7.97[#]
ZF gain GF (singleton gain)[$]	9,092	1,198,588	2.95[#]
ZF loss GF gain	18,178	3,463,227	8.52[#]
ZF loss GF gain (singleton gain)[$]	3,924	997,191	2.45[#]

* means present in the species ortholog or ohnologs.

$ singleton gain means gain in either GF1 or GF2, but not both

# percentage of 2*total number of CNEs

Table S7. Triplets with different number of coexpressed tissues.

Number of co-expressed tissues	ZF-GF1	ZF-GF2	ZF-GF_pair	GF1-GF2
0	487 (5.74%)	543 (6.40%)	93 (1.10%)	872 (10.28%)
1	1040 (12.26%)	1084 (12.78%)	744 (8.77%)	1191 (14.04%)
2	1069 (12.60%)	1051 (12.39%)	1085 (12.79%)	989 (11.66%)
3	711 (8.38%)	699 (8.24%)	703 (8.29%)	726 (8.56%)
4	765 (9.02%)	784 (9.24%)	705 (8.31%)	799 (9.42%)
5	1752 (20.65%)	1645 (19.39%)	1477 (17.41%)	1808 (21.31%)
6	2659 (31.35%)	2677 (31.56%)	3676 (43.33%)	2098 (24.73%)

Table S8. Number and percentage of ohnolog clusters in evolutionary fate categories.

Fate	GF
Double correlated	4539 (53.51%)
Dosage-correlated	5699 (67.18%)
Double correlated or Dosage-correlated	5700 (67.19%)
Double co-expressed	3506 (41.33%)
Dosage co-expressed	5437 (64.03%)
Sub-functionalization	39 (0.46%)
Neo-functionalization	321 (3.78%)
Non-functionalization	672 (7.92%)
Partial Sub-functionalization	6 (0.07%)
Partial Neo-functionalization	286 (3.37%)
Partial Non-functionalization	1169 (13.78%)

Table S9. Comparison of features between ZF-GF1 and ZF-GF2, where "Mean1" and "Mean2" are the mean of features between ZF-GF1 and ZF-GF2, respectively.

Fate	Feature	Mean1	Mean2	Difference	Wilcox rank test P value
Neo-F	Nucleotide identity (%)	83.54	82.88	0.66	0.6376
	Exon gain/loss (%)	7.39	7.60	-0.21	0.8217
	CNE gain/loss (%)	20.53	20.79	-0.26	0.7522
Non-F	Nucleotide identity (%)	84.10	82.19	1.91	0.0130 *
	Exon gain/loss (%)	7.95	24.78	-16.82	0.0000 ***
	CNE gain/loss (%)	19.10	26.16	-7.06	0.0000 ***

Table S10. Comparison of features between different gene evolutionary fate.

Feature	species pair	Fate1	Fate2	Mean1	Mean2	P value
Nucleotide identity	ZF_GF	coexpress	sub-F	85.91	84.9	0.0132 *
Nucleotide identity	ZF_GF	coexpress	neo-F	85.91	82.88	0 ***
Nucleotide identity	ZF_GF	coexpress	non-F	85.91	82.19	0 ***
Nucleotide identity	ZF_GF	sub-F	neo-F	84.9	82.88	0.8899
Nucleotide identity	ZF_GF	sub-F	non-F	84.9	82.19	0.6525
Nucleotide identity	ZF_GF	neo-F	non-F	82.88	82.19	0.3486
Nucleotide identity	GF1_GF2	coexpress	sub-F	88.65	84.96	0.2456
Nucleotide identity	GF1_GF2	coexpress	neo-F	88.65	86.02	0.186
Nucleotide identity	GF1_GF2	coexpress	non-F	88.65	82.62	0 ***
Nucleotide identity	GF1_GF2	sub-F	neo-F	84.96	86.02	0.5309
Nucleotide identity	GF1_GF2	sub-F	non-F	84.96	82.62	0.7804
Nucleotide identity	GF1_GF2	neo-F	non-F	86.02	82.62	0.0212 *
Exon gain/loss	ZF_GF	coexpress	sub-F	5.38	7.49	0.0085 **
Exon gain/loss	ZF_GF	coexpress	neo-F	5.38	7.64	0.0078 **
Exon gain/loss	ZF_GF	coexpress	non-F	5.38	24.4	0 ***
Exon gain/loss	ZF_GF	sub-F	neo-F	7.49	7.64	0.284
Exon gain/loss	ZF_GF	sub-F	non-F	7.49	24.4	0 ***
Exon gain/loss	ZF_GF	neo-F	non-F	7.64	24.4	0 ***
Exon gain/loss	GF1_GF2	coexpress	sub-F	3.14	3.75	0.1529
Exon gain/loss	GF1_GF2	coexpress	neo-F	3.14	4.29	0.013 *
Exon gain/loss	GF1_GF2	coexpress	non-F	3.14	21.9	0 ***
Exon gain/loss	GF1_GF2	sub-F	neo-F	3.75	4.29	0.6709
Exon gain/loss	GF1_GF2	sub-F	non-F	3.75	21.9	0 ***
Exon gain/loss	GF1_GF2	neo-F	non-F	4.29	21.9	0 ***
CNE gain/loss	ZF_GF	coexpress	sub-F	18.19	20.51	0.4655
CNE gain/loss	ZF_GF	coexpress	neo-F	18.19	20.79	0.0168 *

CNE gain/loss	ZF_GF	coexpress	non-F	18.19	26.16	0.0001 ***
CNE gain/loss	ZF_GF	sub-F	neo-F	20.51	20.79	0.7124
CNE gain/loss	ZF_GF	sub-F	non-F	20.51	26.16	0.5096
CNE gain/loss	ZF_GF	neo-F	non-F	20.79	26.16	0.5804
CNE gain/loss	GF1_GF2	coexpress	sub-F	12.01	12.93	0.6682
CNE gain/loss	GF1_GF2	coexpress	neo-F	12.01	12.86	0.1718
CNE gain/loss	GF1_GF2	coexpress	non-F	12.01	17.46	0.0989 *
CNE gain/loss	GF1_GF2	sub-F	neo-F	12.93	12.86	0.9502
CNE gain/loss	GF1_GF2	sub-F	non-F	12.93	17.46	0.9443
CNE gain/loss	GF1_GF2	neo-F	non-F	12.86	17.46	0.9912

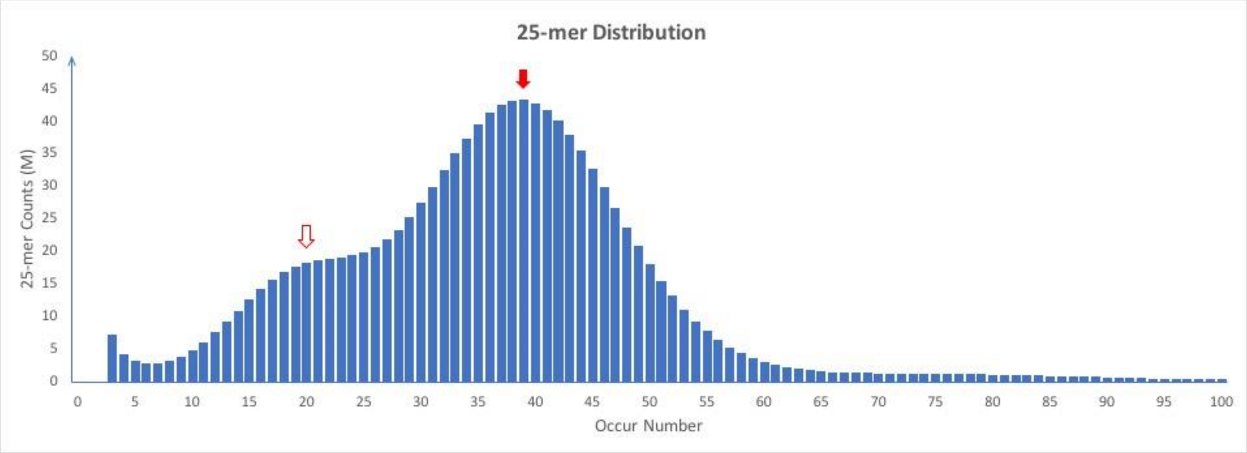


Fig. S1. Twenty-five–nucleotide oligomer occurrence distribution from 2 × 125 bp Illumina paired-end reads. The two peaks indicate that a fraction of the genome was not sequenced to the same depth of coverage, i.e. part of the genome (approximately 16% from The Canu assembly) was at 20X coverage instead of 40X (white arrow *vs*. red arrow). The 20X peak was indicative of regions of the genome that were not homozygous.

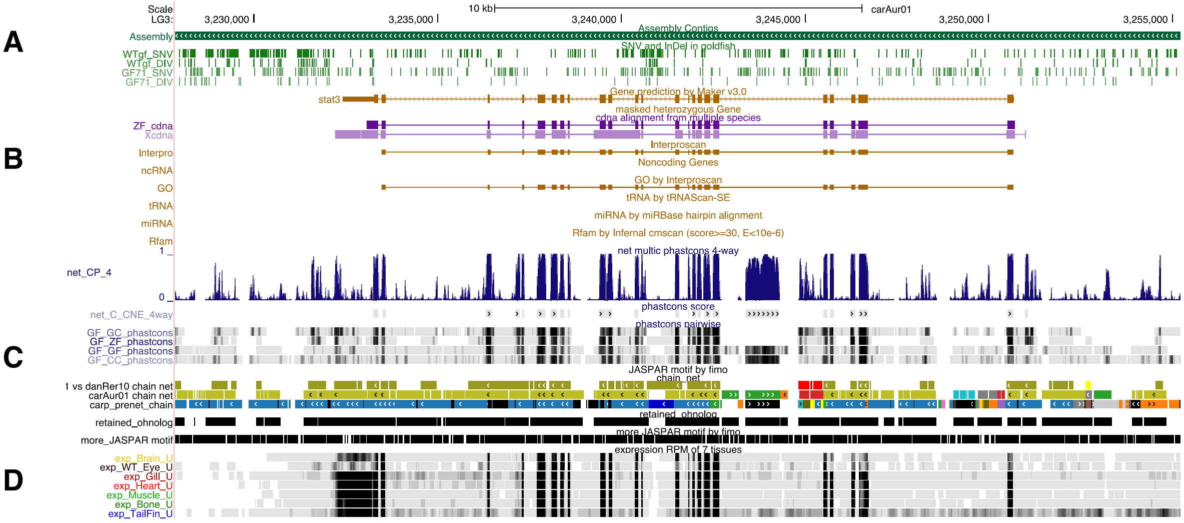


Fig. S2. Screenshot of the UCSC Genome Browser implementation of the carAur01 assembly. Genome annotation includes: A) Assembly, SNV and DIV data from sequencing three “wild-type” Wakin goldfish, B) gene model annotation C) multiple genome alignment tracks that compare goldfish to zebrafish, grass carp, and common carp to identify conserved coding and non-coding (i.e. enhancers/promoters) sequences, D) gene expression from 7 adult goldfish tissues. Hub available at: https://research.nhgri.nih.gov/goldfish/

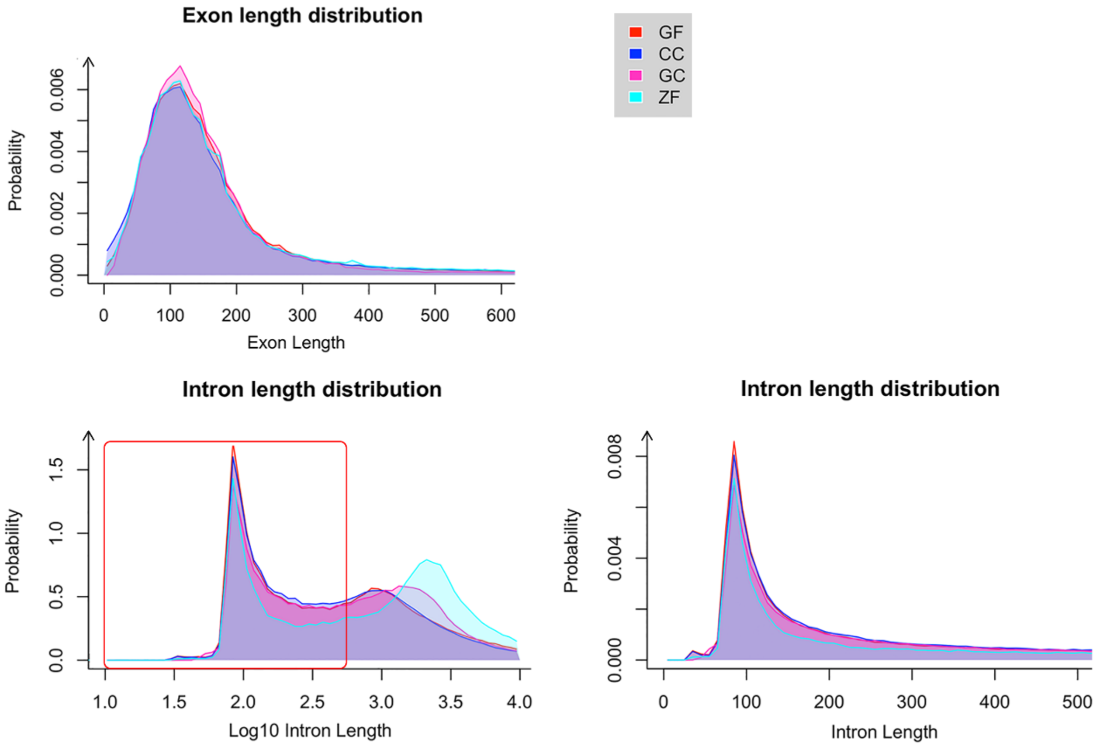
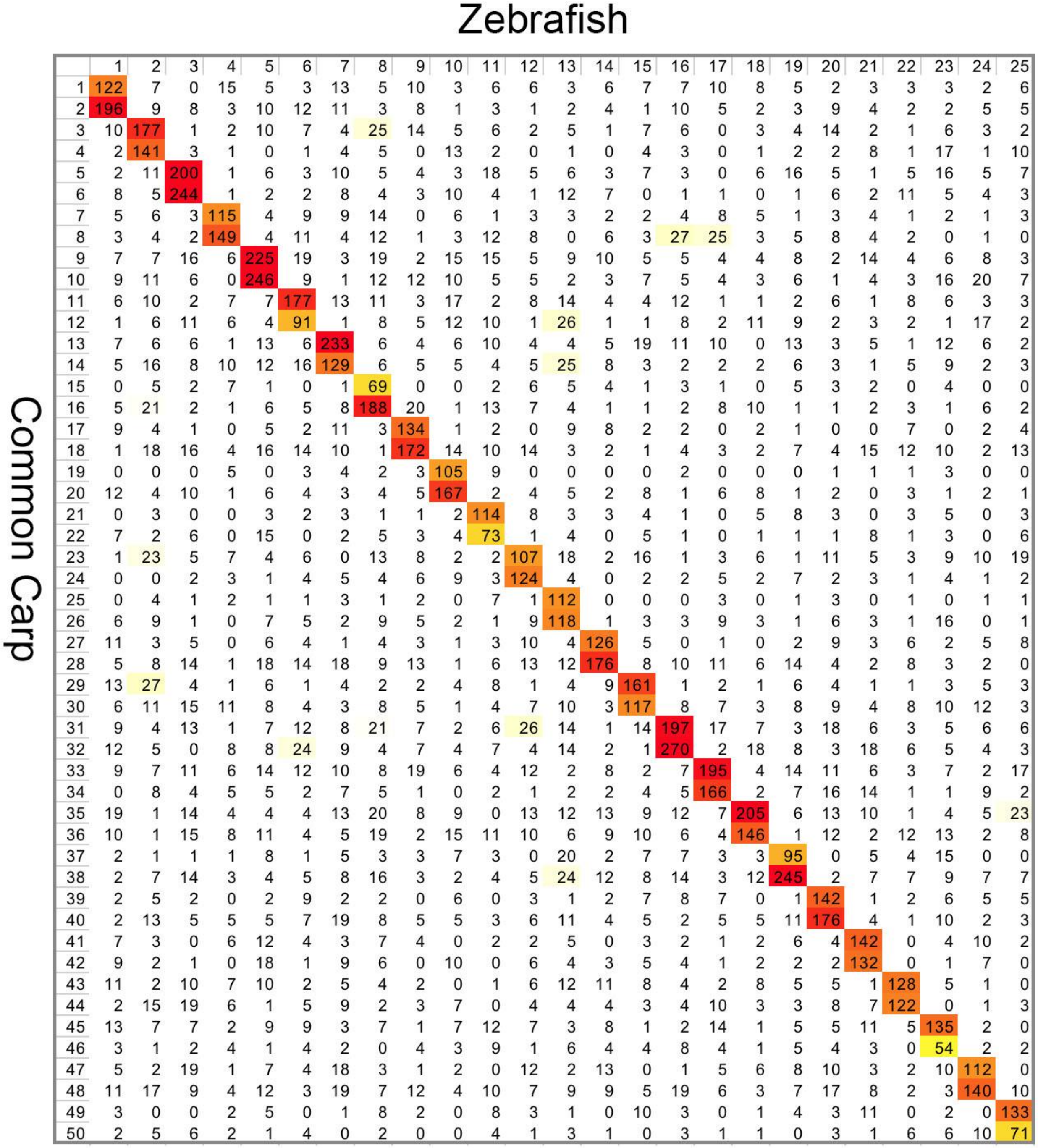
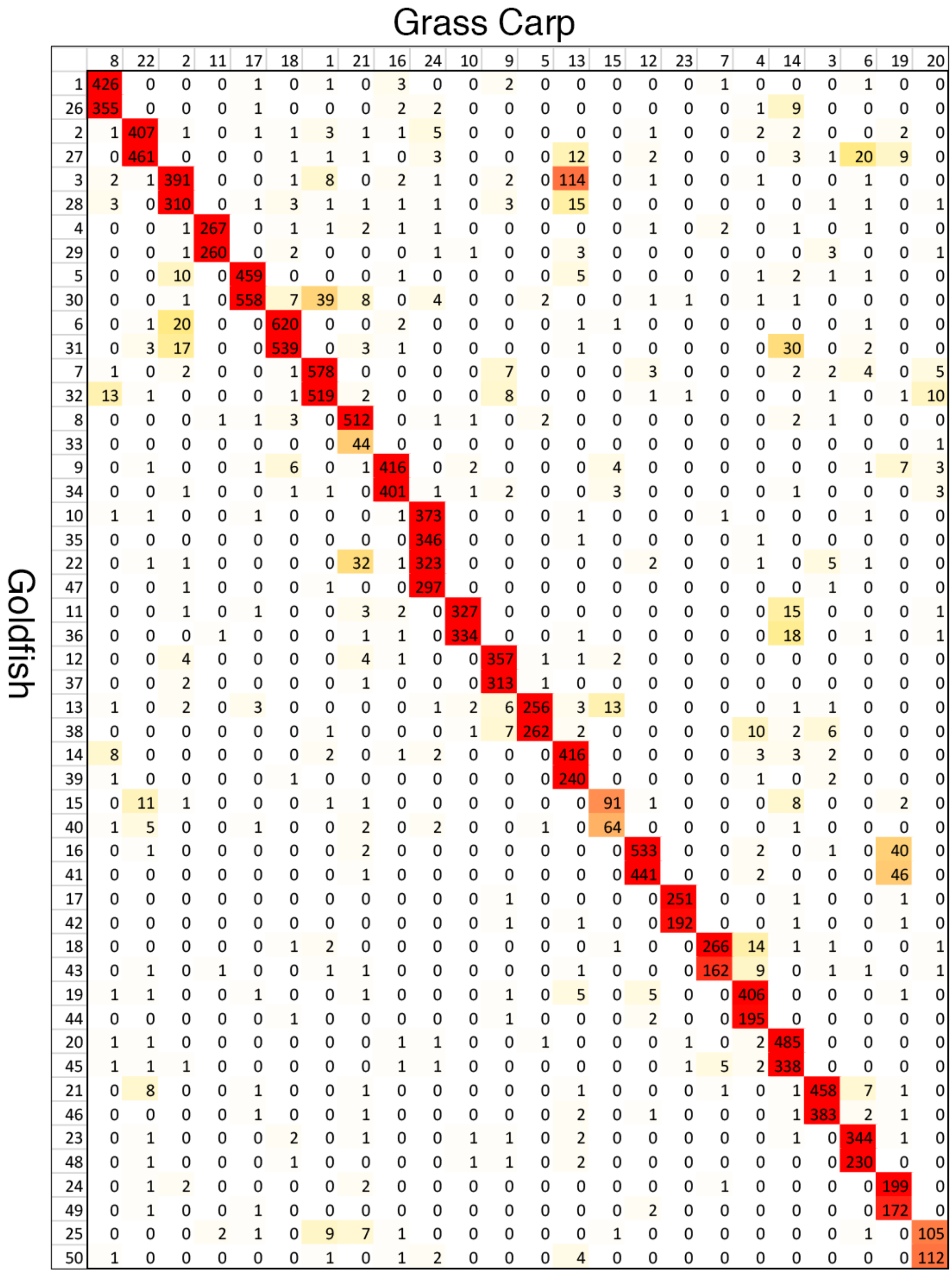


Fig. S3. Distribution of exon and intron lengths. Bottom right panel is an enlargement of the red box in the bottom left panel. GF: goldfish, CC: common carp, GC: grass carp, ZF: zebrafish.


Zebrafish
Common Carp

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	122	7	0	15	5	3	13	5	10	3	6	6	3	6	7	7	10	8	5	2	3	3	3	2	6
2	196	9	8	3	10	12	11	3	8	1	3	1	2	4	1	10	5	2	3	9	4	2	2	5	5
3	10	177	1	2	10	7	4	25	14	5	6	2	5	1	7	6	0	3	4	14	2	1	6	3	2
4	2	141	3	1	0	1	4	5	0	13	2	0	1	0	4	3	0	1	2	2	8	1	17	1	10
5	2	11	200	1	6	3	10	5	4	3	18	5	6	3	7	3	0	6	16	5	1	5	16	5	7
6	8	5	244	1	2	2	8	4	3	10	4	1	12	7	0	1	1	0	1	6	2	11	5	4	3
7	5	6	3	115	4	9	9	14	0	6	1	3	3	2	2	4	8	5	1	3	4	1	2	1	3
8	3	4	2	149	4	11	4	12	1	3	12	8	0	6	3	27	25	3	5	8	4	2	0	1	0
9	7	7	16	6	225	19	3	19	2	15	15	5	9	10	5	5	4	4	8	2	14	4	6	8	3
10	9	11	6	0	246	9	1	12	12	10	5	5	2	3	7	5	4	3	6	1	4	3	16	20	7
11	6	10	2	7	7	177	13	11	3	17	2	8	14	4	4	12	1	1	2	6	1	8	6	3	3
12	1	6	11	6	4	91	1	8	5	12	10	1	26	1	1	8	2	11	9	2	3	2	1	17	2
13	7	6	6	1	13	6	233	6	4	6	10	4	4	5	19	11	10	0	13	3	5	1	12	6	2
14	5	16	8	10	12	16	129	6	5	5	4	5	25	8	3	2	2	2	6	3	1	5	9	2	3
15	0	5	2	7	1	0	1	69	0	0	2	6	5	4	1	3	1	0	5	3	2	0	4	0	0
16	5	21	2	1	6	5	8	188	20	1	13	7	4	1	1	2	8	10	1	1	2	3	1	6	2
17	9	4	1	0	5	2	11	3	134	1	2	0	9	8	2	2	0	2	1	0	0	7	0	2	4
18	1	18	16	4	16	14	10	1	172	14	10	14	3	2	1	4	3	2	7	4	15	12	10	2	13
19	0	0	0	5	0	3	4	2	3	105	9	0	0	0	0	2	0	0	0	1	1	1	3	0	0
20	12	4	10	1	6	4	3	4	5	167	2	4	5	2	8	1	6	8	1	2	0	3	1	2	1
21	0	3	0	0	3	2	3	1	1	2	114	8	3	3	4	1	0	5	8	3	0	3	5	0	3
22	7	2	6	0	15	0	2	5	3	4	73	1	4	0	5	1	0	1	1	1	8	1	3	0	6
23	1	23	5	7	4	6	0	13	8	2	2	107	18	2	16	1	3	6	1	11	5	3	9	10	19
24	0	0	2	3	1	4	5	4	6	9	3	124	4	0	2	2	5	2	7	2	3	1	4	1	2
25	0	4	1	2	1	1	3	1	2	0	7	1	112	0	0	0	3	0	1	3	0	1	0	1	1
26	6	9	1	0	7	5	2	9	5	2	1	9	118	1	3	3	9	3	1	6	3	1	16	0	1
27	11	3	5	0	6	4	1	4	3	1	3	10	4	126	5	0	1	0	2	9	3	6	2	5	8
28	5	8	14	1	18	14	18	9	13	1	6	13	12	176	8	10	11	6	14	4	2	8	3	2	0
29	13	27	4	1	6	1	4	2	2	4	8	1	4	9	161	1	2	1	6	4	1	1	3	5	3
30	6	11	15	11	8	4	3	8	5	1	4	7	10	3	117	8	7	3	8	9	4	8	10	12	3
31	9	4	13	1	7	12	8	21	7	2	6	26	14	1	14	197	17	7	3	18	6	3	5	6	6
32	12	5	0	8	8	24	9	4	7	4	7	4	14	2	1	270	2	18	8	3	18	6	5	4	3
33	9	7	11	6	14	12	10	8	19	6	4	12	2	8	2	7	195	4	14	11	6	3	7	2	17
34	0	8	4	5	5	2	7	5	1	0	2	1	2	2	4	5	166	2	7	16	14	1	1	9	2
35	19	1	14	4	4	4	13	20	8	9	0	13	12	13	9	12	7	205	6	13	10	1	4	5	23
36	10	1	15	8	11	4	5	19	2	15	11	10	6	9	10	6	4	146	1	12	2	12	13	2	8
37	2	1	1	1	8	1	5	3	3	7	3	0	20	2	7	7	3	3	95	0	5	4	15	0	0
38	2	7	14	3	4	5	8	16	3	2	4	5	24	12	8	14	3	12	245	2	7	7	9	7	7
39	2	5	2	0	2	9	2	2	0	6	0	3	1	2	7	8	7	0	1	142	1	2	6	5	5
40	2	13	5	5	5	7	19	8	5	5	3	6	11	4	5	2	5	5	11	176	4	1	10	2	3
41	7	3	0	6	12	4	3	7	4	0	2	2	5	0	3	2	1	2	6	4	142	0	4	10	2
42	9	2	1	0	18	1	9	6	0	10	0	6	4	3	5	4	1	2	2	2	132	0	1	7	0
43	11	2	10	7	10	2	5	4	2	0	1	6	12	11	8	4	2	8	5	5	1	128	5	1	0
44	2	15	19	6	1	5	9	2	3	7	0	4	4	4	3	4	10	3	3	8	7	122	0	1	3
45	13	7	7	2	9	9	3	7	1	7	12	7	3	8	1	2	14	1	5	5	11	5	135	2	0
46	3	1	2	4	1	4	2	0	4	3	9	1	6	4	4	8	4	1	5	4	3	0	54	2	2
47	5	2	19	1	7	4	18	3	1	2	0	12	2	13	0	1	5	6	8	10	3	2	10	112	0
48	11	17	9	4	12	3	19	7	12	4	10	7	9	9	5	19	6	3	7	17	8	2	3	140	10
49	3	0	0	2	5	0	1	8	2	0	8	3	1	0	10	3	0	1	4	3	11	0	2	0	133
50	2	5	6	2	1	4	0	2	0	0	4	1	3	1	0	3	1	1	0	3	1	6	6	10	71

Fig. S4. RBH gene counts between zebrafish and common carp chromosomes. Red to yellow indicates high to low numbers.

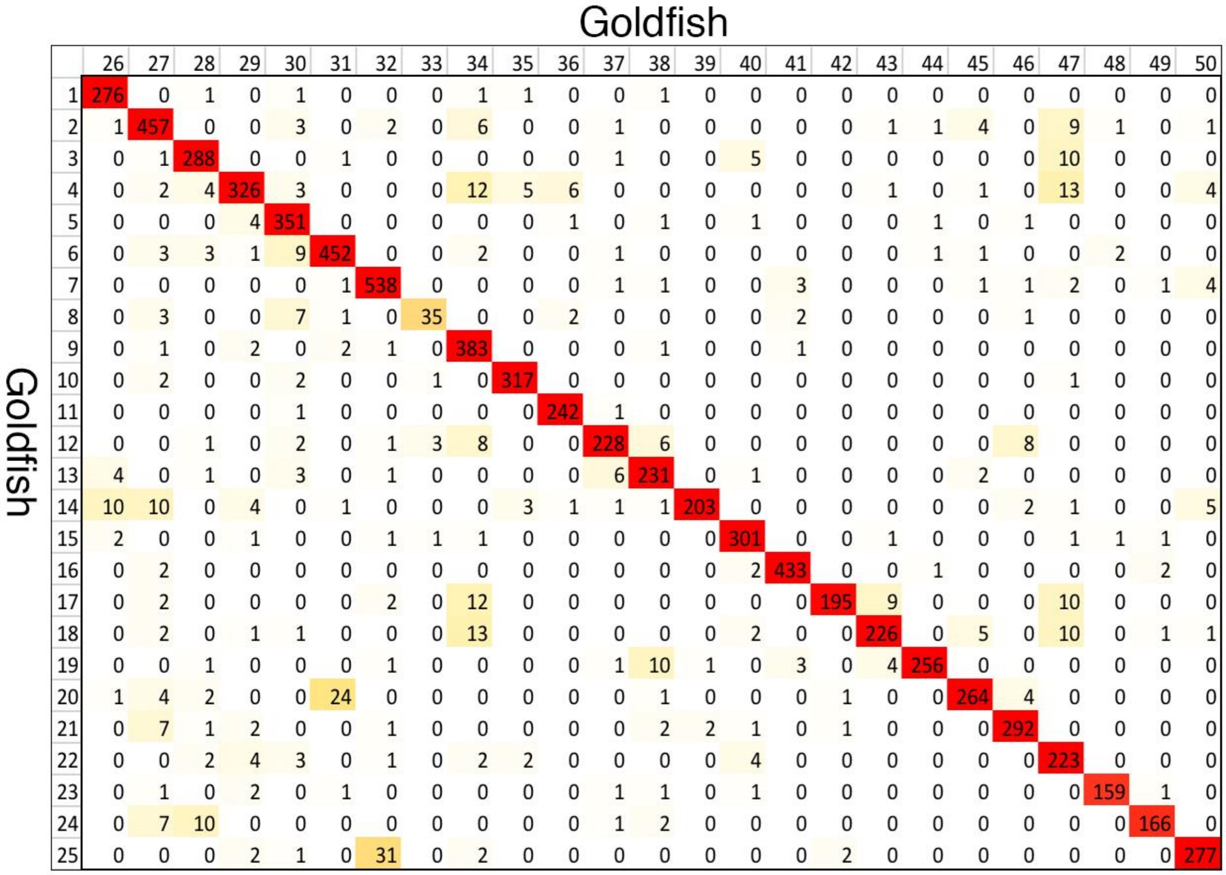


Grass Carp

Goldfish

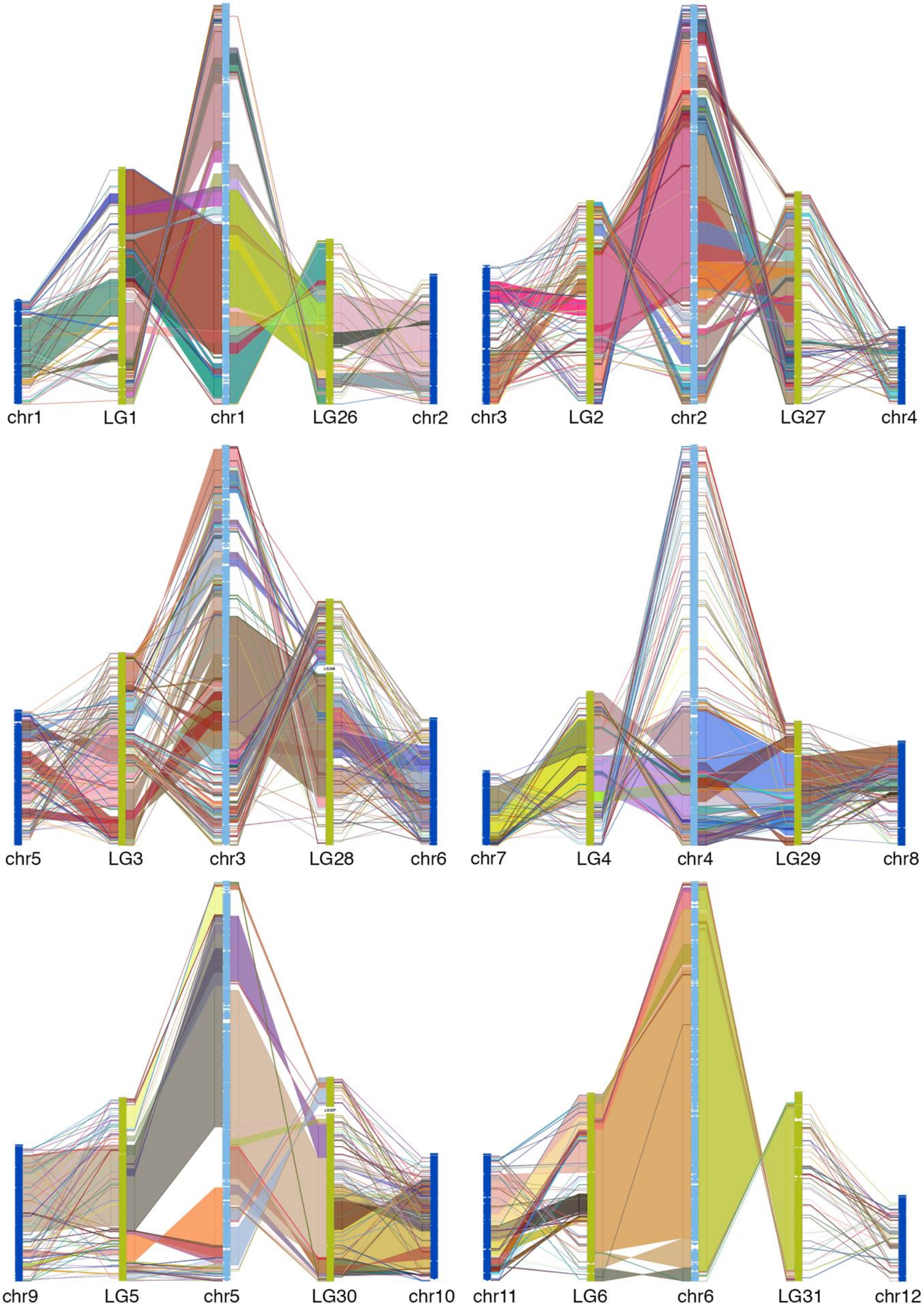
	8	22	2	11	17	18	1	21	16	24	10	9	5	13	15	12	23	7	4	14	3	6	19	20
1	426	0	0	0	1	0	1	0	3	0	0	2	0	0	0	0	0	1	0	0	0	1	0	0
26	355	0	0	0	1	0	0	0	2	2	0	0	0	0	0	0	0	0	1	9	0	0	0	0
2	1	407	1	0	1	1	3	1	1	5	0	0	0	0	0	1	0	0	2	2	0	0	2	0
27	0	461	0	0	0	1	1	1	0	3	0	0	0	12	0	2	0	0	0	3	1	20	9	0
3	2	1	391	0	0	1	8	0	2	1	0	2	0	114	0	1	0	0	1	0	0	1	0	0
28	3	0	310	0	1	3	1	1	1	1	0	3	0	15	0	0	0	0	0	0	1	1	0	1
4	0	0	1	267	0	1	1	2	1	1	0	0	0	0	0	1	0	2	0	1	0	1	0	0
29	0	0	1	260	0	2	0	0	0	1	1	0	0	3	0	0	0	0	0	0	3	0	0	1
5	0	0	10	0	459	0	0	0	1	0	0	0	0	5	0	0	0	0	1	2	1	1	0	0
30	0	0	1	0	558	7	39	8	0	4	0	0	2	0	0	1	1	0	1	1	0	0	0	0
6	0	1	20	0	0	620	0	0	2	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0
31	0	3	17	0	0	539	0	3	1	0	0	0	0	1	0	0	0	0	0	30	0	2	0	0
7	1	0	2	0	0	1	578	0	0	0	0	7	0	0	0	3	0	0	0	2	2	4	0	5
32	13	1	0	0	0	1	519	2	0	0	0	8	0	0	0	1	1	0	0	0	1	0	1	10
8	0	0	0	1	1	3	0	512	0	1	1	0	2	0	0	0	0	0	0	2	1	0	0	0
33	0	0	0	0	0	0	0	44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
9	0	1	0	0	1	6	0	1	416	0	2	0	0	0	4	0	0	0	0	0	0	1	7	3
34	0	0	1	0	0	1	1	0	401	1	1	2	0	0	3	0	0	0	0	1	0	0	0	3
10	1	1	0	0	1	0	0	0	1	373	0	0	0	1	0	0	0	1	0	0	0	1	0	0
35	0	0	0	0	0	0	0	0	0	346	0	0	0	1	0	0	0	0	1	0	0	0	0	0
22	0	1	1	0	0	0	0	32	1	323	0	0	0	0	0	2	0	0	1	0	5	1	0	0
47	0	0	1	0	0	0	1	0	0	297	0	0	0	0	0	0	0	0	0	0	1	0	0	0
11	0	0	1	0	1	0	0	3	2	0	327	0	0	0	0	0	0	0	0	15	0	0	0	1
36	0	0	0	1	0	0	0	1	1	0	334	0	0	1	0	0	0	0	0	18	0	1	0	1
12	0	0	4	0	0	0	0	4	1	0	0	357	1	1	2	0	0	0	0	0	0	0	0	0
37	0	0	2	0	0	0	0	1	0	0	0	313	1	0	0	0	0	0	0	0	0	0	0	0
13	1	0	2	0	3	0	0	0	0	1	2	6	256	3	13	0	0	0	0	1	1	0	0	0
38	0	0	0	0	0	0	1	0	0	0	1	7	262	2	0	0	0	0	10	2	6	0	0	0
14	8	0	0	0	0	0	2	0	1	2	0	0	0	416	0	0	0	0	3	3	2	0	0	0
39	1	0	0	0	0	1	0	0	0	0	0	0	0	240	0	0	0	0	1	0	2	0	0	0
15	0	11	1	0	0	0	1	1	0	0	0	0	0	0	91	1	0	0	0	8	0	0	2	0
40	1	5	0	0	1	0	0	2	0	2	0	0	1	0	64	0	0	0	0	1	0	0	0	0
16	0	1	0	0	0	0	0	2	0	0	0	0	0	0	0	533	0	0	2	0	1	0	40	0
41	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	441	0	0	2	0	0	0	46	0
17	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	251	0	0	1	0	0	1	0
42	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	192	0	0	1	0	0	1	0
18	0	0	0	0	0	1	2	0	0	0	0	0	0	0	1	0	0	266	14	1	1	0	0	1
43	0	1	0	1	0	0	1	1	0	0	0	0	0	1	0	0	0	162	9	0	1	1	0	1
19	1	1	0	0	1	0	0	1	0	0	0	1	0	5	0	5	0	0	406	0	0	0	1	0
44	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	2	0	0	195	0	0	0	0	0
20	1	1	0	0	0	0	0	0	1	1	0	0	1	0	0	0	1	0	2	485	0	0	0	0
45	1	1	1	0	0	0	0	0	1	1	0	0	0	0	0	0	1	5	2	338	0	0	0	0
21	0	8	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	1	0	1	458	7	1	0
46	0	0	0	0	1	0	0	1	0	0	0	0	0	2	0	1	0	0	0	1	383	2	1	0
23	0	1	0	0	0	2	0	1	0	0	1	1	0	2	0	0	0	0	0	1	0	344	1	0
48	0	1	0	0	0	1	0	0	0	0	1	1	0	2	0	0	0	0	0	0	0	230	0	0
24	0	1	2	0	0	0	0	2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	199	0
49	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	172	0
25	0	0	0	2	1	0	9	7	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	105
50	1	0	0	0	0	0	1	0	1	2	0	0	0	4	0	0	0	0	0	0	0	0	0	112

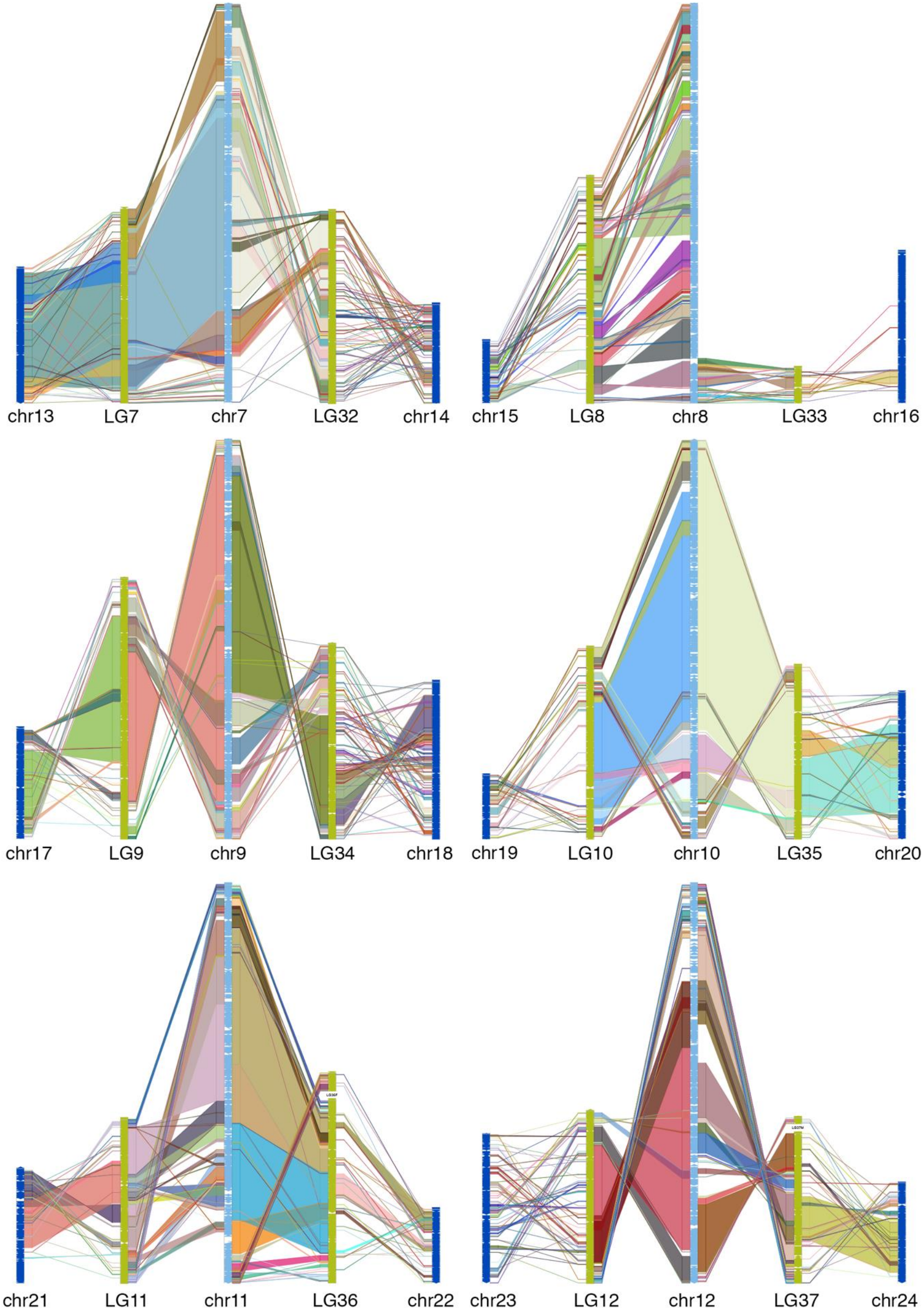
Fig. S5. RBH gene counts between grass carp and goldfish chromosomes. Red to yellow indicates high to low numbers

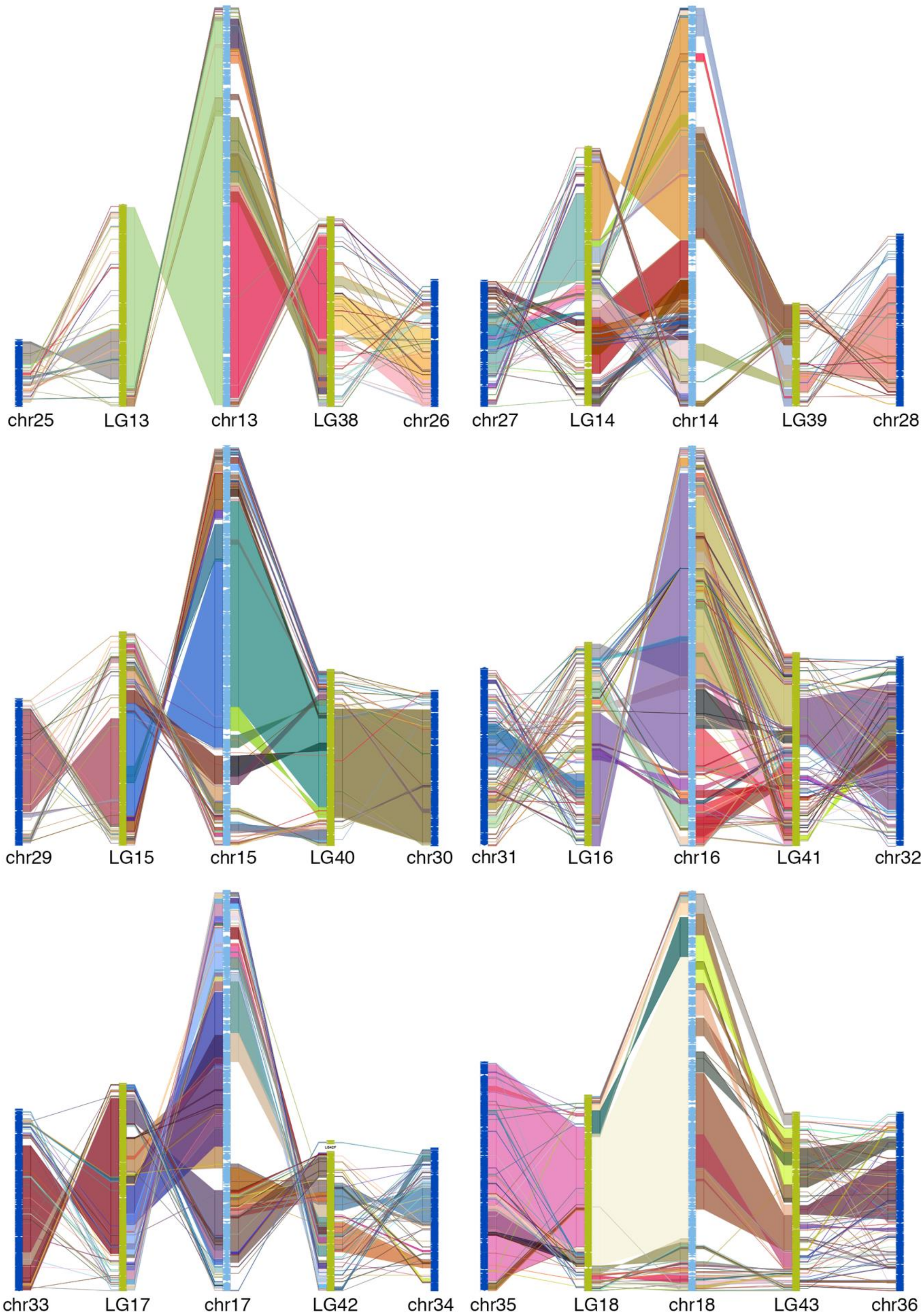

Goldfish
Goldfish

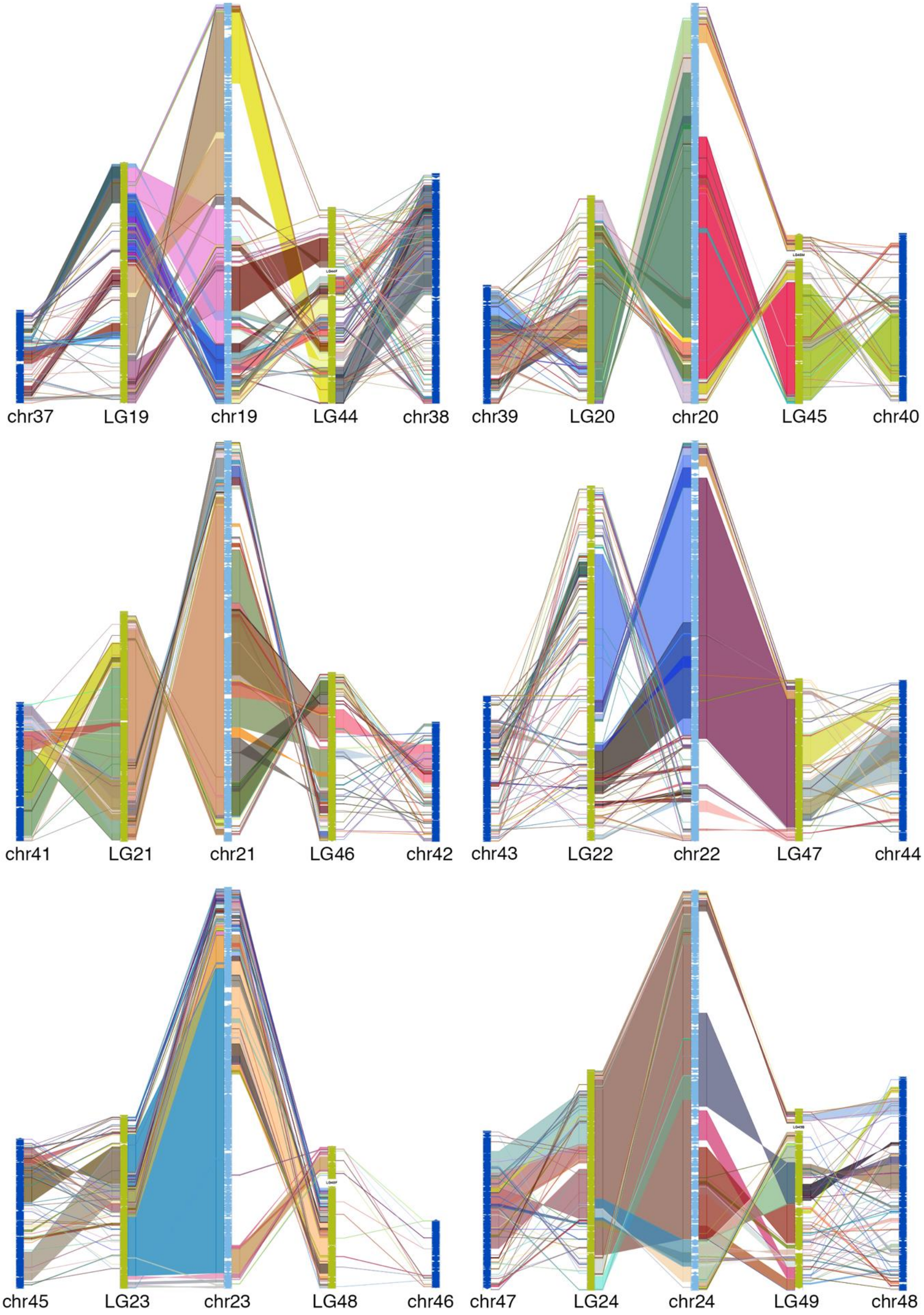
	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
1	276	0	1	0	1	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
2	1	457	0	0	3	0	2	0	6	0	0	1	0	0	0	0	0	1	1	4	0	9	1	0	1
3	0	1	288	0	0	1	0	0	0	0	0	1	0	0	5	0	0	0	0	0	0	10	0	0	0
4	0	2	4	326	3	0	0	0	12	5	6	0	0	0	0	0	0	1	0	1	0	13	0	0	4
5	0	0	0	4	351	0	0	0	0	0	1	0	1	0	1	0	0	0	1	0	1	0	0	0	0
6	0	3	3	1	9	452	0	0	2	0	0	1	0	0	0	0	0	0	1	1	0	0	2	0	0
7	0	0	0	0	0	1	538	0	0	0	0	1	1	0	0	3	0	0	0	1	1	2	0	1	4
8	0	3	0	0	7	1	0	35	0	0	2	0	0	0	0	2	0	0	0	0	1	0	0	0	0
9	0	1	0	2	0	2	1	0	383	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0
10	0	2	0	0	2	0	0	1	0	317	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
11	0	0	0	0	1	0	0	0	0	0	242	1	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	1	0	2	0	1	3	8	0	0	228	6	0	0	0	0	0	0	0	8	0	0	0	0
13	4	0	1	0	3	0	1	0	0	0	0	6	231	0	1	0	0	0	0	2	0	0	0	0	0
14	10	10	0	4	0	1	0	0	0	3	1	1	1	203	0	0	0	0	0	0	2	1	0	0	5
15	2	0	0	1	0	0	1	1	1	0	0	0	0	0	301	0	0	1	0	0	0	1	1	1	0
16	0	2	0	0	0	0	0	0	0	0	0	0	0	0	2	433	0	0	1	0	0	0	0	2	0
17	0	2	0	0	0	0	2	0	12	0	0	0	0	0	0	0	195	9	0	0	0	10	0	0	0
18	0	2	0	1	1	0	0	0	13	0	0	0	0	0	2	0	0	226	0	5	0	10	0	1	1
19	0	0	1	0	0	0	1	0	0	0	0	1	10	1	0	3	0	4	256	0	0	0	0	0	0
20	1	4	2	0	0	24	0	0	0	0	0	0	1	0	0	0	1	0	0	264	4	0	0	0	0
21	0	7	1	2	0	0	1	0	0	0	0	0	2	2	1	0	1	0	0	0	292	0	0	0	0
22	0	0	2	4	3	0	1	0	2	2	0	0	0	0	4	0	0	0	0	0	0	223	0	0	0
23	0	1	0	2	0	1	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	159	1	0
24	0	7	10	0	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0	0	0	0	0	166	0
25	0	0	0	2	1	0	31	0	2	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	277

Fig. S6. RBH gene counts between goldfish whole-genome duplicated chromosomes. Each row or column is one chromosome. Red to yellow indicates high to low numbers.


chr1
LG1
chr1
LG26
chr2
chr3
LG2
chr2
LG27
chr4
chr5
LG3
chr3
LG28
chr6
chr7
LG4
chr4
LG29
chr8
chr9
LG5
chr5
LG30
chr10
chr11
LG6
chr6
LG31
chr12


chr13
LG7
chr7
LG32
chr14
chr15
LG8
chr8
LG33
chr16
chr17
LG9
chr9
LG34
chr18
chr19
LG10
chr10
LG35
chr20
LG36F
chr21
LG11
chr11
LG36
chr22
LG37M
chr23
LG12
chr12
LG37
chr24


chr25
LG13
chr13
LG38
chr26
chr27
LG14
chr14
LG39
chr28
chr29
LG15
chr15
LG40
chr30
chr31
LG16
chr16
LG41
chr32
chr33
LG17
chr17
LG42
chr34
chr35
LG18
chr18
LG43
chr36


chr37
LG19
chr19
LG44F
LG44
chr38
chr39
LG20
chr20
LG45M
LG45
chr40
chr41
LG21
chr21
LG46
chr42
chr43
LG22
chr22
LG47
chr44
chr45
LG23
chr23
LG48F
LG48
chr46
chr47
LG24
chr24
LG49B
LG49
chr48

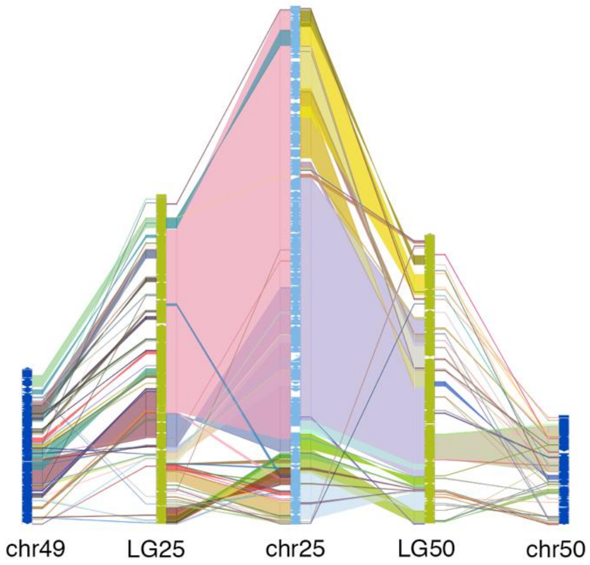


Fig. S7. Chain-net alignment between each zebrafish chromosome (middle light blue bars) and two corresponding whole-genome duplicated goldfish chromosomes (green bars), and goldfish to common carp (blue bars). Lines or blocks between bars show alignments between the two chromosomes. Typically one of goldfish chromosome pairs contained a significantly larger block of conserved col linearity than the other, but both chromosomes show remarkable stability across 60 million years of evolution.

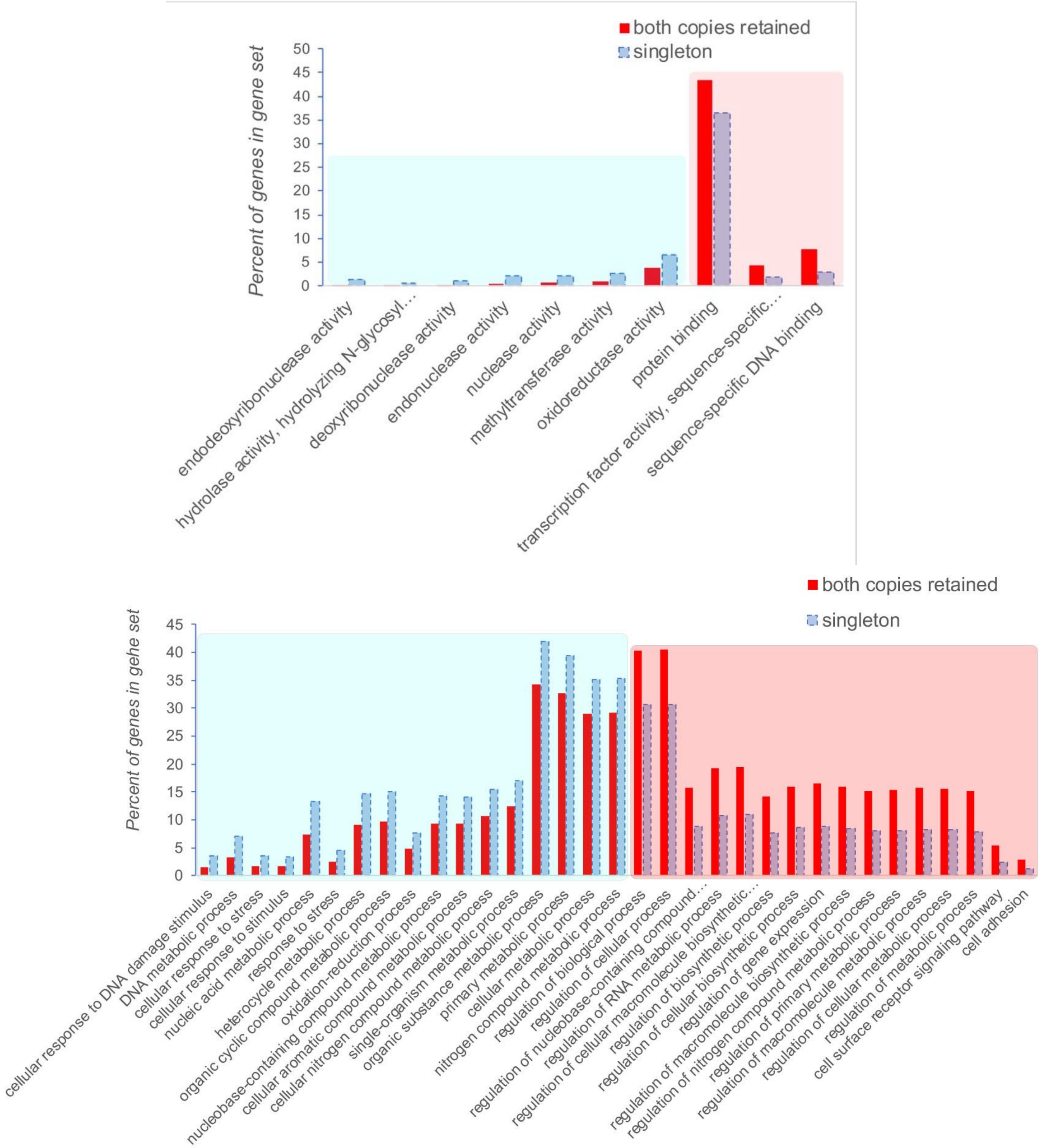


Fig. S8. GO terms prone to retaining both gene copies (blue rectangle) or losing one copy (blue rectangle) after WGD in goldfish. Zebrafish was used as the reference genome (FDR<0.01). Upper: GO molecular functions. Lower: GO biological processes. “Percent of genes in gene set” describes how many genes in each class (both preserved or one copy lost) fall into each GO term, i.e. are some genes in each class over-represented (more likely or less likely to be lost) compared to neutral.

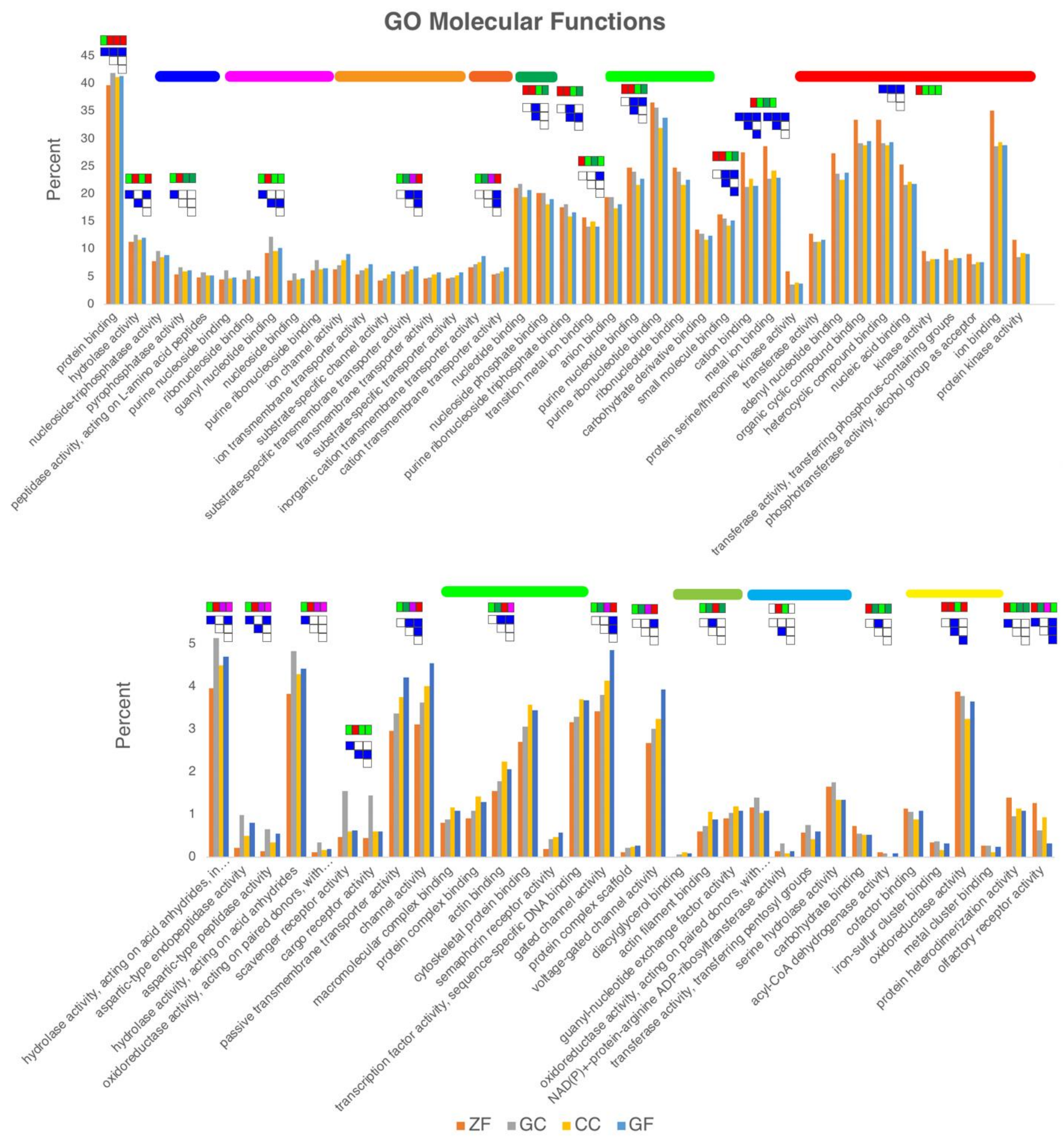


Fig. S9. GO molecular function comparison among zebrafish (ZF), grass carp (GC), common carp (CC), goldfish (GF). The histogram shows the percentage of genes in the gene set. The four colored boxes indicate the relative values among the four species, green for low, red for high, pink, purple or dark green show middle values from higher to lower. The blue or white matrix indicates pair-wise significant values, blue for significant (p-value<0.01 and FDR<0.1), white for non-significant. Color bars indicate clusters with similar trends among the four species.

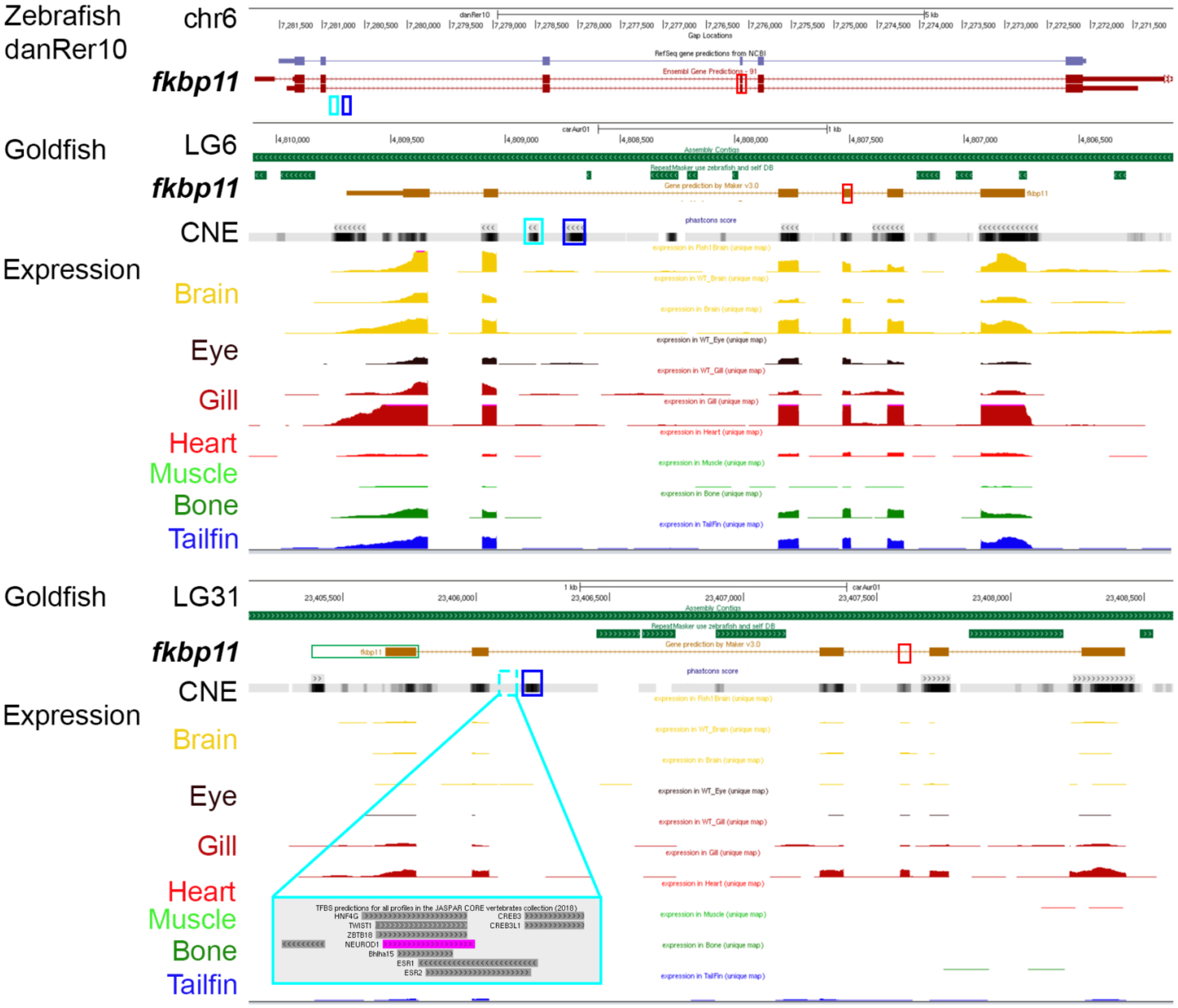


Fig. S10. Example of neo-F. Screenshot example of the fkbp11 gene containing conserved, non-coding elements on linkage group 6, which is mix of non-functionalization and neo-functionalization. The “4-way conservation” peaks are from comparing goldfish, zebrafish, common carp and grass carp, gray bars beneath the peaks are regions satisfying the criteria for CNE. The GF to GF track shows sequences conserved in both chromosomal duplicates. The red dotted box shows the missing sequences on the matching duplicated chromosome (LG31). The remaining tracks are the RNA-seq data from each tissue, showing strong expression in brain, eye, gill, bone, and tail fin, with weaker expression in the muscle and heart. The region on LG31 containing the second copy of *fkbp11*. The red box shows where the missing CNE should be. Expression levels for most of the tissues is very low with the exception of expression in the gill. **c.** Zebrafish *fkbp11* showing the 4-way conservation peaks and the BLAT hit using the goldfish sequences from LG6 (red arrow). **d.** Magnified view of the zebrafish CNE (upper) and goldfish CNE (lower) including JASPAR-predicted transcription factor binding sites. Red arrow marks a highly conserved neurod1 site, a potentially strong enhancer for brain and eye expression.

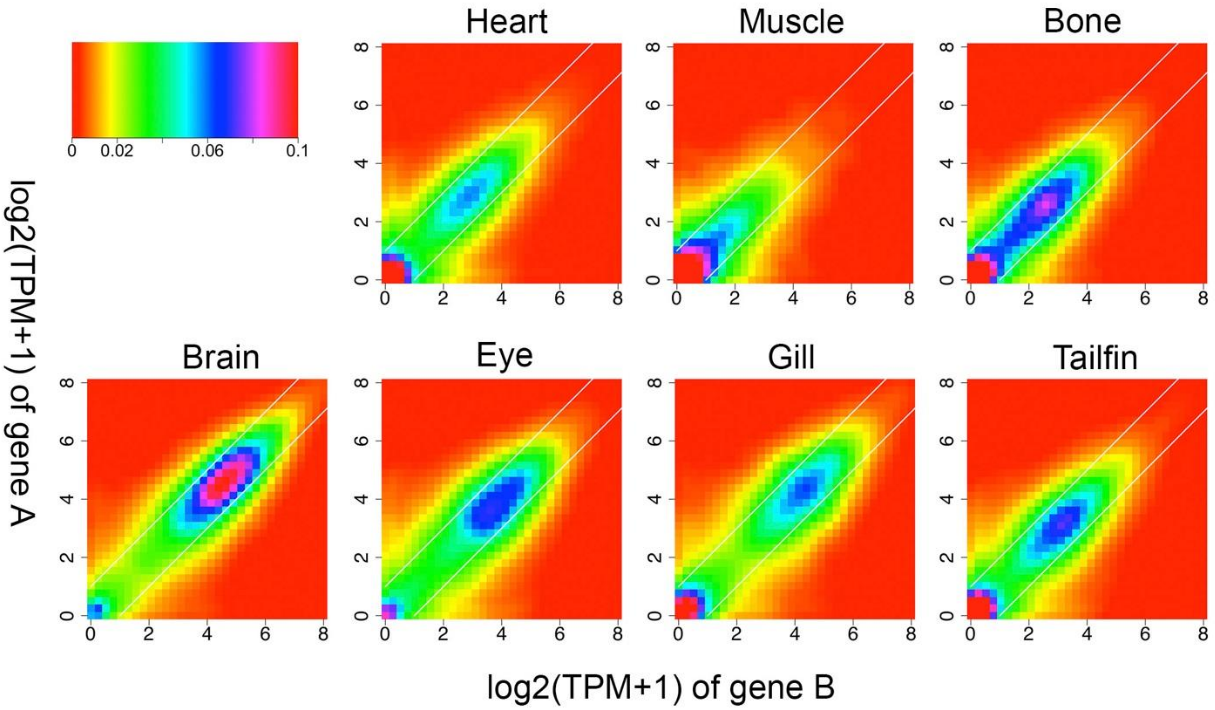
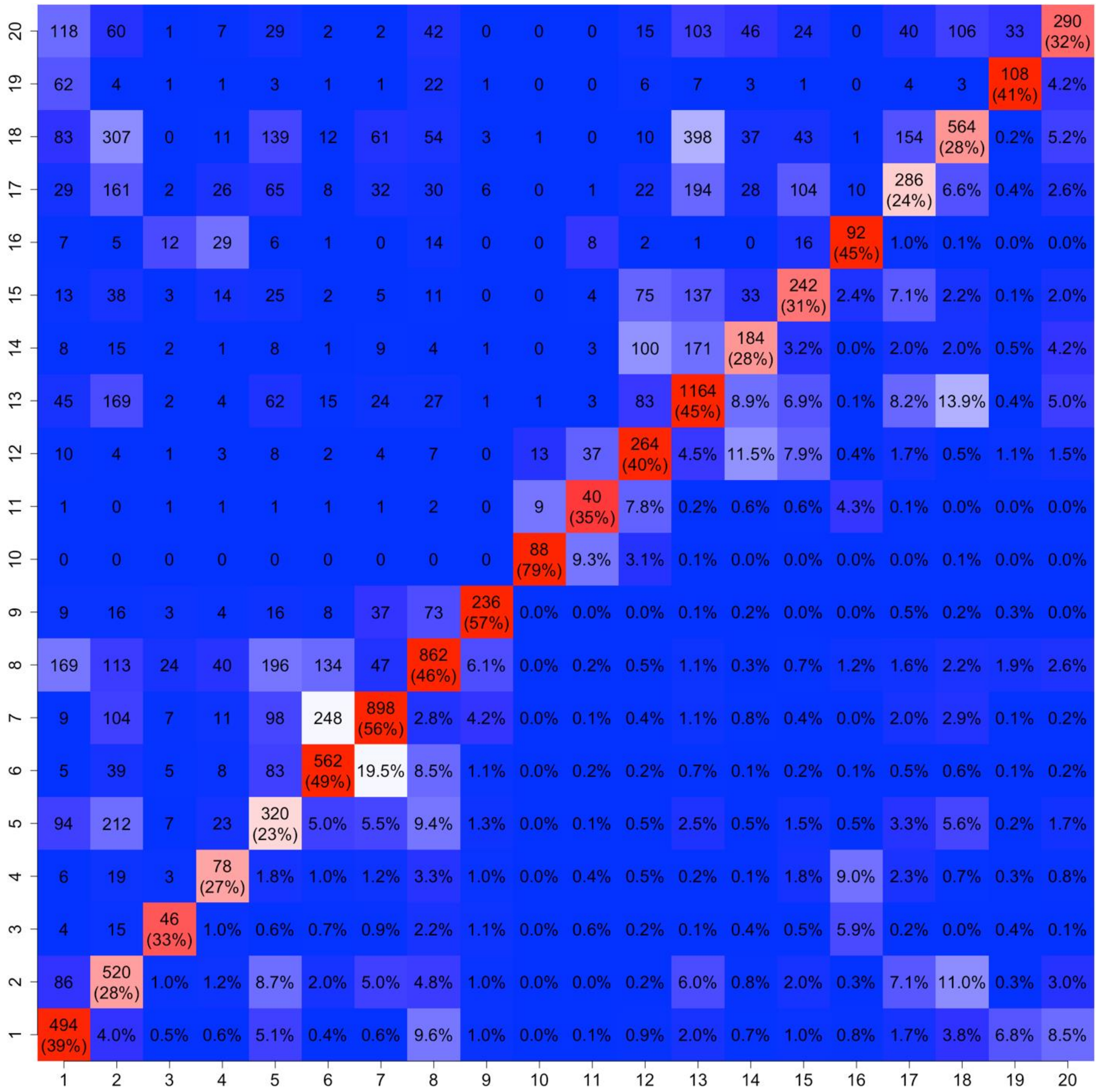
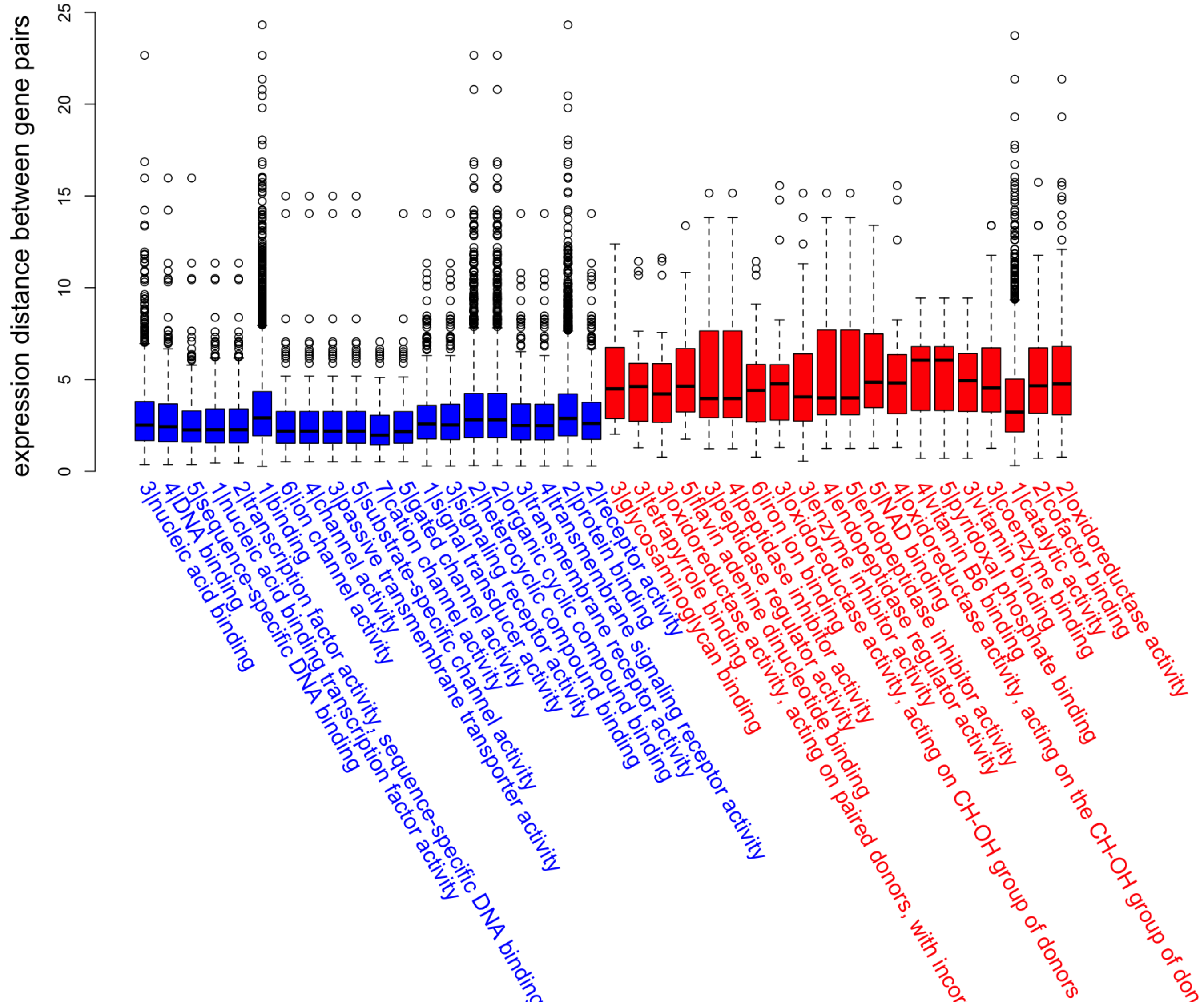


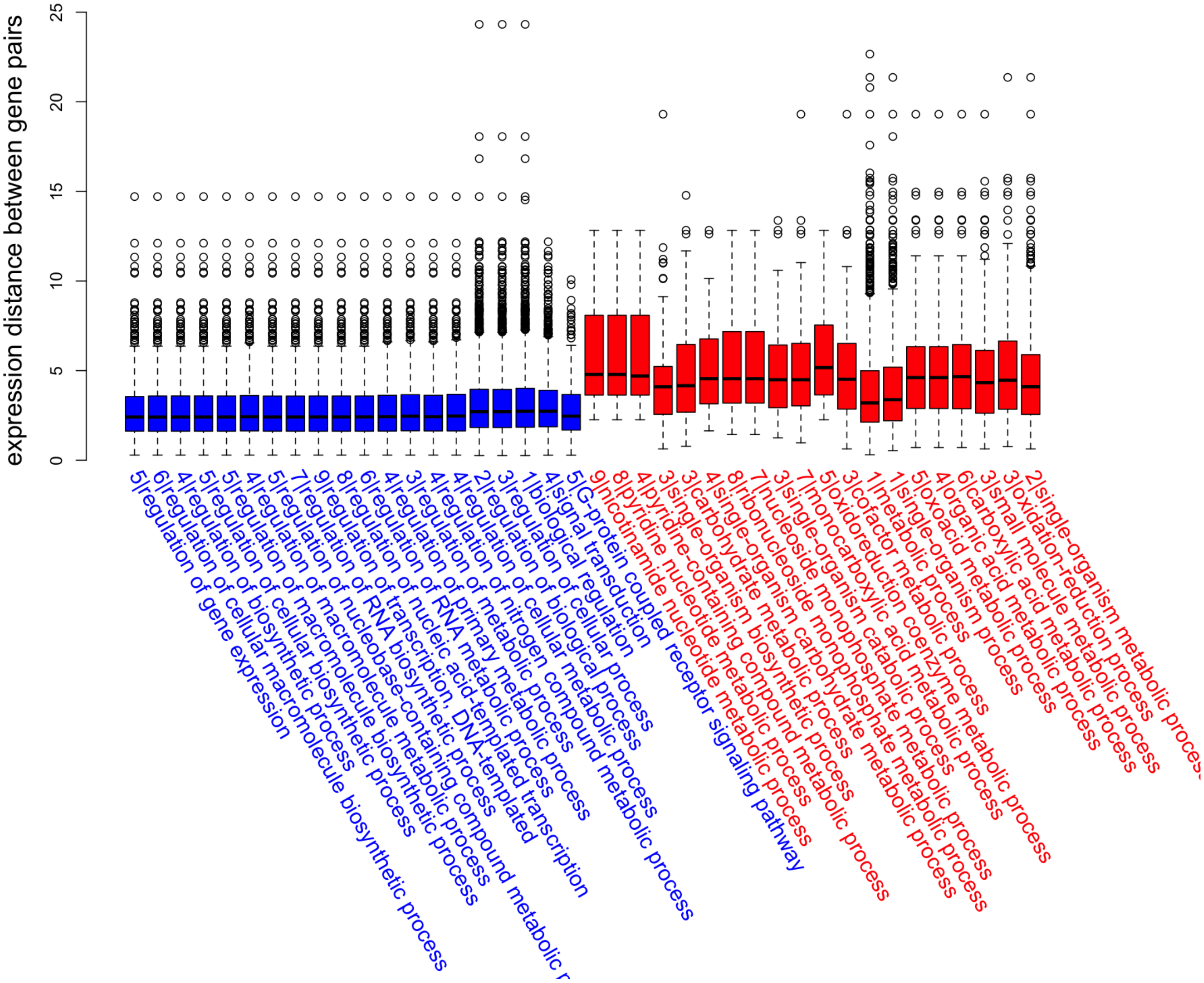
Fig. S11. Expression of ohnolog gene pairs in seven tissues. Histogram is symmetrized. Color indicates percent of gene pairs. For each tissue, the TPM expression difference between most of gene pairs are less than 2-fold (i.e. between white lines).



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
20	118	60	1	7	29	2	2	42	0	0	0	15	103	46	24	0	40	106	33	290 (32%)
19	62	4	1	1	3	1	1	22	1	0	0	6	7	3	1	0	4	3	108 (41%)	4.2%
18	83	307	0	11	139	12	61	54	3	1	0	10	398	37	43	1	154	564 (28%)	0.2%	5.2%
17	29	161	2	26	65	8	32	30	6	0	1	22	194	28	104	10	286 (24%)	6.6%	0.4%	2.6%
16	7	5	12	29	6	1	0	14	0	0	8	2	1	0	16	92 (45%)	1.0%	0.1%	0.0%	0.0%
15	13	38	3	14	25	2	5	11	0	0	4	75	137	33	242 (31%)	2.4%	7.1%	2.2%	0.1%	2.0%
14	8	15	2	1	8	1	9	4	1	0	3	100	171	184 (28%)	3.2%	0.0%	2.0%	2.0%	0.5%	4.2%
13	45	169	2	4	62	15	24	27	1	1	3	83	1164 (45%)	8.9%	6.9%	0.1%	8.2%	13.9%	0.4%	5.0%
12	10	4	1	3	8	2	4	7	0	13	37	264 (40%)	4.5%	11.5%	7.9%	0.4%	1.7%	0.5%	1.1%	1.5%
11	1	0	1	1	1	1	1	2	0	9	40 (35%)	7.8%	0.2%	0.6%	0.6%	4.3%	0.1%	0.0%	0.0%	0.0%
10	0	0	0	0	0	0	0	0	0	88 (79%)	9.3%	3.1%	0.1%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%
9	9	16	3	4	16	8	37	73	236 (57%)	0.0%	0.0%	0.0%	0.1%	0.2%	0.0%	0.0%	0.5%	0.2%	0.3%	0.0%
8	169	113	24	40	196	134	47	862 (46%)	6.1%	0.0%	0.2%	0.5%	1.1%	0.3%	0.7%	1.2%	1.6%	2.2%	1.9%	2.6%
7	9	104	7	11	98	248	898 (56%)	2.8%	4.2%	0.0%	0.1%	0.4%	1.1%	0.8%	0.4%	0.0%	2.0%	2.9%	0.1%	0.2%
6	5	39	5	8	83	562 (49%)	19.5%	8.5%	1.1%	0.0%	0.2%	0.2%	0.7%	0.1%	0.2%	0.1%	0.5%	0.6%	0.1%	0.2%
5	94	212	7	23	320 (23%)	5.0%	5.5%	9.4%	1.3%	0.0%	0.1%	0.5%	2.5%	0.5%	1.5%	0.5%	3.3%	5.6%	0.2%	1.7%
4	6	19	3	78 (27%)	1.8%	1.0%	1.2%	3.3%	1.0%	0.0%	0.4%	0.5%	0.2%	0.1%	1.8%	9.0%	2.3%	0.7%	0.3%	0.8%
3	4	15	46 (33%)	1.0%	0.6%	0.7%	0.9%	2.2%	1.1%	0.0%	0.6%	0.2%	0.1%	0.4%	0.5%	5.9%	0.2%	0.0%	0.4%	0.1%
2	86	520 (28%)	1.0%	1.2%	8.7%	2.0%	5.0%	4.8%	1.0%	0.0%	0.0%	0.2%	6.0%	0.8%	2.0%	0.3%	7.1%	11.0%	0.3%	3.0%
1	494 (39%)	4.0%	0.5%	0.6%	5.1%	0.4%	0.6%	9.6%	1.0%	0.0%	0.1%	0.9%	2.0%	0.7%	1.0%	0.8%	1.7%	3.8%	6.8%	8.5%

Fig. S12. Number of ohnolog gene pairs in the same cluster (diagonal) or between each of the 20 clusters (top triangle). The lower triangle shows the percentages. Blue-white-red Color indicates the percentage, from low to high.


expression distance between gene pairs
0
5
10
15
20
25
3|nucleic acid binding
4|DNA binding
5|sequence-specific DNA binding
1|nucleic acid binding transcription factor activity
2|transcription factor activity, sequence-specific DNA bindinc
1|binding
6|ion channel activity
4|channel activity
3|passive transmembrane transporter activity
5|substrate-specific channel activity
7|cation channel activity
5|gated channel activity
1|signal transducer activity
3|signaling receptor activity
2|heterocyclic compound binding
2|organic cyclic compound binding
3|transmembrane receptor activity
4|transmembrane signaling receptor activity
2|protein binding
2|receptor activity
3|glycosaminoglycan binding
3|tetrapyrrole binding
3|oxidoreductase activity, acting on paired donors, with incor
5|flavin adenine dinucleotide binding
3|peptidase regulator activity
4|peptidase inhibitor activity
6|iron ion binding
3|oxidoreductase activity, acting on CH-OH group of donors
3|enzyme inhibitor activity
4|endopeptidase regulator activity
5|endopeptidase inhibitor activity
5|NAD binding
4|oxidoreductase activity, acting on the CH-OH group of don
4|vitamin B6 binding
5|pyridoxal phosphate binding
3|vitamin binding
3|coenzyme binding
1|catalytic activity
2|cofactor binding
2|oxidoreductase activity


expression distance between gene pairs
0
5
10
15
20
25
5|regulation of gene expression
6|regulation of cellular macromolecule biosynthetic process
4|regulation of biosynthetic process
5|regulation of cellular biosynthetic process
5|regulation of macromolecule biosynthetic process
4|regulation of macromolecule metabolic process
5|regulation of nucleobase-containing compound metabolic r
7|regulation of RNA biosynthetic process
9|regulation of transcription, DNA-templated
8|regulation of nucleic acid-templated transcription
6|regulation of RNA metabolic process
4|regulation of primary metabolic process
3|regulation of metabolic process
4|regulation of nitrogen compound metabolic process
4|regulation of cellular metabolic process
2|regulation of biological process
3|regulation of cellular process
1|biological regulation
4|signal transduction
5|G-protein coupled receptor signaling pathway
9|nicotinamide nucleotide metabolic process
8|pyridine nucleotide metabolic process
4|pyridine-containing compound metabolic process
3|single-organism biosynthetic process
3|carbohydrate metabolic process
4|single-organism carbohydrate metabolic process
8|ribonucleoside monophosphate metabolic process
7|nucleoside monophosphate metabolic process
3|single-organism catabolic process
7|monocarboxylic acid metabolic process
5|oxidoreduction coenzyme metabolic process
3|cofactor metabolic process
1|metabolic process
1|single-organism process
5|oxoacid metabolic process
4|organic acid metabolic process
6|carboxylic acid metabolic process
3|small molecule metabolic process
3|oxidation-reduction process
2|single-organism metabolic proces-

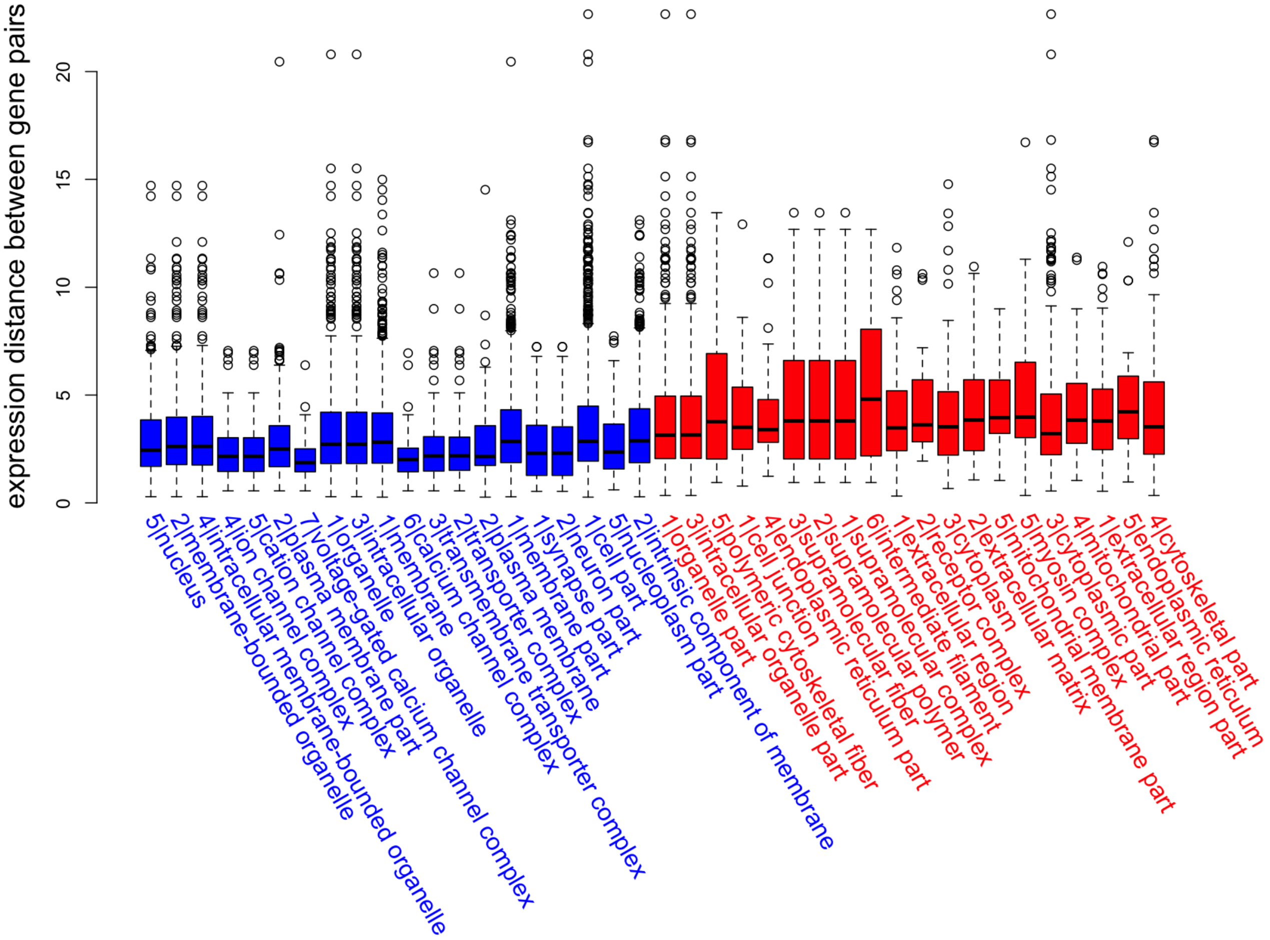


Fig. S13. Function enrichment and reduction in divergent expressed gene pairs. GO molecular function (top), biological process (middle) and cell component (bottom) with significantly low (top 20, blue) or high (top 20, red) expression distances between carp WGD ohnolog gene pairs (one side Wilcoxon rank sum test p<0.01).

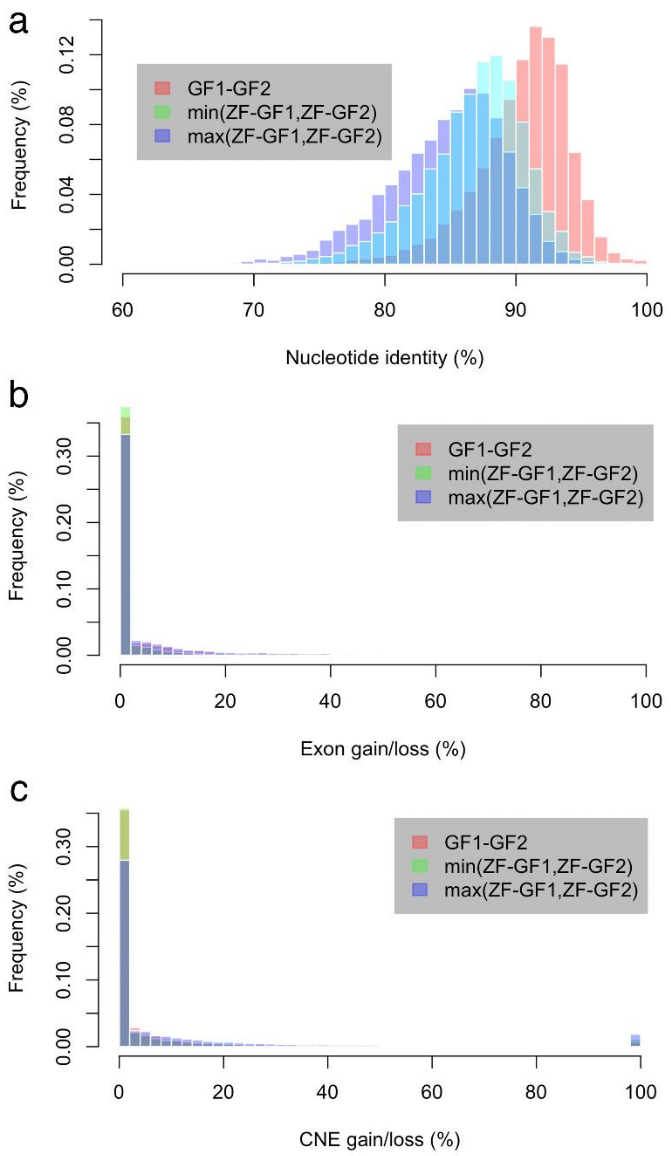


Fig. S14. Sequence divergence among zebrafish-goldfish triplets. Nucleotide sequence divergence between zebrafish ortholog and goldfish orthologs, and between goldfish ortholog pairs by: (a) nucleotide identity; (b) percentage of exon gain/loss length; (c) percentage of CNE gain/loss length. Frequency was counted from ZF-GF1-GF2 gene triplets. One goldfish ohnolog often retains high similarity to the zebrafish ortholog while the other accumulates more mutations. ZF: zebrafish ortholog; GF1,2: goldfish ohnolog, 1 or 2 is randomly assigned.

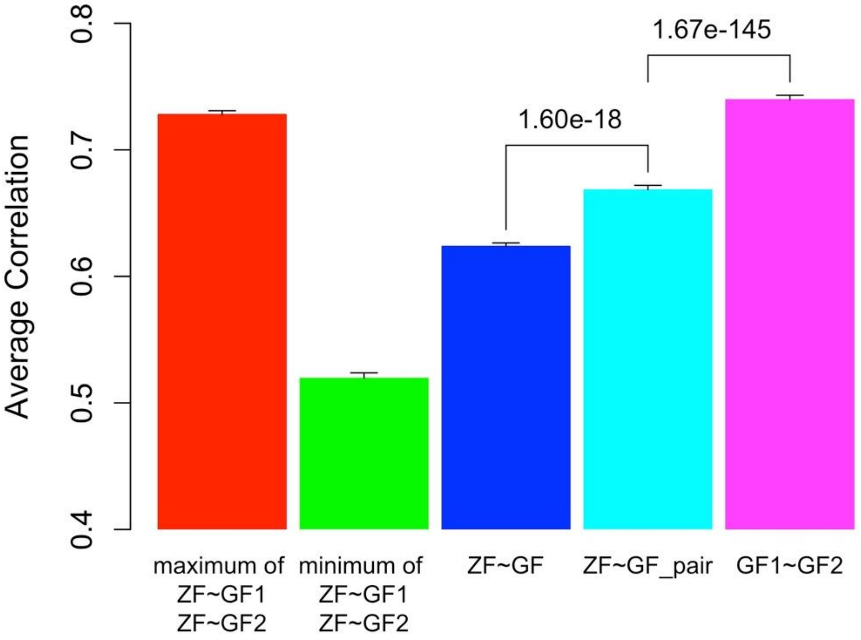
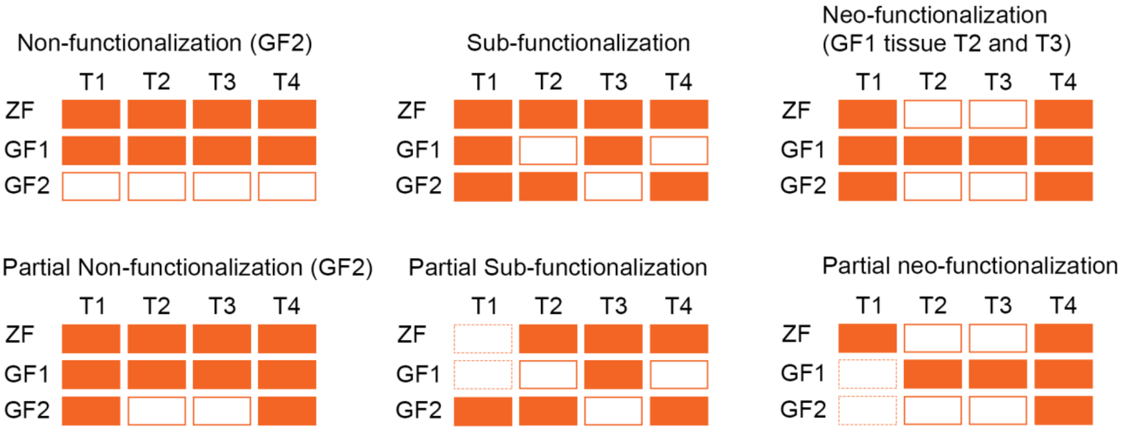


Fig. S15. Pearson's correlation coefficient between zebrafish ortholog (ZF)–goldfish ohnolog (GF) and goldfish ohnolog-ohnolog (GF1-GF2). GF_pair is the sum expression of two GF ohnologs. GF1-GF2 and maximum of (ZF-GF1,ZF-GF2) have the highest expression (log2(FPKM+1)) correlation, while the minimum of (ZF-GF1,ZF-GF2) is the lowest, suggesting one ohnolog maintains high correlation with the ZF ortholog. ZF-GF_pair is higher than ZF-GF indicates that goldfish ohnolog maintains dosage correlation with the zebrafish orthologs.



Non-functionalization (GF2)

	T1	T2	T3	T4
ZF	■	■	■	■
GF1	■	■	■	■
GF2	□	□	□	□

Sub-functionalization

	T1	T2	T3	T4
ZF	■	■	■	■
GF1	■	□	■	□
GF2	■	■	□	■

Neo-functionalization (GF1 tissue T2 and T3)

	T1	T2	T3	T4
ZF	■	□	□	■
GF1	■	■	■	■
GF2	■	□	□	■

Partial Non-functionalization (GF2)

	T1	T2	T3	T4
ZF	■	■	■	■
GF1	■	■	■	■
GF2	■	□	□	■

Partial Sub-functionalization

	T1	T2	T3	T4
ZF	⬚	■	■	■
GF1	⬚	□	■	□
GF2	■	■	□	■

Partial neo-functionalization

	T1	T2	T3	T4
ZF	■	□	□	■
GF1	⬚	■	■	■
GF2	⬚	□	□	■

Fig. S16. Definition of neo-F, sub-F, and neo-F. Filled box indicates expressed or 'on' (FPKM>=2), while open box indicates silence or 'off' (FPKM<=1). Dashed boxes in the partial- definition is for distinguishing from non-partial- definition for displaying purpose.

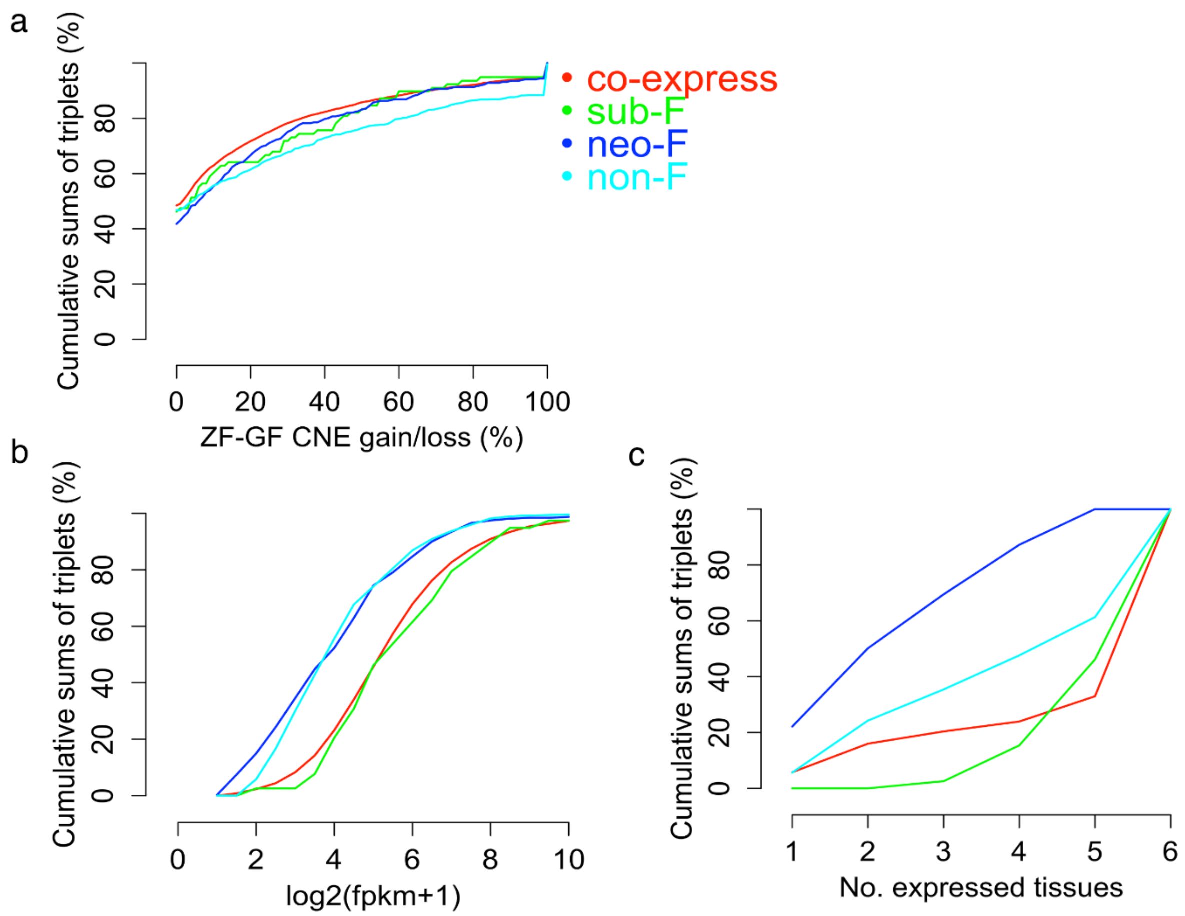


Fig. S17. Correlation between different classes of gene expression changes and gain/loss of CNEs. (a) Cumulative sum of triplets for different CNE gain/loss categories between zebrafish (ZF) and goldfish (GF). New-expressed genes in the neo-F groups have more CNE gain/loss compared to zebrafish. (b) Zebrafish genes in the non-F and neo-F groups have more lower expressed genes than those in the co-expressed and sub-F groups (expression FPKM is the highest FPKM across all six tissues). (c) Zebrafish genes in sub-F group expressed in more tissues than those in non-F group and neo-F groups, suggesting genes expressed in more tissues are more likely to be subfunctionalized.

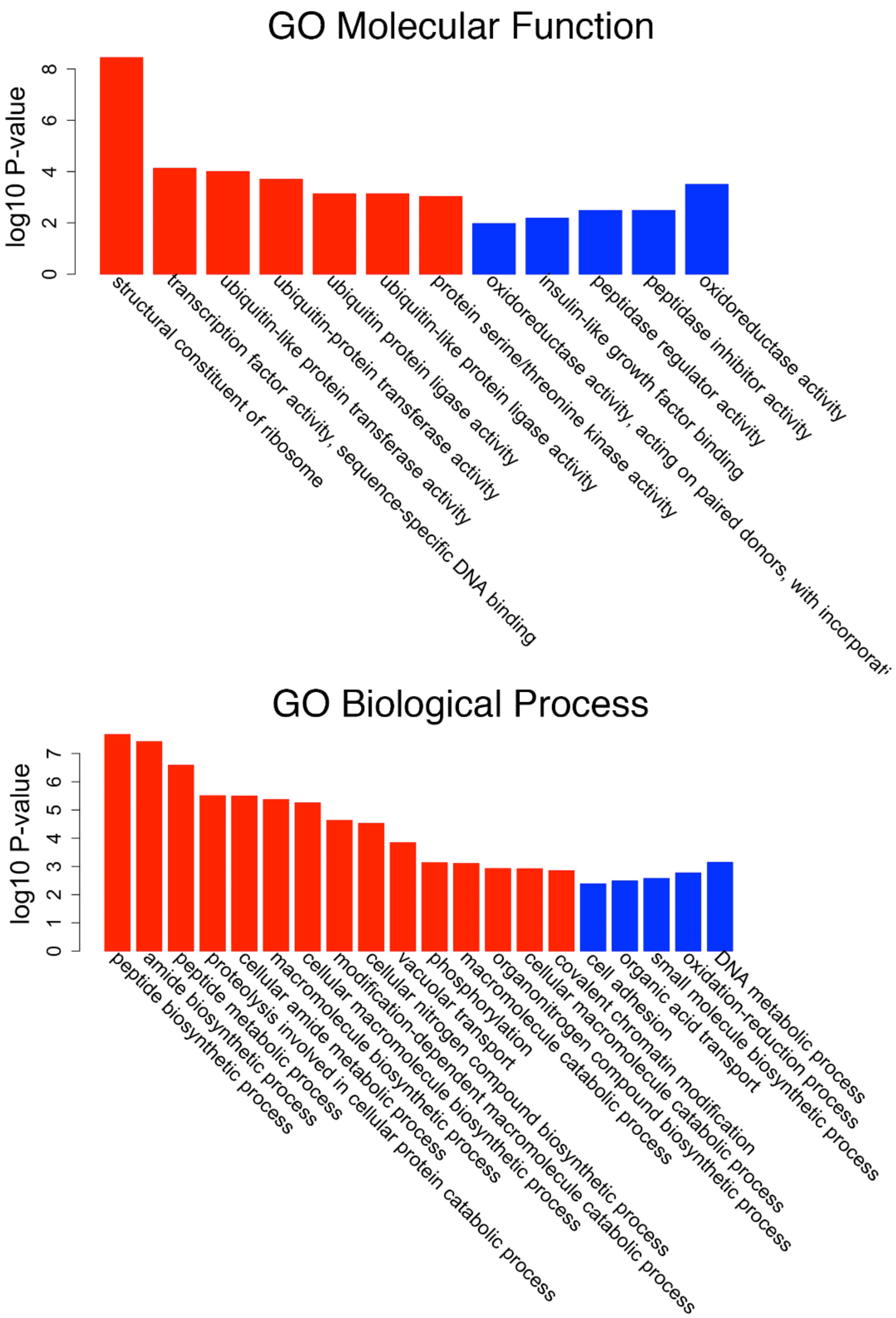


Fig. S18. Function enrichment (red) or reduction (blue) of genes in coexpressed groups. Genes are enriched in functions involved in macromolecule, biosynthetic process, metabolic process, ribosome function, etc.

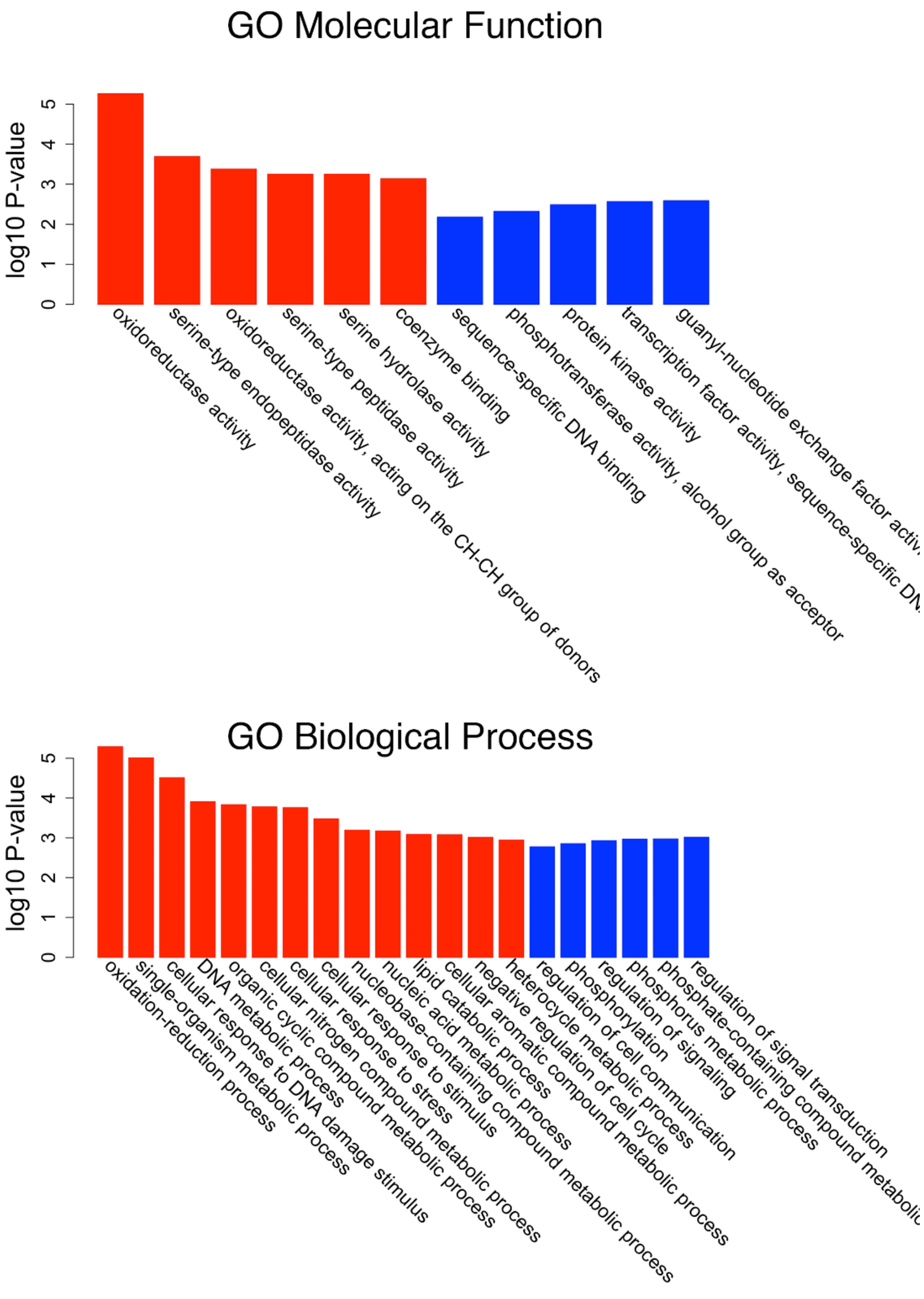


Fig. S19. Function enrichment (red) or reduction (blue) of genes in nonfunctionalized groups. Genes are enriched in functions involved in oxidoreductase activity, cellular nitrogen compound metabolic process and serine hydrolase activity, etc.

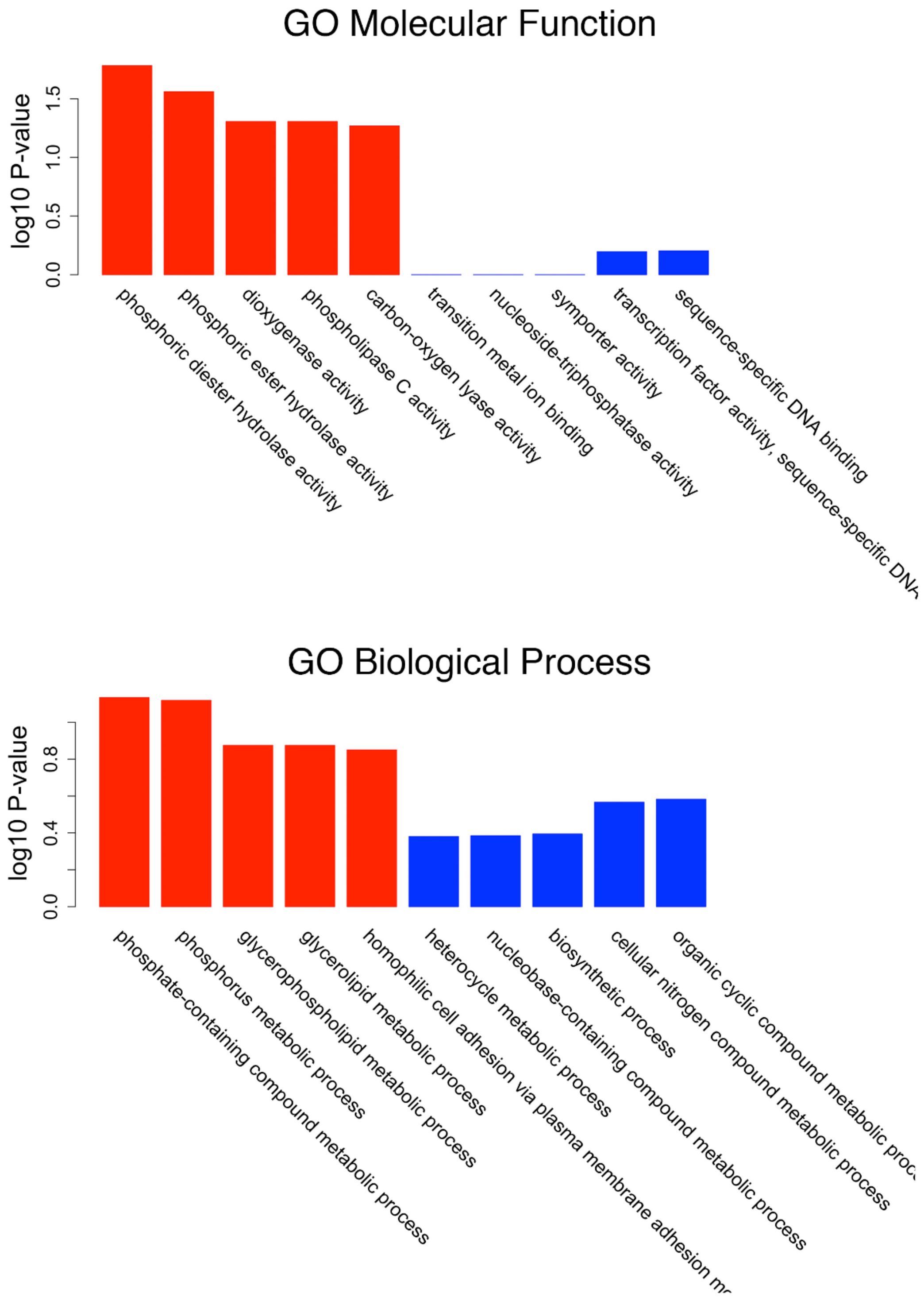
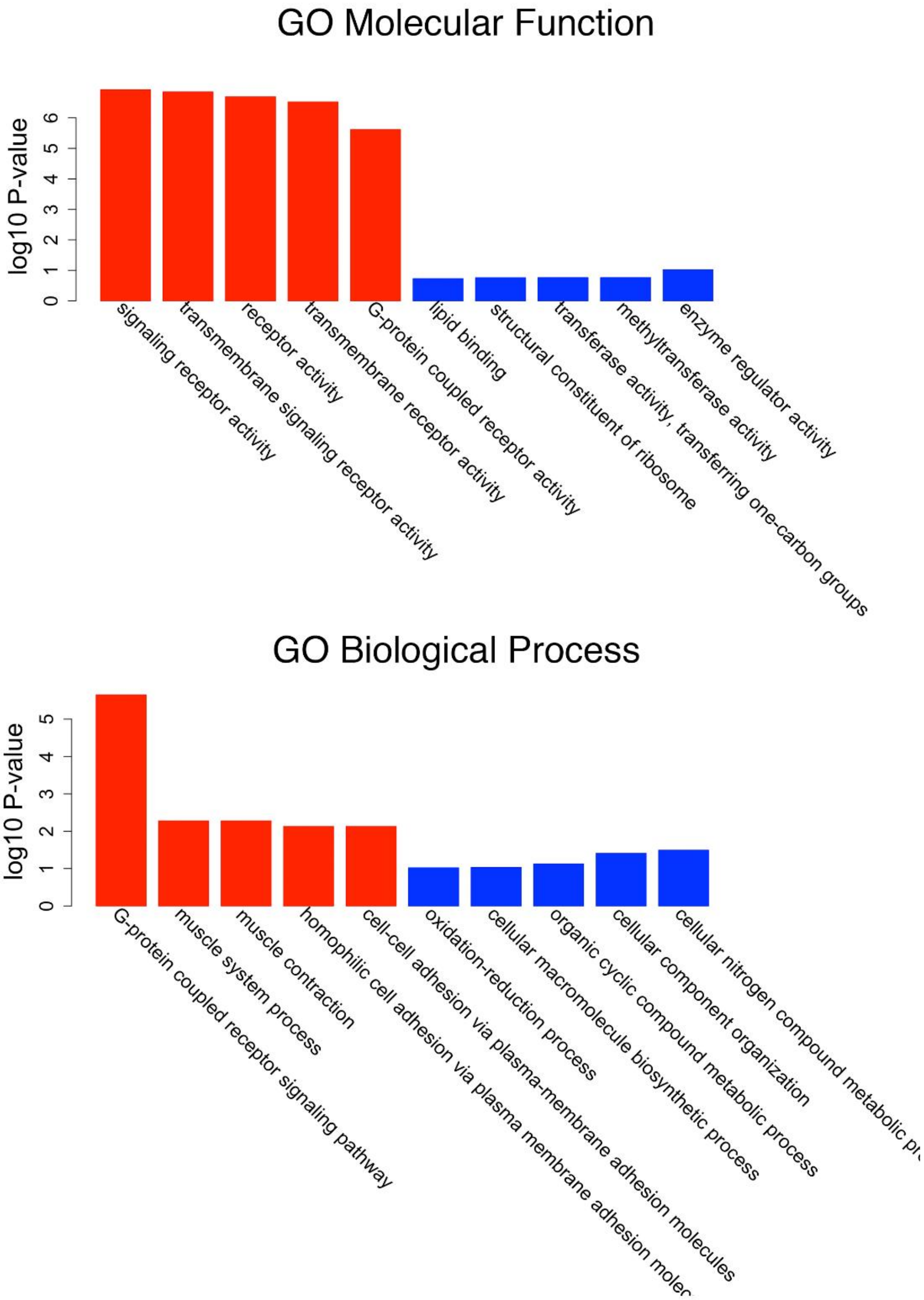


Fig. S20. Function enrichment (red) or reduction (blue) of genes in subfunctionalized groups. The top five significant GO terms are selected because the number of sub-functionalized group is too small. Genes are enriched in functions involved in signal transduction and dioxygenase activity, etc.



GO Molecular Function

GO Biological Process

Fig. S21. Function enrichment (red) or reduction (blue) of genes in neofunctionalized groups. Genes are enriched in functions involved in cell-cell adhesion, signaling receptor and transductor activity, etc.

Supplementary Methods and Analysis

Goldfish Genome Homepage

https://research.nhgri.nih.gov/goldfish/

De novo Assembly

Goldfish husbandry

Fertilized goldfish eggs were incubated at 20°C. After 3 to 5 days post-fertilization (dpf), hatched goldfish larvae were fed brine shrimp (Artemia) twice per day. The water in tanks for larvae was changed with fresh water incubated at 20°C every week. After 14 dpf, goldfish were fed pellets once per day. The water in tanks for adult goldfish was changed with fresh water every month. All procedures using goldfish were approved by the Animal Experimental Committees of the Institute for Protein Research at Osaka University (approval ID 29-03-0), and were performed according to the Guidelines for Animal Experiments of Osaka University.

Genome Assembly

We obtained 16,671,136 reads longer than 1kbp, containing a total of 130 Gb with an N50 length of 9,889 bases (Table 1). All reads were corrected and assembled into 9415 contigs using Canu (*21*) and consensus accuracy improved using Arrow from the PacBio software package. Total length of the Canu assembly is 1,848 Mb and N50 reached 816.8kbp, the longest contig was 12.8Mbp. We found that 6,937 contigs (~497Mbp) had relative read coverage less than 0.6, which may be from the heterogeneous diploid region of our fish sample, compared to 2,393 contigs (total length ~1347Mbp) with read coverage in the range of 0.6 to 1.8, most likely from the homologous regions (Table 2) This is consistent with the 25-mer spectrum from our Illumina HiSeq2500 short read sequencing (Fig. 1). By summing all contig lengths normalized by read coverage, we determined the actual haploid genome size was at least 1.6Gbp. Contigs were aligned to self by using nucmer (*61*). 928 contigs contained in other contigs with low read coverage were removed, which was 27.3Mbp in total. All other contigs were retained.

Linkage Group Construction

RNA-seq data from two goldfish parents and their F_1 offspring were download from NCBI (bioproject:PRJEB12518) (*22*). All reads were trimmed using Trimmomatic (*62*) (same configuring as in Gene Annotation) and aligned to the Canu assembly using hisat2 (*53*). Variant calling was performed via samtools mpileup and bcftools call (parameter '-m') (*54*). We obtained ~5.6 M variants in total. SNPs with missing genotype or low read depth (<4) in more than 25% samples or with missing genotype in the two parents were removed (other filter: bcftools filter -g 10 -Ov -i 'TYPE="snp" && QUAL>=10 && INFO/DP>=50'). SNPs that were homozygous in both parents or failing a Mendelian test were removed. We also required two SNPs on the same contig to be separated by at least 10Kbp. 14022 SNPs were kept after filtering and used for constructing genetic maps.

SNPs from the same contigs were grouped and ordered using ‘group’ and ‘seq.order’ from the R package ‘onemap’, with a LOD threshold of 5.5. Contigs with two or more groups (with each >= 3 markers) were broken at position with read depth valley and depth < 20 and depth < 20% quantile. In total, 16 contigs were broken. All SNPs were grouped using ‘group’ in the ‘onemap’ package. SNPs in each group were ordered using ‘seq.order’. Contigs were placed in each linkage group according to the ordered SNPs using chromonomer (http://catchenlab.life.illinois.edu/chromonomer/, v1.06). After manual corrections, 50 long linkage groups were retained and named according their alignment to the zebrafish genome (LG1 and LG26 map to zebrafish chr1, LG2 and LG27 map to zebrafish chr2, and so on). Several short linkage groups, which were named according to their zebrafish alignment, were also retained. This assembly was named ‘carAur01’.

Genome Annotation

Repeat Masking and Gene Structure Annotation

A custom repeat library for goldfish was built using RepeatModeler (http://www.repeatmasker.org/) based on the Canu assembly. Zebrafish and the custom repeat library were used to mask the genome by RepeatMasker (http://www.repeatmasker.org/, performed in MAKER3).

RNA-seq from seven goldfish tissues were performed to aid with gene annotation, include bone, brain (3 samples), eye, gill (2 samples), heart, muscle and tailfin. RNA libraries were prepared and sequenced on HiSeq2000 sequencer by NISC. All 2x125 pair-end reads were trimmed using Trimmomatic (ILLUMINACLIP:adapters/TruSeq3-PE-2.fa:2:30:10:8:true LEADING:3 SLIDINGWINDOW:20:20 MINLEN:40) and assembled via Trinity assembler without a genome-guide (*56*). All assemblies were clustered via CDHIT (-c 0.95 -aS 0.95 -uS 0.05), as EST evidence for Maker 3.0.

cDNA sequences from the Ensembl database (version 85, 69 species), NCBI vertebrate RefSeq and common carp (http://www.carpbase.org/gbrowse.php) were used as alternative RNA evidence. Proteins from the Ensembl database, common carp, and UniProt database (uniref90) were used as protein evidence. To annotate gene structure, we performed MAKER 3.0 (*28*) on the Canu assembly with Augustus prediction and the EST, RNA, protein evidence. Gene structures were lifted over to the carAur01 assembly using liftover (*57, 63*) or crossmap (https://sourceforge.net/projects/crossmap/files/).

Because our fish was not fully homozygous, we needed to identify those genes in the heterozygous diploid regions. All cDNA sequences from Maker gene models were aligned to self by megablast. Alignments with identity ≥ 97.5% and coverage of both sequences ≥ 70% were kept. Alignments were retained if they satisfied one of the following restrictions: (1) identity >= 99.5% and the relative coverages of both contigs where the two genes were located were less than 0.8, (2) the relative coverage of both contigs was less than 0.75, (3) the relative read coverage of either contig was less than 0.6. DNA sequences from all remaining aligned genes were fetched and aligned using lastz and chained with axtChain. All alignments with matched basepairs covering less than 0.6 of both genes or with identity less than 95% were discarded. Only the shorter of the two genes in the retained alignments was masked and not used for following analysis.

MAKER3 generated 81,778 coding gene models, of which 80,062 were liftover'ed to carAur01, and 9,738 genes were masked as one allele of the heterozygous genes. The average exon and intron length was ~202bp and ~174bp. The distribution of exon and intron size is similar to zebrafish, grass carp and common carp (fig. S2).

Non-coding RNA annotation

Non-coding RNA sequences from other species were downloaded from NONCODE (*64*) (zebrafish and human), RNAcentral (*65*) and Ensembl ncrna (ver. 85) (*66*). All sequences were first aligned to the genome using blastn in the NCBI-BLAST+ package (*67*) (-evalue 1e-4 –perc_identity 80). All genomic target regions were fetched and refined using exonerate(*68*) for each query. Exonerate alignments for each query RNA were kept if they satisfied: (1) score ≥ 0.9 best score for the query; (2) query coverage ≥ 0.6; (3) query identity ≥ 0.7; (4) non-canonical splice site ≤ 3.

Trinity genome-guided assembly was performed on the RNA-seq data from the seven tissues. 'align_and_estimate_abundance.pl' from the Trinity package was used to estimate the expression of each transcript. Transcripts with expression lower then 1 TPM were filtered. All remaining transcripts were aligned to the Canu assembly using the same BLASTN-exonerate approach except using a higher identity 90%. Exonerate alignments for each query RNA were kept if they satisfied: (1) score ≥ 0.95 best score for the query; (2) query coverage ≥ 0.75; (3) query identity ≥ 0.9; (4) non-canonical splice sites ≤ 3. All Trinity transcripts with no alignment to any MAKER genes or with Trinotate PFam/Spot annotation were also removed (*69*). Coding potential of the remaining transcripts were predicted by using CPC (*70*). Transcripts with 'coding' labels were removed. All the remaining exonerate results were transformed to GFF3 and merged using 'cuffcompare' from cufflinks package.

Hairpin sequences from miRBase were also aligned to the genome using the BLASTN-exonerate approach. Alignments were retained if they satisfied: (1) score ≥ 0.9 best score for the query and (2) query coverage >90%, identity >90%.

The genome was scanned against the Rfam database using cmscan from the Infernal package (version 1.1.1) (*71, 72*). Only hits with bit score ≥ 30 and E-value ≤ 10e-6 were kept. When dealing with overlapping hits, we kept the hit amongst all overlapping hits that had the highest bit score.

Conserved Noncoding Elements (CNE) Identification

All-to-all pairwise genomic alignment was performed using lastz (--gapped --ambiguous=n --step=3 --strand=both --masking=100 --maxwordcount=100 --identity=70..100 --format=axt) and axtToChain for four species (goldfish, common carp, grass carp, zebrafish) and transformed to pairwise MAF format and split at gaps longer than 30bp (chainToAxt –maxGap=30, then axtToMaf -score). All the pairwise MAF files were transformed to multiple alignment MAF files using roast (P=multic). Phylogenetic models were fit for each chromosome, linkage group or scaffold using phyloFit (--tree '(ZF,(GC,(GF,CC)))' --subst-mod REV --nrate 4), which was used by phastCons for computing conservation scores and most conserved regions. The most conserved regions out of exons (of coding or noncoding genes) were defined as CNE (conserved noncoding element). goldfish (or common carp) CNE that overlapped at least 50% of the goldfish-goldfish (or common carp-common carp) self chain-net alignment regions were retained either as both WGD copies or as singletons.

Gene Functional Annotation

Interproscan5 (*58*) was used to annotate the Interpro/GO/Pathway function for all protein-coding genes.

SNV and DIV

2x250 read pairs from a second gynogenic goldfish (GF71, 73X coverage) and a wild-type goldfish (WTGF, 70X coverage) were aligned to the carAur01 assembly using bwa mem (bwa mem -t 16 -I 538.,149.3). Most Probable Genotype (MPG) (https://research.nhgri.nih.gov/software/bam2mpg/index.shtml, https://github.com/nhansen/bam2mpg) (*73*) was used to call variants from the bwa mem produced bam files. The MPG output variant calls were converted to VCF for variants with a minimum Most Probable Variant (MPV) score of 10 or greater with a MPV-score/read-coverage ≥0.5

Functional Enrichment

Fisher exact tests were performed to identify significantly enriched GO molecular functions among goldfish, common carp, grass carp and zebrafish. We also performed the same tests between duplicated retained genes and single-copy-lost genes in goldfish for each GO terms in the 'molecular function' and 'biological process' domain (Fig. 8). Compared to the other three species, goldfish show enriched function in channel activity and depressed function in olfactory receptor activity (Fig. 9).

Evolution Analysis

Ohnolog Gene Clusters

Protein and cDNA sequences of zebrafish (GRCz10) were downloaded from the Ensembl database. Grass carp sequences were downloaded from Grass Carp Genome Database (GCGD) (*74*). Common carp sequences were downloaded from NCBI

(GCF_000951615.1). We performed all-to-all Blastn on the cDNAs from the four species. Non-overlapping alignments from the same cDNA pairs were concatenated. We identified synteny blocks for each pair of species through iteratively merging nearby aligned gene pairs with, at most, five unaligned genes between them. Alignments were used as an edge to group genes into clusters with constrained gene numbers for each species according to whether it was before or after the carp WGD event (zebrafish : grass carp : common carp : goldfish = 1:1:2:2). Two genes or gene clusters were merged if the number of edge between them was > $50\%\mathrm{N}_1\mathrm{N}_2$, or > $20\%\mathrm{N}_1\mathrm{N}_2$ and there were edges linked between the two genes to a matching outgroup gene according to the species tree '(zebrafish, (grass carp, (common carp, goldfish)))', where N_1 and N_2 were the number of genes in each gene cluster. The priority for the edge for aggregate genes or gene clusters were edges in synteny blocks and then 'reciprocal best hit' edge. Other edges were used to rescue and merge some genes into those non-full-size (i.e. 1:1:2:2) clusters.

Phylogenetic Analysis

Proteins from all 1:1:2:2 ohnolog clusters were multiple aligned using MAFFT (*75*) with '--auto' option, then transformed to codon alignment using 'tranalign' from EMBOSS Suite (*76*). Poorly aligned codon regions were eliminated using Gblocks (*77*). The third position of all codons was filtered out into separated alignments. All third-codon sequences from the same chromosomes were concatenated for building phylogenetic trees. ML tree was built using RAxML (*78*) with the model GTRGAMMA. Pairwise synonymous substitutions were computed by using 'codeml' from the PALM package (runmode = -2, method = 0) (*79*). Divergence time of the carp WGD event was estimated by 20.5*L(WGD)*2/L(grass_carp,carp), where 20.5 is the divergence time of grass carp and common carp in unit Mya, L(WGD) is the average branch length from WGD event to goldfish and common carp, L(grass_carp, carp) is the average branch length between grass_carp and common carp or goldfish. Similar estimation was performed for the speciation of common carp and goldfish.

Computation of gene loss rate in goldfish and salmon

The number of ohnolog gene clusters with retained duplicated goldfish genes or blat aligned loci was 16,455, while those with singletons was 2,341. Divided by the WGD time of 14.4 My, duplicated genes were lost after the WGD at a rate of ~2,341/((2,341+16,455)*2)/14.4 = 0.43%. In salmon, the number of ohnologous gene clusters with retained duplicated (and singleton) salmon genes was (22,803+5,125)/2=13,964 and 9,278 (last row of table S11 in Lien 2016). Divided by the WGD time of ~80My, the rate was 9,278/((9,278+13,964)*2)/80=0.25%. Goldfish lost genes faster than salmon after their respective WGD (Chi square test p<2.2e-16). There were 2,964 ohnolog clusters without zebrafish homologs and 521 without grass carp homologs out of 23,592 clusters, i.e. 7.4% of gene gain/loss occurred between zebrafish and grass carp (over 120 *My*), which resulte in a rate of 7.4%/120 = 0.061% (Chi square test p=4.861e-05).

Expression Comparison between Retained WGD Gene Pairs

Co-linear blocks were fetched from the goldfish self chain-net alignment. Gap larger than 20kbp was broken. Blocks shorter than 50kbp were removed. Blocks were removed if it overlaps other longer blocks. The two sequences in each collinear block were presumed to be derived from the same sequence before the carp WGD event. WGD gene pairs/exons/CNES were fetched from these collinear blocks for follow-up analysis. Exons/CNEs were map to their paralogs using crossmap based on the self chain co-linear blocks. Exons/CNEs with more than 50% failure to map to any genomic region is considered as exon/CNE loss, denoted as singleton exons/CNE, and counted as one exon/CNE loss in its paralogous gene. The number of lost exon/CNEs for gene pairs was defined as the total number of lost exon/CNEs of both ohnologous genes. Some CNEs may have been annotated as exons because of annotation errors. We labeled CNEs that mapped to a paralogous region containing an exon as exons to remove these errors. Genes that the CNEs were predicted to regulate were defined as the nearest gene(s) in 5kbp windows in either direction.

RNA-seq reads from the seven tissues were mapped to the carAur01 assembly using STAR (default settting and two pass). Expression levels (TPM or FPKM) were estimated using RSEM (rsem-calculate-expression --paired-end --forward-prob 0.0 --alignments -p 16 --seed 987347 --calc-ci --calc-pme --estimate-rspd --time --no-bam-output) and transformed to logTPM=log2(TPM+1). Euclidean distances or correlation coefficients of the expression between WGD gene pair were calculated in R. 449 gene pairs were silenced, another 649 gene pairs contained exactly one silenced gene. The remaining 19,500 genes (9,750 gene pairs with both genes expressed) were hierarchically clustered using the ‘hcluster’ and ‘ward.D2’ method in R, based on the logTPM value and Euclidean distance. Tissue specific expressed gene pair was defined as gene pair with TPM>=4 in one gene and TPM<0.5 in the other gene in at least one tissue. Expression standard deviations across the seven tissues were also calculated for each gene.

Gene pairs were divided into 6 groups according to their pairwise cDNA identity (≤86%, 86-88%,88-90%,90-92%,92-94%,>94%). Histogram of expression distances for each group were computed in R using ‘hist’ with bin size 2. In order to illuminate the relationship between exon loss and expression distance, gene pairs were divided into 4 groups: no exon loss, one exon loss, two exon losses, three or more exon losses. One sided Wilcoxon rank sum tests were performed for each pair of groups. For CNE lost, Wilcoxon rank sum test was performed on the expression standard deviation between genes in the no-CNE-lost group and those in the CNE-lost group, using only gene pairs with CNE loss but no exon loss.

In order to find out which biological functions were prone to diverging after the WGD, we performed Wilcoxon rank sum tests on the expression distance between genes inside the GO terms and genes outside the GO terms. The top 20 and bottom 20 GO terms with p < 0.1 were plotted in Fig. 13.

Software and Databases

Software	Version	URL
Trinity	2.8.4	https://github.com/trinityrnaseq/trinityrnaseq/wiki
MAKER	3.0	http://www.yandell-lab.org/software/maker.html
CrossMap	0.2.7	https://sourceforge.net/projects/crossmap/files/
Canu	1.4	http://canu.readthedocs.io/en/latest/index.html
onemap	2.0.8	https://cran.r-project.org/web/packages/onemap/index.html
RepeatMasker	4.0.7	http://www.repeatmasker.org
NCBI-BLAST+	2.6.0+	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/
Exonerate	2.2.0	https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate
lastz	1.04	https://www.bx.psu.edu/~rsharris/lastz/
PHAST	1.3	http://compgen.cshl.edu/phast/
Trinotate	3.0.1	https://trinotate.github.io
Infernal	1.1.2	http://eddylab.org/infernal
InterProScan	5.27	ftp://ftp.ebi.ac.uk/pub/software/unix/iprscan/5
samtools	1.9	http://samtools.sourceforge.net
Bam2mpg	1.0.1	https://research.nhgri.nih.gov/software/bam2mpg/index.shtml
GBlocks	0.91b	http://molevol.cmima.csic.es/castresana/Gblocks.html
RAxML	8.2.11	https://sco.h-its.org/exelixis/web/software/raxml/index.html
PAML	4.9e	http://abacus.gene.ucl.ac.uk/software/paml.html
EMBOSS	6.6.0	http://emboss.sourceforge.net/index.html
STAR	2.5.2b	https://github.com/alexdobin/STAR
RSEM	1.3.0	https://deweylab.github.io/RSEM
Trimmomatic	0.36	http://www.usadellab.org/cms/?page=trimmomatic
HISAT2	2.2.1.0	https://ccb.jhu.edu/software/hisat2/index.shtml
BUSCO	3.0.2	https://busco.ezlab.org
CHROMONOMER	1.06	http://catchenlab.life.illinois.edu/chromonomer/

Database	Version	URL
Ensembl	Release-85	http://ensembl.org
NONCODE	V5	http://www.noncode.org/
RNACentral	8.0	http://rnacentral.org
PFam	February 2015	http://pfam.xfam.org
Uniprot	2016-10-11	https://www.ebi.ac.uk/uniprot
RFam	12.3	http://rfam.xfam.org
UCSC genome browser	-	http://genome.ucsc.edu