Chapter 2.

Nearly half of the genes duplicated through fish-specific genome doubling underwent lineage-specific loss or retention

2.1. Introduction

It is widely supposed that rounds of ancient polyploidization, or whole genome doubling (WGD), have been involved in shaping genomic architecture of eukaryotes, including yeasts, plants, and vertebrates (Blanc and Wolfe, 2004; Kellis et al., 2004; Adams and Wendel, 2005; Dehal and Boore, 2005; Panopoulou and Poustka, 2005; Froschauer et al., 2006). To understand the genome evolution in eukaryotes, therefore, it would be important to investigate the evolution of duplicated genes derived from WGD events. The formation of new genomic features following WGD should have been accompanied by both loss of function and acquisition of new roles (sub/neo-functionalization; Ohno, 1970; Force et al., 1999) of duplicated genes.

Vertebrates, which have the most complex body plan and behavioral characteristics, are considered to have underwent several rounds of WGD early in their evolution (Ohno, 1970; Lundin, 1993; Holland et al., 1994). This notion was largely supported by the data from recent genome studies (Panopoulou et al., 2003; Dehal and Boore, 2005). For example, multiple *Hox* gene clusters and MHC paralogous regions in vertebrate genomes appeared to have been generated through the ancient WGD (Abi-Rached et al., 2002; Hoegg and Meyer, 2005), implying the involvement of WGD events in the evolution of unique characteristics of vertebrates.

Recently, analysis of teleost fish genomes showed that an additional third-round (3R) WGD has occurred in the common ancestor of teleosts (Christoffels et al., 2004; Jaillon et al., 2004; Vandepoele et al., 2004; Christoffels et al., 2006; Kasahara et al., 2007), while the other

vertebrates such as tetrapods underwent only 1R- and 2R-WGDs (Panopoulou et al., 2003; Dehal and Boore, 2005; Panopoulou and Poustka, 2005). This surprising finding suggests that the number of encoding genes of teleost genome is larger than that of tetrapod genome. However, the estimated number of protein-coding genes in teleost genome is not quite different from that of mammalian genomes (mammals: 22,000 in average among human, mouse, dog, and cow; teleosts: 23,000 in average among pufferfish, stickleback, medaka, zebrafish; Ensembl Genome Browser, July 2007). Most of the duplicate genes generated by the 3R-WGD are therefore considered to have been lost in the evolution of teleosts.

It is unclear, however, how the duplicated genes generated through 3R-WGD have been lost or persisted during diversification of the teleost lineages. Previous studies have investigated the duplicated genes in teleost genomes only by pairwise comparisons of single teleost species versus human (Jaillon et al., 2004; Kasahara et al., 2007). These studies suggested that the total number of 3R-WGD-derived duplicate gene pairs per genome was 750 to 2,100, however, these estimates lacked evolutionary information. It may be assumed that almost all of the duplicated genes have been rapidly lost subsequent to the 3R-WGD, but it is also possible that there had been a large number of 3R-WGD-derived duplicated genes persisted in the common ancestor of teleosts. Subsequent loss or persistence of the duplicate genes that occurred independently among teleosts may have contributed to architectural differences among teleost genomes, and then, distinct characteristics among teleost lineages. To address these concerns, comparative evolutionary analysis of multiple teleost genomes is required on the basis of a reliable phylogenetic framework for teleosts and divergence time estimates between their lineages.

In this chapter, I investigated four teleost fishes, zebrafish, stickleback, medaka, and pufferfish, for which the genome sequence data are available. For comparative genomic analyses, I focused on several signal transduction pathways involved in learning and memory, and sensory system, which may have played an important role in vertebrate evolution, and

metabolic pathways involved in energy metabolism, which are common to eukaryotes. As representatives of these pathways, molecular interaction networks of long-term potentiation of symaptic transmission (LTP), taste transduction (TT), olfactory transduction (OT), and TCA cycle (TCA) were examined. Through the analysis of genes involved in these networks, I estimated the total number of genes that have remained in duplicate in common ancestor of the four teleosts since the 3R-WGD. Then the process of lineage-specific loss of these 3R-WGD-derived duplicated genes were inferred. I found that an unexpectedly larger number of duplicated genes has been maintained in the ancestor. I further examined the characteristics of these duplicated genes that have persisted since the 3R-WGD, by focusing on gene product length, gene function, and a number of interaction partners in the molecular interaction networks.

2.2. Materials and Methods

2.2.1. Database survey, phylogenetic analysis, and identification of orthologous gene groups

Network diagrams for LTP, TT, OT, and TCA, and cDNA sequences of human genes that comprise these networks were obtained from KEGG pathway database (Kanehisa et al., 2004). The obtained human cDNA sequences were used as queries to BLASTN search against the Ensembl genome database (Birney et al., 2006) of next organisms (database versions); human *Homo sapiens* (NCBI 36, Oct 2005), chicken *Gallus gallus* (WASHUC2, May 2006), clawed frog *Xenopus tropicalis* (JGI 4.1, Aug 2005), zebrafish *Danio rerio* (Zv7, Apr 2007), medaka *Oryzias latipes* (HdrR, Oct 2005), stickleback *Gasterosteus aculeatus* (BROAD S1, Feb 2006), pufferfish *Tetraodon nigroviridis* (TETRAODON 7, Apr 2003), ascidian *Ciona intestinalis* (JGI 2, Mar 2005), and fruit fly *Drosophila melanogaster* (BDGP 4.3, July 2005). The resulting BLAST hits were manually screened (*E*-value cut-off of $<10^{-3}$) and evaluated

for their gene product length and Ensembl annotations to confirm their similarity with the human gene queries. When only a partial cDNA was found in the Ensembl transcript databases, I predicted full length cDNA by using the program WISE2 (Birney et al., 2004) from the genomic sequence data.

Amino acid sequences of the proteins obtained by above procedure were aligned using ClustalW (Thompson et al., 1994). All gap-containing sites were removed. For each alignment, a preliminary neighbor-joining (NJ) analysis was performed based on Poisson-corrected genetic distances by using the MEGA 3.1 software (Kumar et al., 2004). Based on the resultant NJ trees, sequences that comprise clades, which were apparently distinct to human query gene, were excluded. In addition, a proper outgroup gene for the remaining sequences was selected. After these selections, amino acid sequences were aligned again and subjected to the program ProtTest 1.4 (Abascal et al., 2005) to choose proper amino acid substitution model for phylogenetic analysis. A maximum-likelihood (ML) tree was estimated using the TREEFINDER software package (version of June 2007; Jobb, 2007) with 1,000 approximate bootstrap tests (LR-ELW edge support: the Expected-Likelihood Weights applied to Local Rearrangements of tree topology; Strimmer and Rambaut, 2002; Jobb, 2007). When the resultant ML tree was ambiguous, a further ML analysis was performed based on nucleotide sequences and a nucleotide substitution model selected by ModelTest 3.06 (Posada and Crandall, 1998). The phylogenetic information contained within nucleotide sequences is roughly larger than that contained within amino acid sequences, and in many cases, increases the resolution of resulting phylogeny. The analysis using nucleotide sequences, however, take much time and effort, and therefore cannot be applied to all the analyses undertaken in this study.

Based on the final ML trees estimated (shown in Appendix 1: Fig. S1–S63), the genes examined were classified into four categories: (*i*) the genes that have a 1:1 orthologous relationship (Waterston et al., 2002) between human and teleosts (Fig. 2-1A); (*ii*) the genes

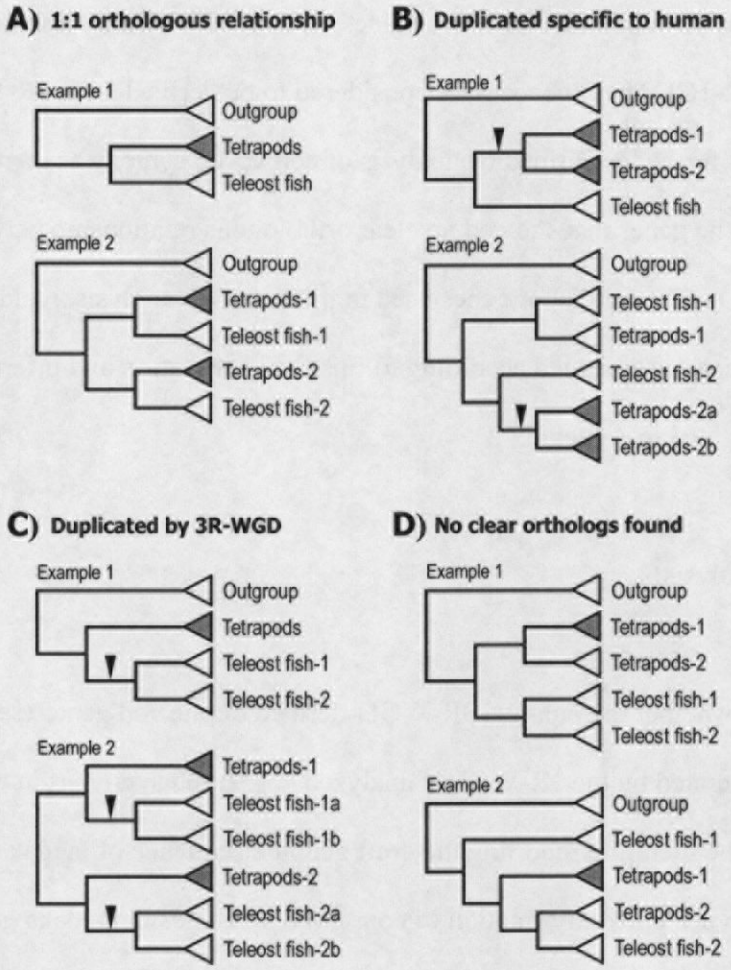

Fig. 2-1. Schematic view of classification of the identified orthologous gene groups into four categories; **A**: gene groups that show 1 to 1 orthologous relationship between tetrapods and teleost fish; **B**: gene groups that were duplicated specifically within human or tetrapod lineage; **C**: gene groups that were duplicated through third-round whole genome doubling (3R-WGD); **D**: gene groups that show no clear orthologous relationship between tetrapods and teleosts. Arrow denotes gene duplication. Triangles denote orthologous gene clades derived from each of tetrapods, teleosts, or outgroup organisms (*Ciona* and *Drosophila*). Shaded triangles denote orthologous gene clades that contains a networks-related gene of human examined in this study.

that have duplicated in human or tetrapods but not in teleosts (Fig. 2-1B); (*iii*) the genes that teleost orthologs were subdivided into two clades, in both of which the fish phylogeny reappeared (Fig. 2-1C). These genes were considered to be derived from 3R-WGD; this identification was further confirmed by analysis of conserved synteny as mentioned in the next section; (*iv*) the genes that showed no clear orthologous relationship between human and teleosts (Fig. 2-1D). The full list of genes used in the final ML analysis, including Ensembl IDs and gene names denominated according to Ensembl annotation and inferred gene phylogeny, were given in Appendix 2.

2.2.2. Synteny analysis

To confirm whether the putative 3R-WGD-derived duplicated genes (see Fig. 2-1C) were actually generated by the 3R-WGD, I analyzed conserved syntenies in the medaka genome. Within the literature reporting the draft genome sequence of medaka (Kasahara et al., 2007), the authors presented information of conserved syntenies in medaka genome in its Supplementary Table 14. They described whether medaka genes corresponding to 20,352 protein-coding genes of human were member of a block of doubly conserved synteny (DCS), which were assumed to be derived from the 3R-WGD. Using this information, I could ascertain whether the inferred teleost-specific duplicated genes were originated from the 3R-WGD, even when one of the duplicate gene was lost in medaka.

2.2.3. Measuring a number of interaction partners of the network-related proteins

For each protein involved in LTP, TT, OT, and TCA, a number of interaction partners in networks was counted on the basis of the network diagrams presented in KEGG pathway database (Kanehisa et al., 2004). In the case of proteins that were related to more than one of

the four networks (CAMK2B and 2G, PRKACB, CAMK2A and 2D, ADCY8, PLCB2, ITPR3, PRKACA and CG, and PRKX and Y) and proteins that have a general function with multiple partners within a network (calmodulin [CaM] and protein kinase [PKA]), molecular interactions that perform substantially identical function were counted as one.

For example, ADCY8 (adenylate cyclase 8 [EC:4.6.1.1]) was involved in LTP and TT, and interacts with CaM and cyclic AMP (cAMP), and G-protein and cAMP in the former and latter, respectively. In this case, the total number of interaction partners of ADCY8 was counted as three (CaM, cAMP, and G-protein). In the case of CaM, although CaM protein interacts with a number of types of proteins, a number of interaction partners of CaM was counted as two (Ca^{2+} and another protein). This is owing to that the Ca^{2+}-activated CaM has a general function as a calcium sensor and signal transducer for a multitude of different proteins. This view is supported by the fact that the amino acid sequences of CaM are extremely conserved among vertebrates (Friedberg and Taliaferro, 2005) while the interaction partners of CaM are numerous. Amino acid sequences of members of CALM1/2 family matches completely among tetrapods and teleosts. Due to the similar reason as in CaM, the number of interaction partners of PKA was counted as two (cAMP and another protein). Olfactory receptor genes (ORs) in OT and taste receptor type 2 genes (T2Rs) in TT were excluded from the analysis as an exception, because the OR and T2R gene family has preferentially expanded by specific duplications in mammals, and orthologous relationships between tetrapods and teleosts were not clear (Gilad et al., 2005; Niimura and Nei, 2005; Shi and Zhang, 2006). If the large numbers of OR and T2R gene loci were incorporated to the analysis, the resulting statistics should be biased and therefore be out of scope for this study.

2.3. Results

2.3.1. Identification of orthologous relationships between human and teleost fish gene

In this study, protein-coding genes considered to be involved in LTP, TT, OT, and TCA were analyzed to search for the possible trace of gene duplication by the 3R-WGD and lineage-specific loss of the duplicated genes. According to the KEGG pathway database (Kanehisa et al., 2004), LTP, TT, OT, and TCA were comprised of 67, 24, 32, and 27 human gene loci, respectively (OR and T2R genes were excluded; see *section 2.2. Materials and Methods*). Among these genes, 3, 8, and 5 genes were repeatedly involved in LTP and TT, LTP and OT, and LTP, TT, and OT, respectively. After removing these overlaps, the total 130 human genes (listed in Table 2-1) were subjected to comparative genomic analysis.

Putative orthologous genes in pufferfish, stickleback, medaka, zebrafish, human, chicken, clawed frog, ascidian, and fruit fly genome, which were obtained from BLAST-based database searches, were subjected to several steps of phylogenetic analysis including preliminary NJ and rigid ML analyses (for details, see *section 2.2. Materials and Methods*). As a result, orthologous relationships with teleost genes were found for 119 human genes, while no clear orthologous relationship was found for 11 human genes (PPP3CB, PPP3R2, CALM1, 2, 3, and 6, PRKCG, GUCA1C, CLCA1, 2 and 4; Appendix 1: Fig. S14, S15, S16, S25, S45, and S47; see Fig. 2-1D). Among the 119 human genes, 3 were appeared to be duplicated specific to human (Appendix 1: Fig. S7 and S53B). Such human-specific duplicated genes were considered to be derived from a single ancestral gene shared with teleosts (see Fig. 2-1B). Accordingly a total of 116 orthologous relationships between human and teleosts were identified for the network-related genes examined in this study (Table 2-2).

Table 2-1.

List of network-related genes examined in this study.

Abbreviated name: function of molecule	Name(s) of gene locus(loci) in human (related network(s); no. of interaction partners)
AMPAR: glutamate receptor	GRIA1 (LTP; 5), GRIA2 (LTP; 5)
AC: adenylate cyclase	ADCY1 (LTP; 2), ADCY3 (OT; 4), ADCY4 (TT; 2), ADCY6 (TT; 2), ADCY8 (LTP, TT; 2)
NMDAR: glutamate receptor	GRIN1(LTP; 5), GRIN2A (LTP; 5), GRIN2B (LTP; 5), GRIN2C (LTP; 5), GRIN2D (LTP; 5)
VDCC: calcium channel	CACNA1C (LTP, 1)
mGluR: glutamate receptor	GRM1(LTP; 3), GRM5 (LTP; 3)
PKA: protein kinase	PRKACA (LTP, TT, OT; 2), PRKACB (LTP, TT, OT; 2), PRKACG (LTP, TT, OT; 2), PRKX (LTP, TT, OT; 2), PRKY (LTP, TT, OT; 2)
I-1 or IPP1: protein phosphatase	IPP1 (LTP; 3)
EPAC1: guanyl-nucleotide exchange factor	EPAC1 (LTP; 2)
Rap1: GTPase	Rap1 (LTP; 3)
PP1: protein phosphatase	PPP1R12A (LTP; 2), PPP1CA (LTP; 2), PPP1CB (LTP; 2), PPP1CC (LTP; 2)
CaMK2: protein kinase	CAMK2A (LTP, OT; 3), CAMK2B (LTP, OT; 3), CAMK2D (LTP, OT; 3), CAMK2G (LTP, OT; 3)
CaN: calcium binding/protein phosphatase	CHP (LTP; 2), PPP3CA (LTP; 2), PPP3CB (LTP; 2), PPP3CC (LTP; 2), PP3R1 (LTP; 2), PP3R2 (LTP; 2)
CaM: calmodulin	CALM1 (LTP, OT; 2), CALM2 (LTP, OT; 2), CALM3 (LTP, OT;

2), CALML6 (LTP, OT; 2)

Ras: GTPase

HRAS (LTP; 2), KRAS (LTP; 2), NRAS (LTP; 2)

Raf: protein kinase

ARAF (LTP; 3), BRAF (LTP; 3), RAF1 (LTP; 3)

MEK1/2: protein kinase kinase

MAP2K1 (LTP; 3), MAP2K2 (LTP; 3)

ERK1/2: protein kinase

MAPK1 (LTP; 3), MAPK3 (LTP; 3)

Rsk: protein serine/threonine kinase

RPS6KA1 (LTP; 2), RPS6KA2 (LTP; 2), RPS6KA3 (LTP; 2), RPS6KA6 (LTP; 2)

CREB: transcription factor

ATF4 (LTP; 3)

CBP: histone acetyltransferase

CREBBP (LTP; 2), EP300 (LTP; 2)

CaMK4: protein kinase

CAMK4 (LTP; 2)

PKC: protein kinase

PRKCA (LTP; 2), PRKCB (LTP; 2), PRKCG (LTP; 2)

Gq: G protein

GNAQ (LTP; 2)

PLCβ: phospholipase, beta

PLCB1 (LTP; 3), PLCB2 (LTP, TT; 3), PLCB3 (LTP; 3), PLCB4 (LTP; 3)

IPR: calcium ion transporter

ITPR1 (LTP; 2), ITPR2 (LTP; 2), ITPR3 (LTP, TT; 2)

TAS1R: taste receptor, type 1

T1R1 (TT; 2), T1R2 (TT; 2), T1R3 (TT; 2)

Gα: G protein alpha subunit

GNAS (TT; 2), GNAT3 (TT; 4)

Gβγ: G protein beta and gamma subunit

GNB1 (TT; 4), GNB3 (TT; 4), GNG3 (TT; 4), GNG13 (TT; 4)

PDE1: phosphodiesterase 1

PDE1A (TT; 2), PDE1C (OT; 2)

CACN: calcium channel

CACNA1A (TT; 3), CACNA1B (TT; 3)

TRPM5: potential cation channel

TRPM5 (TT; 2)

KCN: potassium voltage-gated channel

KCNB1 (TT; 3)

CNG: cyclic nucleotide gated channel

CNGB1 (OT; 4), CNGA3 (OT; 4), CNGA4 (OT; 4)

Arrestin

ARRB2 (OT; 1)

GRK: receptor kinase

ADRBK2 (OT; 1)

Phd: phosducin

PDC (OT; 2)

Golf: G protein alpha subunit, olfactory type

GNAL (OT; 3)

PKG: cGMP-dependent protein kinase

PRKG1 (OT; 2), PRKG2 (OT; 2)

GCAP: guanylate cyclase activator

GUCA1A (OT; 2), GUCA1B (OT; 2), GUCA1C (OT; 2)

pGC: guanylate cyclase

Gucy2d (OT; 3), Gucy2f (OT; 3)

CLCA: calcium-dependent chloride channel

CLCA1 (OT; 3), CLCA2 (OT; 3), CLCA4 (OT; 3)

PYC: pyruvate carboxylase

PYC (TCA; 2)

PCK: phosphoenolpyruvate carboxykinase

PCK1 (TCA; 2), PCK2 (TCA; 2)

MDH: L-malate dehydrogenase

MDH1 (TCA; 2), MDH2 (TCA; 2)

FH: fumarate hydratase

FH (TCA; 2)

SDH: succinate dehydrogenase

SDHA (TCA; 2), SDHB (TCA; 2), SDHC (TCA; 2), SDHD (TCA; 2)

SUCLG: succinate-CoA ligase

SUCLG1 (TCA; 2), SUCLG2a (TCA; 2), SUCLG2b (TCA; 2)

SUCLA2: succinate-CoA ligase

SUCLA2 (TCA; 2)

CS: citrate synthase

CS (TCA; 2)

ACLY: ATP citrate synthase

ACLY (TCA; 2)

CLYBL: citrate lyase

CLYBL (TCA; 3)

ACO: aconitase

ACO1 (TCA; 2), ACO2 (TCA; 2)

IDH: isocitrate dehydrogenase (NADP+)

IDH1 (TCA; 2), IDH2 (TCA; 2)

IDH3: isocitrate dehydrogenase (NAD+)

IDH3A (TCA; 2), IDH3B (TCA; 2), IDH3G (TCA; 2)

OGDH: oxoglutarate dehydrogenase

OGDH (TCA; 2), OGDHL (TCA; 2)

DLD: dihydrolipoyl dehydrogenase

DLD (TCA; 2)

DLST: dihydrolipoyllysine-residue succinyltransferase

DLST (TCA; 3)

Abbreviations: LPT, long-term potentiation; TT, Taste transduction; OT, olfactory transduction; TCA, TCA cycle

Table 2-2.

Number of orthologous relationships between human and teleosts identified in this study.

	LTP	TT	OT	TCA	Grand total[a]
No. of network-related loci in human	67	24	32	28	130
Duplicated specific to tetrapods or human	3	2	2	1	3
No clear orthologs	6	0	8	0	11
No. of orthogous relationships identified	58	22	22	27	116

[a]An overlap of the genes that were involved in more than one network was controlled.

Abbreviations: LPT, long-term potentiation; TT, Taste transduction; OT, olfactory transduction; TCA, TCA cycle

2.3.2. Persistence rate of duplicated genes generated by the 3R-WGD

The careful and systematic analyses of molecular phylogeny of each gene family mentioned above verified gene duplication events during evolution of tetrapods and teleosts. Along with available information of doubly conserved synteny block from medaka genome (Kasahara et al., 2007), gene duplication events accompanied by the 3R-WGD were reliably identified (for details, see Appendix 1: Fig. S1—S63). A schematic plot, derived primarily from this result, was shown in Fig. 2-2. This figure indicates the presence or absence of gene locus/loci belonging to each of the identified orthologous gene groups in genomes of the species examined. Circles colored in magenta and green denote the genes belong to the orthologous gene groups in which duplication by the 3R-WGD was detected, and the uncolored (gray) circles denote the genes belong to the orthologous gene groups in which duplication by the 3R-WGD was not detected. These data are summarized in Table 2-3; forty-five of the 116 orthologous gene groups exhibited gene duplication by the 3R-WGD, indicating that the genome of common ancestor of zebrafish, medaka, stickleback, and pufferfish harbored the 161 (116+45) gene loci belonging to the 116 orthologous gene groups. It deduces that 56.5% ([45×2]/161) of the gene loci in common ancestor of the four teleosts were duplicated as a result of the 3R-WGD (see Table 2-3).

Fig. 2-2 also present information of secondary loss of the 3R-WGD-derived duplicate gene in the extant teleost genomes. For example, zebrafish do not have one of the duplicated pair of PPP1CA in LTP. Considering these lineage-specific absence of genes, a number of total gene loci and 3R-WGD-derived duplicate gene loci in each of the four teleost genome were counted (Table 2-4); the four teleost genomes possessed average 133 gene loci belonging to the 116 orthologous gene groups. Among them, average 55 (55/133=41.4%) loci were remained in duplicate since the 3R-WGD (see Table 2-4).

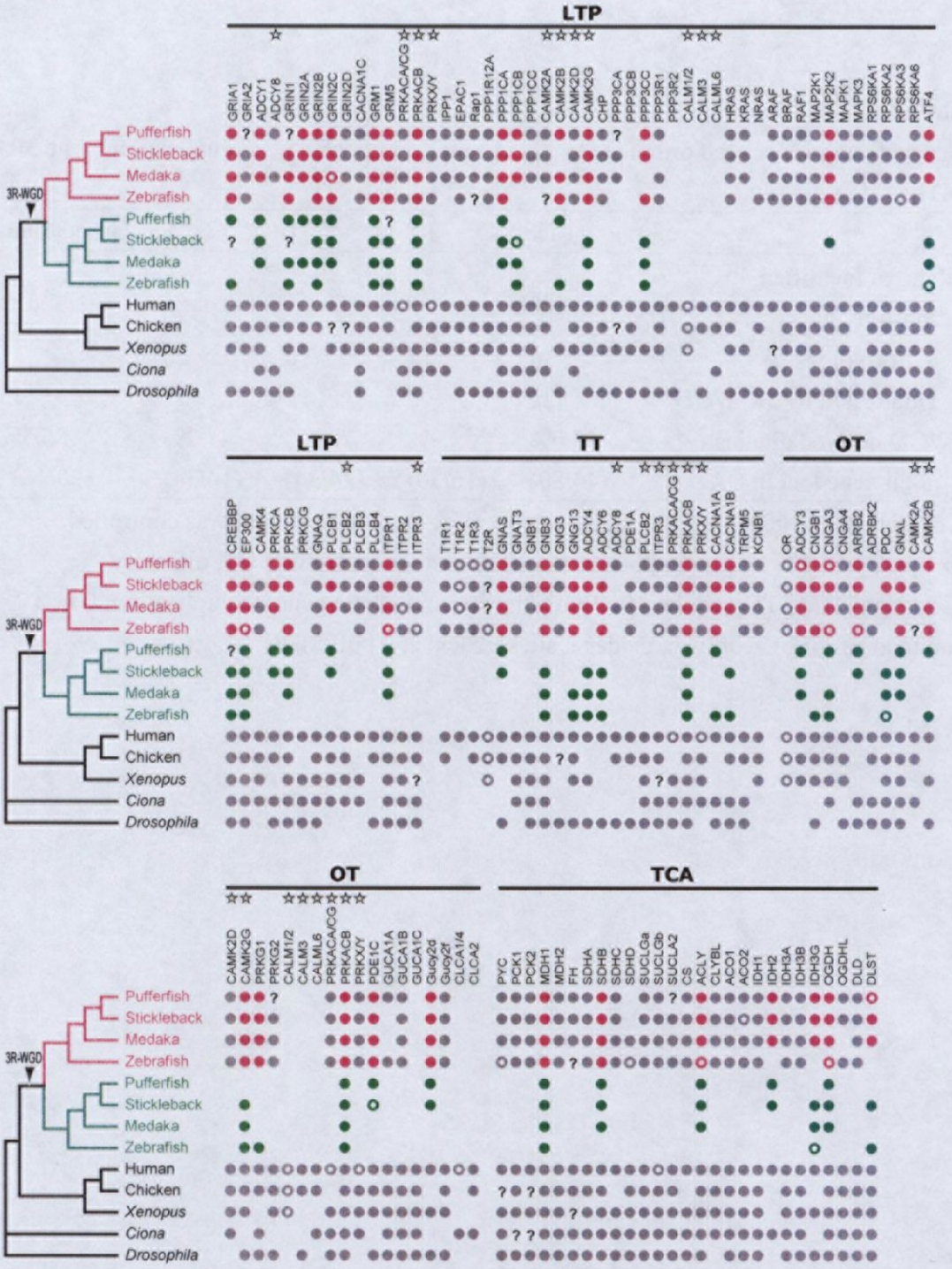

Fig. 2-2. A schematic representation of the results of the database mining and phylogenetic analysis of this study. Circle denotes the gene locus/loci belonging to each of the identified orthologous gene groups for each species examined. Circles colored in magenta and green denote the genes belonging to the orthologous gene groups in which duplication by the 3R-WGD was detected. Open circles denote the gene loci that were duplicated specifically within the species. Question marks show that a partial sequence of putative ortholog was found but phylogenetically unsorted. Star marks denote the gene groups that were involved in more than one network examined. *Abbreviations*: LPT, long-term potentiation; TT, Taste transduction; OT, olfactory transduction; TCA, TCA cycle; 3R-WGD, third-round whole genome doubling.

Table 2-3.

Number of network-related orthologous gene groups, in which a gene duplication by he 3R-WGD was detected.

	LTP	TT	OT	TCA	Grand total[a]
Total no. of identified Relationships	58	22	22	27	116
1:1 orthologs	36	14	10	20	71
Duplicated by 3R-WGD	22	8	12	7	45
3R-WGD-derived duplicate-loci / total gene loci in CA	55.0% (44/80)	53.3% (16/30)	70.6% (24/34)	41.2% (14/34)	56.5% (90/161)

[a]An overlap of the genes that were involved in more than one network was controlled.

Abbreviations: LPT, long-term potentiation; TT, Taste transduction; OT, olfactory transduction; TCA, TCA cycle; 3R-WGD, third-round whole genome duplication; CA, a common ancestor of zebrafish, medaka, stickleback, and pufferfish

Table 2-4.

Total number of network-related gene loci and 3R-WGD-derived duplicated loci identified in each of the teleost fish genome.

	LTP		TT		OT		TCA		Grand total[a]	
	No. of loci	3R-derived	No. of loci	3R-derived	No. of loci	3R-derived	No. of loci	3R-derived	No. of loci	3R-derived
CA	58	NA	22	NA	22	NA	27	NA	116	NA
Pufferfish	67 (67)	28 (28)	24 (27)	6 (6)	29 (31)	16 (18)	30 (31)	10 (10)	134 (140)	54 (56)
Stickleback	72 (73)	36 (37)	27 (34)	12 (12)	27 (33)	16 (22)	33 (34)	14 (14)	145 (160)	72 (79)
Medaka	71 (73)	30 (31)	25 (27)	10 (10)	26 (26)	12 (12)	31 (31)	10 (10)	136 (140)	56 (57)
Zebrafish	55 (61)	22 (23)	24 (25)	10 (10)	26 (29)	12 (13)	26 (31)	2 (2)	117 (131)	38 (40)
Average	66.3 (68.5)	26.5 (29.8)	25.0 (28.3)	9.5 (9.5)	27.0 (29.8)	14.0 (16.3)	30.0 (31.8)	30.0 (31.8)	133.0 (142.8)	55.0 (58.0)

Numerals in parentheses indicate the number of gene loci when the lineage-specific gene duplications were incorporated.

[a]An overlap of the genes that were involved in more than one network was controlled.

Abbreviations: LPT, long-term potentiation; TT, Taste transduction; OT, olfactory transduction; TCA, TCA cycle; CA, a common ancestor of tetrapods and teleost fish; 3R, third-round whole genome duplication; NA, not applicable

From the information of presence/absence of the 3R-WGD-derived duplicate loci (see fig. 2), numbers of gene loss events were parsimoniously drawn on the basis of teleost phylogeny. The inferred numbers were then assigned to the phylogenetic tree with branch lengths proportional to estimated divergence time derived from a molecular clock analysis of mitochondrial genome sequences (Azuma et al., unpublished; fig. 3A). In these inferences, if one of the pair of 3R-WGD-derived duplicates was absent in zebrafish, but both were present in medaka and stickleback and/or pufferfish, "−1" was assigned to the branch between node c and zebrafish (for example, PPP1CA locus of LTP). If both of the pair were present in zebrafish, but one of the pair was absent in medaka, stickleback, and pufferfish, "−1" was assigned to the branch between node c and b.

The above assignment yielded the total number of the 3R-WGD-derived duplicate gene pairs persisted in each internal node. These data were tied up to estimated divergence time between teleost lineages inferred from a molecular clock analysis of mitochondrial genome sequences (Azuma et al., unpublished). Then, the derived two-dimensional data points, and the estimated occurrence time of the 3R-WGD (average 353.0 MYA; Christoffels et al., 2004; Vandepoele et al., 2004; Christoffels et al., 2006) and data points of current (0 MYA) teleost genomes (Table 2-4), were approximated to the neutral model of loss-of-function of duplicated genes, $e^{-2\mu t}$, where μ is the null mutation rate and the t is the number of generations since the duplicated genes have become independent loci (Nei and Roychoudhury, 1973; Force et al., 1999; Borenstein et al., 2007). In this approximation, absolute time (year) was used alternative to the number of generations. Least square fitting of the equation to the data points of 3R-WGD and node c yielded an exponential decay curve; its slope α was 1.4936 (Fig. 2-3B gray line). On the other hand, approximation to all data points yielded a relatively moderate curve; its slope α was 26.05 (Fig. 2-3B black line).

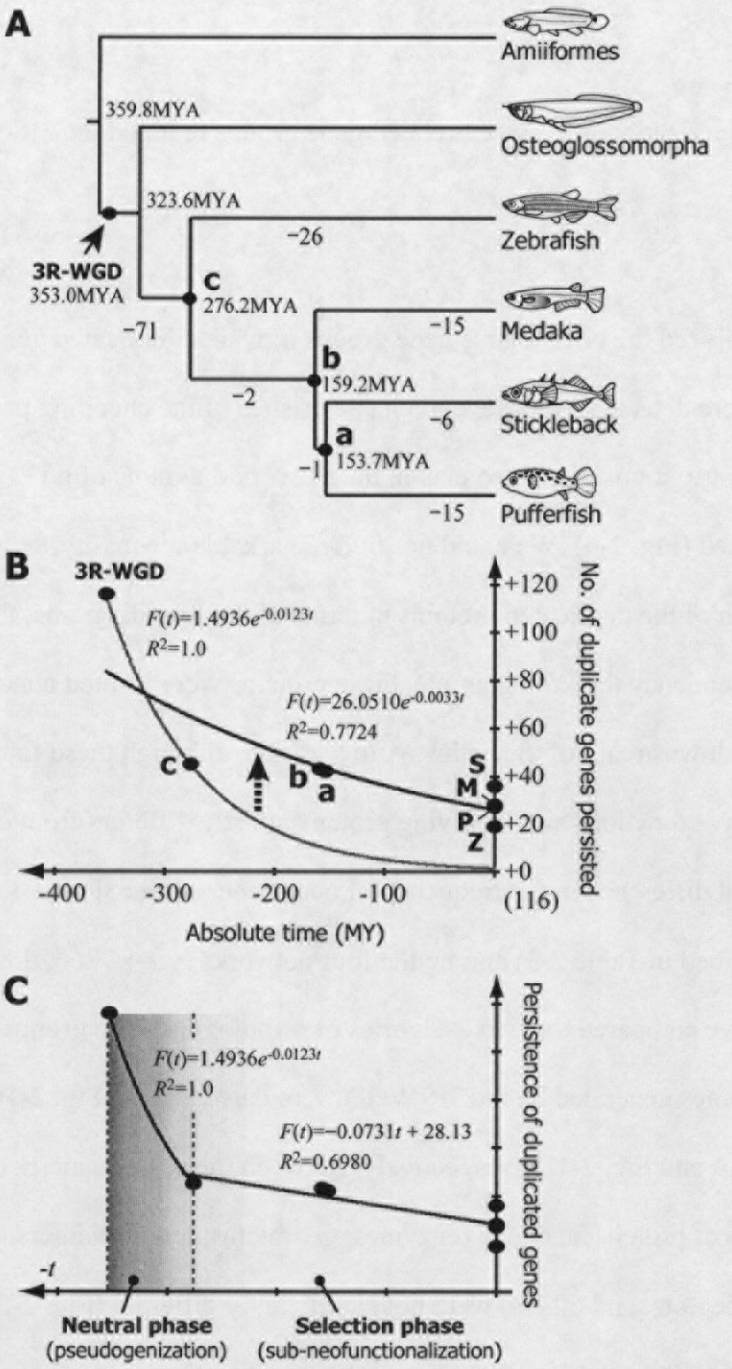

Fig. 2-3. The estimated process of gene loss after the occurrence of third-round whole genome doubling (3R-WGD) in teleost ancestor. (**A**) Numbers of gene loss events, which were drawn from presence/absence data of 116 orthologous genes groups, were assigned to phylogenetic tree of the four teleosts. Branch lengths are proportional to estimated divergence time among lineages (Azuma et al., unpublished). (**B**) An approximation to the neutral model of loss of gene function of duplicated genes, $e^{-2\mu t}$ (Nei and Roychoudhury, 1973; Force et al., 1999). The gray and black lines show the approximation from 3R-WGD and point c, and all data points, respectively. (**C**) The supposed phase transition in duplicate gene preservation, which is reasonably understood by the model of sub-neofunctionalization of duplicate gene evolution (Force et al., 1999; Lynch and Force, 2000; He and Zhang, 2005; Rastogi and Liberles, 2005). In this panel, data points except for the 3R-WGD were approximated to linear equation.

2.3.3. Characterization of network-related gene groups retained the 3R-WGD-derived duplicates

We analyzed the orthologous gene groups in which duplicated genes generated by the 3R-WGD were detected, from several characteristics of the encoding protein. First, the locational distributions of the proteins in the network diagrams of LTP, TT, OT, and TCA were examined (Fig. 2-4). We could not find remarkable inequality, segregation, or concentration of the duplicated proteins in terms of the four diagrams; there was no remarkable tendency that, for example, these proteins were located concentratedly on either upstream or downstream of the pathway. In addition, although these four network systems have respective functions and involving proteins mostly different from each other, there was no significant difference in the frequency of occurrence of the 3R-WGD-derived duplicated genes (described in Table 2-3) among the four networks (χ^2=1.7266, d.f.=3, P=0.6310).

Next, we compared the two categories of orthologous gene groups in which the duplicated genes generated by the 3R-WGD were detected (see Fig. 2-1C) and not detected (see Fig. 2-1A and Fig. 2-1B), respectively. Between them, frequencies of occurrence of the several types of protein functions (enzymes, G proteins, ion implanters, phosphorylation enzymes, receptors, and others) were not significantly different (Fig. 2-5A; χ^2 = 4.3984, d.f. = 5, P = 0.4936). However, length of the protein coding-sequence was clearly different. The orthologous gene groups in which the 3R-WGD-derived duplicated genes were detected were more frequently appeared in the class of long size (>1000 amino acids) and less in the class of short size (<200 amino acids) on the basis of the human protein data as a hypothetical ancestral state (Fig. 2-5B; χ^2 = 10.4317, d.f. = 2, P = 0.0052). An average number of interaction partners of the encoding protein in molecular interaction networks (for detailes, see *section 2.2 Materials and Methods*) was also significantly larger in the gene groups in which the 3R-WGD-derived duplicated genes were detected (Table 2-5).

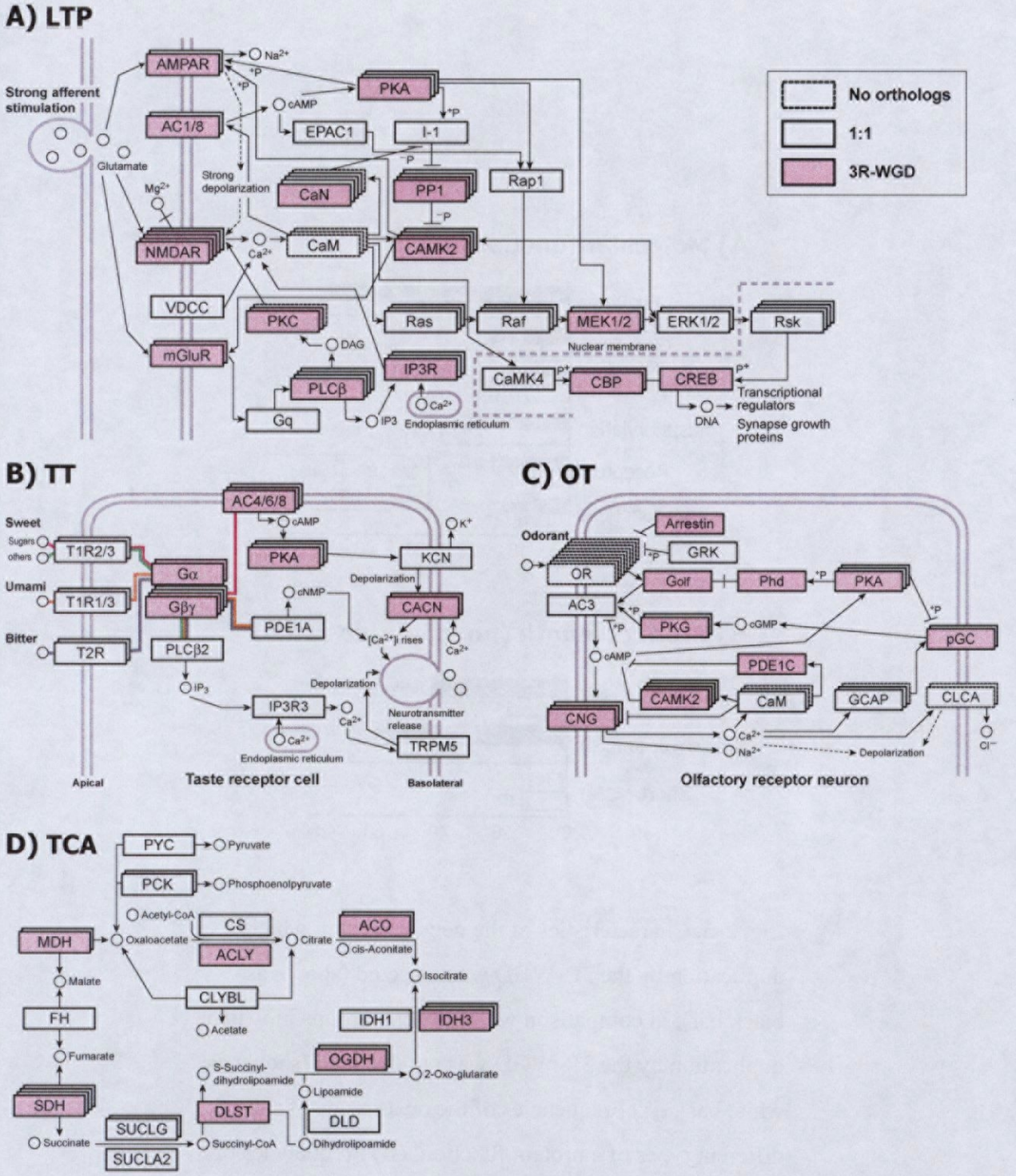

Fig. 2-4. Locational distributions of network-related gene groups, in which duplicate genes generated through third-round whole genome doubling (3R-WGD) were detected (shown as boxes colored in pink). Panel **A**, **B**, **C**, and **D** show the network diagram of long-term potentiation (LTP), taste transduction (TT), olfactory transduction (OT), and TCA cycle (TCA), respectively. White boxes denote the gene groups in which 1 to 1 orthologous relationship was detected between tetrapods and teleosts. Dashed boxes denote the gene groups in which no clear orthologous relationship was detected between tetrapods and teleosts.

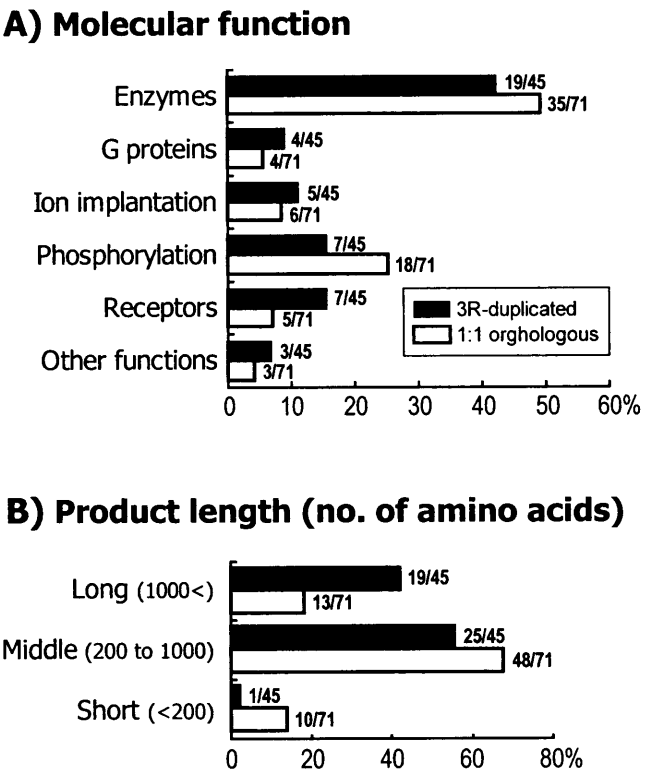

Fig. 2-5. Characteristics of the gene groups in which duplication by the 3R-WGD was detected (shown as black bar), in comparison with the gene groups in which duplication by the 3R-WGD was not detected (shown as white bar). (**A**) Frequencies of the occurrence of six different types of a protein function. (**B**) Frequencies of the occurrence of three classes of a product length (no. of amino acids).

Table 2-5.

Difference in numbers of interaction partners in molecular interaction networks examined.

Molecular interaction network	No. of interaction partners		*P*-value	
	1:1[b]	3R-WGD[c]	Student's *t*	Welch's *t*
Long-term potentiation	2.50 ± 0.14 (*n*=36)	3.00 ± 0.25 (*n*=22)	0.0627	0.0924
Taste transduction	2.62 ± 0.23 (*n*=14)	2.75 ± 0.31 (*n*=8)	0.6460	0.6518
Olfactory transduction	2.40 ± 0.27 (*n*=10)	2.75 ± 0.28 (*n*=12)	0.3810	0.3750
TCA cycle	2.05 ± 0.05 (*n*=20)	2.14 ± 0.14 (*n*=7)	0.4390	0.5576
Grand average[a]	2.40 ± 0.09 (*n*=71)	2.80 ± 0.16 (*n*=45)	0.0183*	0.0288*

*$P<0.05$

[a]An overlap of the genes that were involved in more than one network was controlled.

[b]Gene groups in which 1 to 1 orthologous relationship between tetrapopds teleost fish was detected.

[c]Gene groups in which duplicated genes generated through 3R-WGD were detected.

2.4. Discussion

2.4.1. Contribution of the 3R-WGD to the genomic composition of teleost fishes

In this study, a careful analysis of more than 100 gene families in the zebrafish, medaka, stickleback, and pufferfish genomes, which have experienced fish-specific genome doubling or the 3R-WGD, revealed that the genome of the common ancestor of the above four teleost fish contained a large number of duplicated gene loci (90/161=56.5%) persisted since the 3R-WGD (Table 2-3). This new estimate was derived by extensive data mining of a genomic database, and a latest maximum-likelihood phylogenetic analysis of each of gene family that are involved in the system of LTP, TT, OT, and TCA. Given the above-estimated value represents the whole genome, it is suggested that the large number of 3R-WGD-derived duplicate genes occupied more than half of the protein-coding loci of the common ancestor and underwent lineage-specific gene loss or retention along with the diversification of teleosts. Such lineage-specific evolution of a large number of 3R-WGD-derived genes should have been involved in the differential genomic constitution among teleosts. These genetic differences derived from the 3R-WGD would be correlated with the diversity of teleosts, which have been expanded to include more than 26,000 species and 500 families from 40 orders (Nelson, 2006).

The present study not only demonstrated that a large number of 3R-WGD-derived duplicated genes persisted in the common ancestor of the four teleosts, but also revealed that many of the duplicate genes have been retained in current genomes of teleosts. According to the estimates shown in Table 2-4, gene loci duplicated through the 3R-WGD comprises 41.4%, on average, of the protein-coding loci within teleost genomes. This estimate is higher than that reported in previous studies based on a pairwise comparison between human versus teleost genome; 15.2% ([2134×2]/28005) to 19.2% ([2009×2]/20899) (these reference values

were according to Kasahara et al., 2007 and Ensemble Genome Browser, July 2007). I suggest a plausible reason why I obtained higher persistence rate for genes than previous studies as follows. The analytical procedure carried out in this study could have identified gene duplication events more effectively, because genome sequence data from multiple teleost species and rigorous molecular phylogenetic methods were used. Previous studies may underestimated the contribution of 3R-WGD to the formation of teleost genomes, probably because of a semi-automatic analysis based on BLAST searches which is less sensitive than the analysis implemented in this study. The present study could, therefore, show that the 3R-WGD have contributed largely to the current genomic constitution of teleosts. This conclusion emphasizes the significance of the 3R-WGD in the evolution of teleosts, and further, the importance of ancient WGD events in the evolution of vertebrates.

According to the findings in this study, the number of protein coding genes in teleost genomes is expected to be larger than that in tetrapod genomes. Such increase in the number of genes, however, was not found in the public genome data of teleosts, as mentioned in the Introduction (mammals: 22,000 in average among human, mouse, dog, and cow; teleosts: 23,000 in average among pufferfishes, stickleback, medaka, zebrafish; Ensembl Genome Browser, July 2007). This unexpected similarity of number of genes in teleosts and tetrapods may attributed to the genes putatively specific to tetrapods, in which the clear orthologs of teleosts were not found. Eleven of 130 (8.5%) human genes fell into this category in the present analysis (Table 2-2 and also Fig. 2-2). These genes may underlie some specific traits of human or tetrapods. On the other hand, the data mining process of this study started with the knowledge of biological pathways in human. Therefore, many of the teleost-specific genes may have not been discovered. If such genes exist, teleosts should possess a larger number of genes than tetrapods. This notion seems to be consistent with the estimation that the number of protein-coding genes in pufferfish (*Tetraodon nigroviridis*) genome is much larger than in human genome (28,000 versus 22,700; Ensemble Genome Browser, July 2007), although the

genome size of *T. nigroviridis* is much smaller than human (C value of 0.35 and 3.50, respectively; Gregory et al., 2007). An adequate answer to the question whether the teleost genome harbor more genes than tetrapod genome would be provided from further larger scale analysis of multiple teleost genomes.

The process of gene loss after the 3R-WGD estimated on the basis of teleost phylogeny provided a new evidence compatible with the sub-neofunctionalization model of duplicate gene evolution. As indicated in Fig. 2-3B, 61.2 % (71/116) of duplicated genes arose through the 3R-WGD were lost during the initial 75 MY before the divergence of zebrafish (Fig. 2-3B gray line), whereas the genes remained in the node c rather persisted for about 280 MY (Fig. 2-3B black line). This seems not to be compatible with the neutral model of loss of gene function (Nei and Roychoudhury 1973; Force et al. 1999), which predicts the exponential loss of genes after duplication. Alternatively, the above trajectory can be reasonably understood on the basis of the sub-neofunctionalization model of fate of duplicated genes (Force et al., 1999; Lynch and Force, 2000A; He and Zhang, 2005; Rastogi and Liberles, 2005), indicating the validity of this model. That is, many of duplicated genes are neutrally and thus rapidly lost following a WGD, whereas the genes that persisted for a long period of time have acquired a new role via sub-neofunctionalization, and thus maintained by natural selection (Fig. 2-3C).

2.4.2. Subfunctionalization of duplicate genes and the evolution of teleost genomes

It was further proposed that subfunctionalization has led to the long-time persistence of the 3R-WGD-derived duplicate genes, based on other lines of evidence focusing on the multifunctionality of encoding proteins. Multifunctionality of gene and/or proteins is assumed to be highly correlated with the occurrence of subfunctionalization. The reason for this assumption is that subfunctionalization depends on mutations that leads to loss of subsets of gene function (Force et al., 1999; Lynch and Force, 2000A). Therefore, genes with many

functions are expected to have higher probability to undergo subfunctionalization after duplication (Force et al., 1999). The results shown in Fig. 2-5B and Table 2-5 go along with this prediction; protein properties probably linked to multifunctionality, the length of the protein sequence (longer protein sequence contains more protein domains and motifs) and a total number of interaction partners, were significantly longer and larger in gene groups in which the 3R-WGD-derived duplicates have been retained. This implies that the 3R-WGD-derived duplicate genes that persisted for long evolutionary periods of time were maintained through subfunctionalization. This concept of subfunctionalization has been widely accepted in the studies of duplicated gene evolution (Prince and Pickett, 2002; Zhang, 2003; Hughes, 2005). However, most of the empirical data supporting this subfunctionalization model were obtained from the spatiotemporal pattern of expression or repertoire of protein domains of particular genes (Altschmied et al., 2002; Yu et al., 2003; de Souza et al., 2005; Tocchini-Valentini et al., 2005). Therefore, this study provide a new data supporting the validity of subfunctionalization model from the analysis of a large number of gene families and a multifunctionality of a protein.

If the larger proteins are selectively maintained via subfunctionalization following WGD as suggested above, the average size of proteins in a genome become larger every time after WGD. At present, there is no actual evidence for such tendency. On the contrary, subfunctionalization in coding sequence of duplicated genes often accompanies truncations or deletions of protein domains or motifs, according to hitherto studies (Altschmied et al., 2002; Yu et al., 2003; de Souza et al., 2005). If this phenomenon is widespread, there would be no trend toward increasing average size of the proteins through WGD events. To pursue the above possibility further, however, it will be interesting to compare the genome sequences of two related-species with and without experiences of WGD events, such as yeasts (*Saccharomyces cerevisiae* versus *Kluyveromyces waltii*; Kellis et al., 2004) and vertebrates (e.g., tetrapods versus teleosts; Jaillon et al., 2004; Kasahara et al., 2007).

Moreover, if the proteins with more functions were selectively maintained via subfunctionalization, a number of functions per protein should be decreased each time after a WGD occurs. This will result in a relaxation of functional constraint on a gene, leading to increase in the rate of molecular evolution. Interestingly, it was reported that the molecular evolutionary rates of nuclear genes in teleosts are generally faster than in tetrapods based on pairwise analyses of 6,000 to 9,000 orthologous gene pairs (Jaillon et al., 2004; Kasahara et al. 2007). Generation time effect may be responsible for this acceleration, however, the 3R-WGD event followed by subfunctionalization of a large numbers of duplicated genes may also related to the acceleration by bringing relaxation of functional constraint as suggested above. To examine this possibility, extensive assessment of evolutionary rates of the 3R-WGD-derived duplicate genes would be needed. It can be also hypothesized that a number of involving-genes per function was increased in teleosts via subfunctionalization following the 3R-WGD. Therefore, teleosts may be constrained to use substantially more genes than commonly used in tetrapods to perform a same function. This implies that teleosts may be burdened with many genes with a narrower function than in tetrapods, because of the increasing cost of regulation, transcription, and translation. Teleosts may tolerate such a "cost of subfunctionalization", whereas the 3R-WGD and subsequent subfunctionalization may have contributed to teleost-specific neofunctionalization of genes, and consequently, the genetic and physiological diversity among teleost species.

2.4.3. Gene families expanded and evolved independent of WGD

Among network-related gene groups examined in this study, a few gene families were found to have expanded dramatically by numbers of gene duplication independent of the 3R-WGD. One representative of them is the well known OR (olfactory receptor) gene family. OR gene family in vertebrate is composed of about ten subfamilies (Niimura and Nei, 2005), and

the different subfamily has expanded enormously in each genome of tetrapods or teleosts (Gilad et al., 2005; Niimura and Nei, 2005). Furthermore in the OT (olfactory transduction) system, phosphodiesterase 1C (PDE1C) genes were found to be duplicated specifically in the stickleback lineage for six times, surprisingly (see Appendix 1: Fig. S44). PDE is a key enzyme in the cyclic AMP pathway, as it regulates the localization, duration, and amplitude of intracellular cAMP signaling (Jeon et al., 2005). Sticklebacks are therefore may have multiple bypass circuits to regulate the cAMP pathway probably in the OT system, via sub-neofunctionalization of these highly duplicated PDE1C genes. These multiple PDE1C genes may be a genetic basis for some adaptive traits of stickleback.

In the CaM (or CALM) gene family, which is involved in LTP and OT, there was no clear orthologous relationship identified between tetrapods and teleosts (Appendix 1: Fig. S16). The estimated phylogeny for CaM gene family suggests that the family had consisted of diverse members before the split of tetrapods and teleosts, and distinct members have persisted in current genomes of tetrapods or teleosts, respectively. Despite such a long-term historical divergence, amino acid sequences were completely conserved among the CALM1 and 2 of tetrapods and their closest CALMs of teleosts (data not shown; see also Friedberg and Taliaferro, 2005). This suggests that protein structures of these CALMs have been under strong purifying selection and thus strictly conserved for hundreds of million years. This interesting, unique characteristic and history of CaM gene family may provide a good model system for studying neofunctionalization through changes in gene expression and regulation, but not in coding sequence or protein structure.