東京大学大学院新領域創成科学研究科
情報生命科学専攻

平成 17 年度

修士論文

メダカ遺伝子の比較解析

Comparative Analysis of Medaka Genes

2006 年 3 月提出
指導教員　森下　真一　教授

47-46911　曲　薇（Wei Qu）

Contents

Authors

Wei Qu[1], Yoichiro Nakatani[1], Ahsan Budrul[1], Kiyoshi Naruse[2], Masahiro Kasahara[1], Shin Sasaki[1], Yukinobu Nagayasu[1], Tomoyuki Yamada[1], Koichiro Doi[1], Takanori Narita[3], Tadasu Shin-I[3], Shinichi Hashimoto[5], Asao Fujiyama[4], Yuji Kohara[3], Hiroyuki Takeda[2], Shinichi Morishita[1]

1 Department of Computational Biology, Graduate School of Frontier Sciences, the University of Tokyo,

2 Department of Biological Sciences, Graduate School of Science, the University of Tokyo,

3 Center for Genetic Resource Information, National Institute of Genetics,

4 National Institute of Informatics,

5 Department of Molecular Preventive Medicine, School of Medicine, the University of Tokyo

Introduction

Fish is the biggest group in Vertebrata. Currently, more than 28,000 species of fish have been identified. Not only is fish provided as food, it is a good model organism as well, especially in developmental biology and environment research. With the power of forward genetics and the elegant embryo manipulation techniques, the small teleost species which are a large group of fishes with bony skeletons and including most common fishes, have been used for various aspects of basic biology.

Furthermore, genomic data of fish have an exponential increase in the recent ten years. These data present us a much wider point of view about the divergence of fish species via comparative genomic approaches. Fig. 1 shows a phylogenic tree of vertebrate species. A whole genome duplication event is supposed to have happened in fish more than 110My ago. Since two copies of many genes are present in bony fishes while only one in other vertebrate, several reports indicated the ancestor of fish and mammals only had half of chromosomes of present fish's.

Medaka, *Oryzias latipes*, is a small, egg-laying freshwater fish native to Asia that is found primarily in Japan, but also in Korea and eastern China. As a far eastern cousin of zebrafish, which is a most familiar experimental fish and has already secured its place in the field of developmental genetics, medaka is emerging as an important model fish with increasing genetic toolkit[1]. The physiology, embryology and genetics of medaka have been extensively studied for the past 100 years as well. In evolutionary point of view, medaka holds an important position in the teleost lineage, being in between zebrafish and *Takifugu*. The last common ancestor of medaka and zebrafish lived more than 110 - 160 million years ago[2], and the phylogenic distance of medaka to *Takifugu* is much closer than that of zebrafish to *Takifugu*.

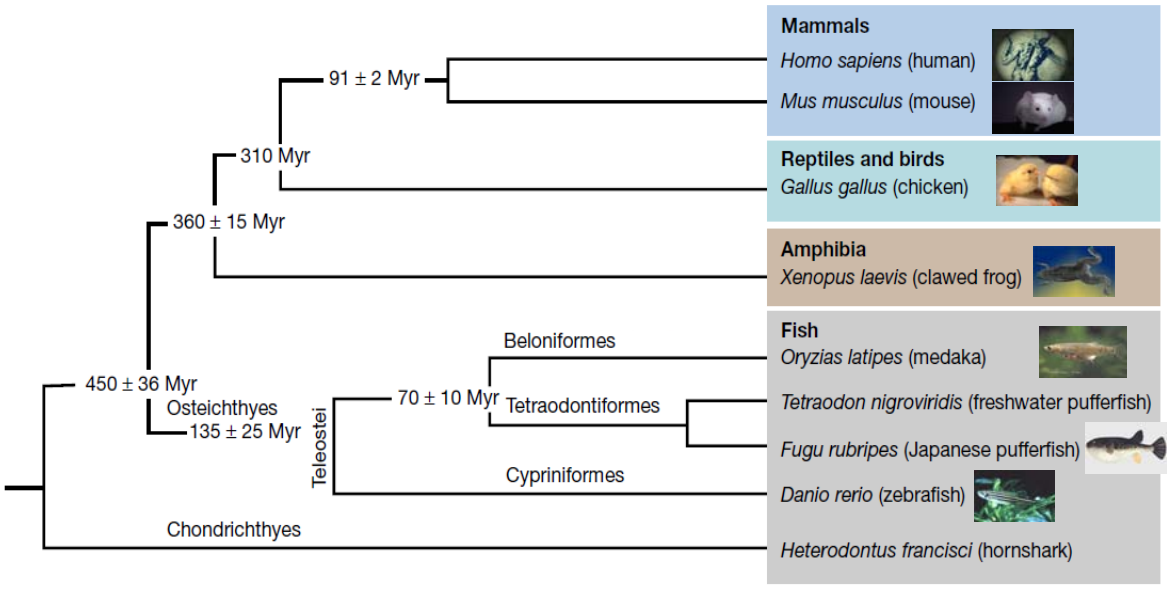


Figure 1. Shima, A. *et al.* (2003) Fish genomes flying. EMBO reports, 4, 2:121-125. Pictures are

obtained from the UniGene web site.

Access to whole genome sequences is an extremely important step to understand how genetic information is stored, organized and has evolved in DNA sequences. Compared to zebrafish's genome 1,700Mb and *Takifugu*'s 390Mb that is among the smallest vertebrate genomes, medaka's genome is supposed to be 700-800Mb. We started the MEDAKA GENOME SEQUENCING PROJECT in 2002, and expected that more detailed genetic information would be offered with high accuracy. We adopted the whole genome shotgun approach and chose the inbred strain, Hd-rR as a main target for sequencing because it is derived from the southern Japanese population and most medaka mutants are of southern origin. I am involved in this MEDAKA GENOME SEQUENCING PROJECT, entirely responsible for comparative analysis of medaka genes to other fish species' genes and human genes. Biological experiment part was completed by our collaboration laboratories.

Here I report comparative analysis of medaka genes' evolution, which is mainly based on the analysis on draft genome sequence of medaka and prediction of medaka genes.

Abstract

We started the MEDAKA GENOME SEQUENCING PROJECT in 2002. I am involved in this project, entirely responsible for comparative analysis of medaka genes to other fish species' genes and human genes. The assembly of the genome sequence of medaka came over its final stage and 20,141 gene clusters of medaka were identified by a novel method, which took advantage of the comprehensive transcription start site information collected by the high-throughput 5'SAGE method. I performed a comparative analysis on the evolution of medaka genes and provided a further perspective on vertebrate evolution.

Annotation of genes

Among the 20,141 gene clusters predicted, over two thousand genes have no homologues with other fish genes, amphibian genes, human genes, *Takifugu* genome or medaka ESTs. About 67% of Medaka genes possess homologues and 58% have strong corresponding orthologues with human Refseq genes.

Half of the human disease genes have medaka orthologues, which represents the importance of the medaka fish as a model organism in experimental medical science especially in the embryology field.

Experimental evidence for novel genes

Dozens of these novel genes were confirmed by RT-PCR, TA cloning and *in situ* hybridization. Furthermore, a pioneer morpholino-based gene knockdown experiment to elucidate the function of these novel genes was designed and performed.

Evolution of orthologues according to Gene Ontology (GO)

2,292 of the 4,342 medaka-human 1:1 orthologues could be identified with one or multiple GO 'biological process' IDs. Orthologues involved in carbohydrate metabolism, alcohol metabolism, and catabolism were more conserved than those implicated in immune response, transcription, apoptosis, DNA repair and response to stress. This result was similar to that from the comparison of 1:1 chicken-human orthologues that genes related to adaptation to the environment seem to be less conserved in their protein-coding sequences.

Genome size difference in Fish

I computed gene size ratios of reciprocal best matches of medaka and *Takifugu*. The average ratio is about 3 the median ratio is about 2, which is supposed to be an important factor effects about 1 time increase of genome size from *Takifugu* to medaka. This helped us to understand how the remarkable difference of genome size arose in fishes.

Evolution of Paralogues

I investigated the evolutionary relationships of medaka genes to *Tetraodon* genes and their paralogous pairs, and I found large portion of medaka paralogous pairs are conserved as paralogous pairs *Tetraodon*. These "pair-to-pair" paralogous pairs involved in transport, catabolism, alcohol metabolism, transcription factor NF-*kappa*B, carbohydrate metabolism *etc*. were significantly conserved than those implicated in RNA metabolism, DNA metabolism, transcription, response to DNA damage stimulus, phosphorus metabolism *etc.* , which is quite associated with the fore mentioned analysis of evolution of medaka-human orthologues.

Similarity of expression pattern

I studied the similarity of expression pattern between medaka duplicate genes. A positive correlation of expression levels was found. However, more or less expression divergences were detected in quite a large portion of duplicated genes, which enables tissue or developmental specialization to evolve.

Local gene duplications

Local gene duplications were detected by comparing each gene with its preceding genes and subsequent genes on the chromosome. The number of identified clusters was considerably lower than that in mammals. There are no big clusters of immunoglobulin or olfactory receptor. Keratin genes, which form the scales of fishes, account for a large proportion.

Genome duplication

1,730 pairs of medaka 1:1 duplicated genes were identified, which is a clear signature of whole genome duplication in medaka. With the homologous information of 1:1 human-medaka orthologues, 572 human-medaka synteny blocks were identified. Ancestor chromosomes' blocks on the human chromosomes were detected manually, which is another evidence for the whole genome duplication in fishes.

Transcriptome map

Over one million 5' SAGE (serial analysis of gene expression) tags were collected and 90% of them were mapped to medaka genome successfully. This high-throughput 5'SAGE method provided us both transcription start site information which used to predict genes and genome-wide messenger RNA expression profile as well. I sketched the transcriptional landscape of medaka genome and found domains with highly or weakly expressed genes scattering on the chromosomes. The landscape of expression levels agrees with that of gene density very well.

Other important parts of MEDAKA GENOME SEQUENCING PROJECT with strong correlation

1. Medaka genome assembly

Medaka genome assembly was completed mainly by Kasahara, M. & Sasaki, S.

Recently, genomes of many species have been reported[3-8]. Quite a bit of them was assembled by the whole-genome shotgun (WGS) method[6] (Fig. 2), which we also adopted for the medaka genome sequencing. Two inbred strains of medaka, Hd-rR and HNI[9] were used to construct shotgun libraries. From the Hd-rR library, a total of 13.8 million reads were obtained, amounting to ~10.6-fold redundancy in sequence coverage (Table 1). Shotgun reads from short plasmid ends of HNI strain, which amounts to ~2.80-fold redundancy in sequence coverage, were also obtained and assembled for SNPs analysis

.

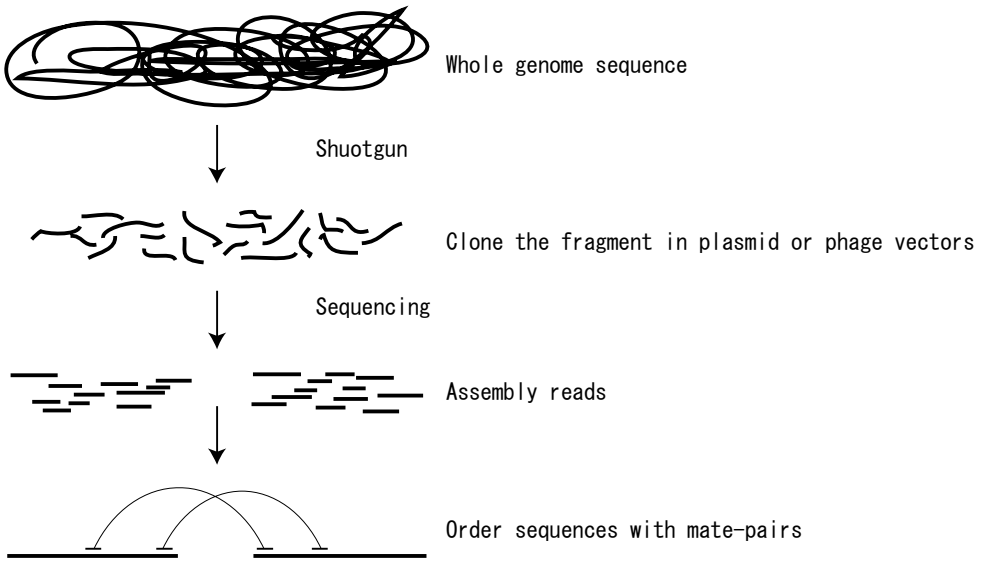


Figure 2. Diagram of whole-genome shotgun (WGS) method.

Table1. The list of shotgun library from Hd-rR strain. The genome size was assumed to be 700.4Mb.

type	size	sequence coverage	clone coverage
Plasmid	2.6kb	9.69	23.1
Plasmid	7.5kb	0.48	4.3
Fosmid	35.5,37.5kb	0.26	11.1
BAC	130-210kb	0.12	15.0
Total		10.55	53.5

Since the two strains have roughly 3% difference at DNA level[1], we were able to detect 16.4 million reliable SNPs by aligning HNI assembly against the Hd-rR assembly. At least one SNP marker was attempted in each scaffold longer than 60kb where possible. Nevertheless, some unanchored scaffolds remained. As these were often replete with repetitive sequences, the unique primers required for genetic mapping could not be designed on them. Most of the non-repetitive sequences were mapped on the chromosomes, and 89.6% of the nucleotides were anchored to the chromosomes (Table 2).

Table2. Ultracontig anchoring statistics. Oriented ultracontigs were mapped on the chromosome. Unoriented ultracontigs were associated to the genetic map (i.e. their positions on the chromosome is known), but the direction on the chromosome was not. Unordered ultracontigs arise when several ultracontigs having genetic markers of same genetic distance. Note that bases of this class are not only unordered but also unoriented.

		bases (Mb)	Percentage
Anchored	oriented	584.0	.83.37
	unoriented	14.7	2.10
	unordered	29.1	4.16
Unanchored		72.6	10.37
Total		700.4	100.00

A comparison with ten reference BACs with total length as 2.3Mb, revealed that 92~99% of the BAC sequences were covered by the assembly, as expected from the high sequence coverage (~10.4x), which is significantly higher than previous fish WGS projects. No global mis-joining was found in those BACs. Although a mis-assembly (insertion / deletion / mis-orientation / mis-ordering) was observed at every 77.3kb on average, it should be noted that not all of the BAC sequences were finished; it may include the inconsistency arising from the mis-assemblies in the reference BACs. To assess the base level accuracy of the whole genome assembly, the reference BACs and the assembly were aligned. The overall base level accuracy was 99.86%, whereas it increased to 99.95% when 100bp ends of the contigs were excluded.

2. Medaka gene prediction

Medaka gene prediction was completed mainly by Ahsan, B.

Hashimoto, S. *et al.*[10] recently showed that the 5'SAGE method could detect transcription start sites (TSS) in the human genome with 99% accuracy. Thus, we constructed a medaka transcription start site (TSS) catalogue by collecting 1,186,742 tags from mixture of 0-7 day old embryos and adult body tissues. Each 5'SAGE tag is a 19-20-nt-long 5'-end of mRNA. 1,186,742 tags were largely redundant. Of the tags, 382,341 were confirmed to be non-redundant and were mapped to the medaka draft genome. This located 344,266 unique loci, to which tags were matched fully or with one mismatch. Tags that mapped to multiple loci or had more than one mismatch were eliminated to exclude false-positive transcription start sites.

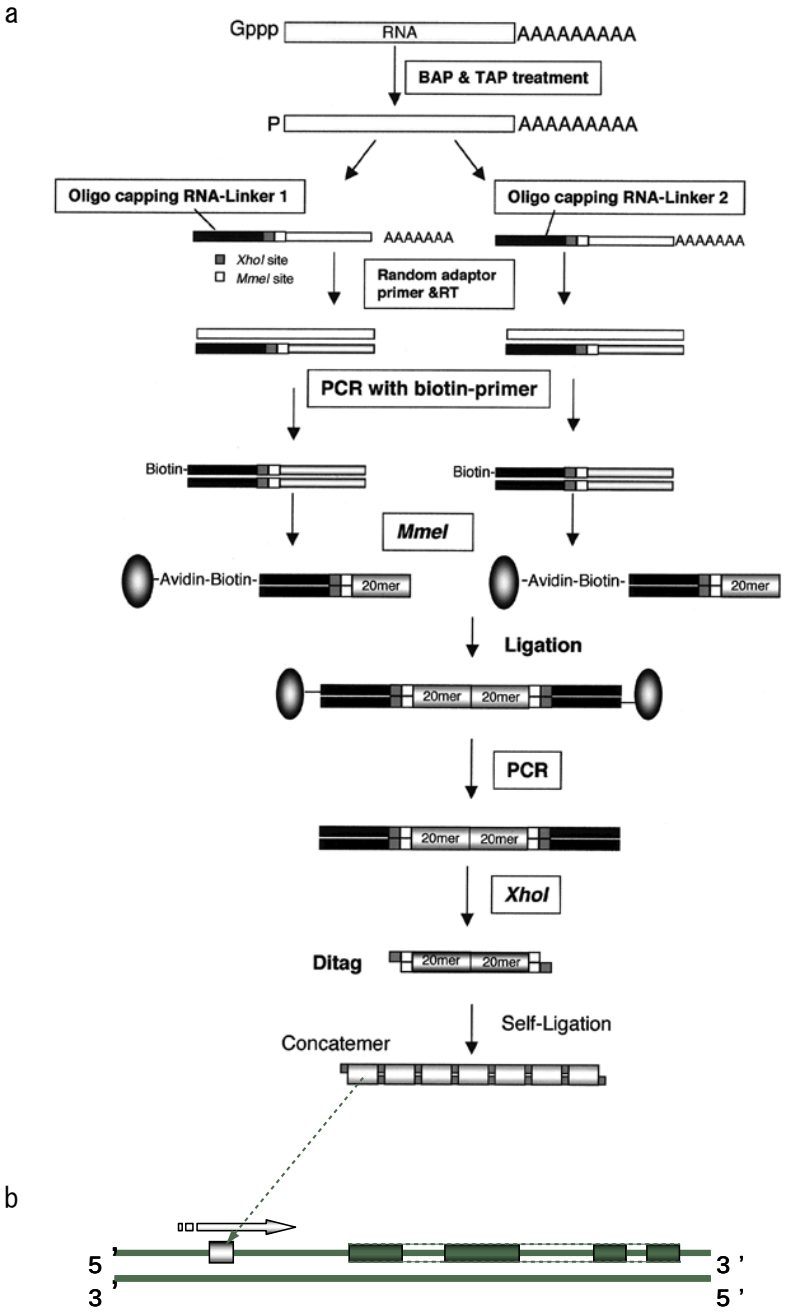


Figure 3. a, Scheme for construction of a 5'-end SAGE library. Hashimoto, S. *et al.* 5'-end SAGE for the analysis of transcriptional start sites. First, collected the 5' SAGE tags of cDNAs, then mapped them to genome, got the information of transcription start site, and at last predicted genes at the downstream of them. A PCR confirmation experiment showed that 90% of the predicted genes were actually transcribed. This method improved the start codon prediction accuracies by 20-30%. b, Predicted genes with transcription start site.

The estimated number of genes in sequenced draft genomes[4-8] continues to fluctuate. One reason for this ambiguity is that genes are neither well defined nor easily recognizable. To predict medaka genes, we used the 5'SAGE method, a novel approach to gene prediction (Fig. 3). After first determining the TSSs on the medaka draft genome, we used Genscan[11] to predict a gene structure for each one. The initial exon was complemented by a heuristic algorithm if Genscan missed it.

This evidence-based transcriptional mapping afforded us a long-awaited and accurate gene profiling method that is not biased by known sequence information. As a 5'SAGE tag predicts the precise (~99%) transcription position that is transcribed by RNA polymerase II (PolII) with cap structure, a gene search can be performed downstream of the tag. Using TSS information and the *ab initio* method, we constructed the medaka gene set of 20,141 genes with 344,266 transcripts and 158,219 exons. The genes contain 7.8 exons on average, with a total coding region of 28,485,099 bp, which is just 4.07% of the medaka draft genome of 700.4 Mb.

Because every predicted gene of the medaka carries expression-based biological evidence, we are confident of the existence of protein-coding genes in the predicted regions. However, at the exon level, our predicted genes have some inaccuracy, which is unavoidable with an *ab initio* method. To test expression and exon level accuracy, we used RT-PCR (Fig. 4) and medaka expressed sequence tags (ESTs). In the RT-PCR test, we obtained 23 expression bands from 25 predicted genes, demonstrating that our gene set is a valid gene catalogue, at least for fishes. We also calculated that our gene set covered 6950 (85%) of the 8,158 representative EST sequences stored in Unigene. The remaining 1,208 representative EST sequences and 20,141 predicted genes supported by 5'SAGE tags constitute the current set of medaka genes.

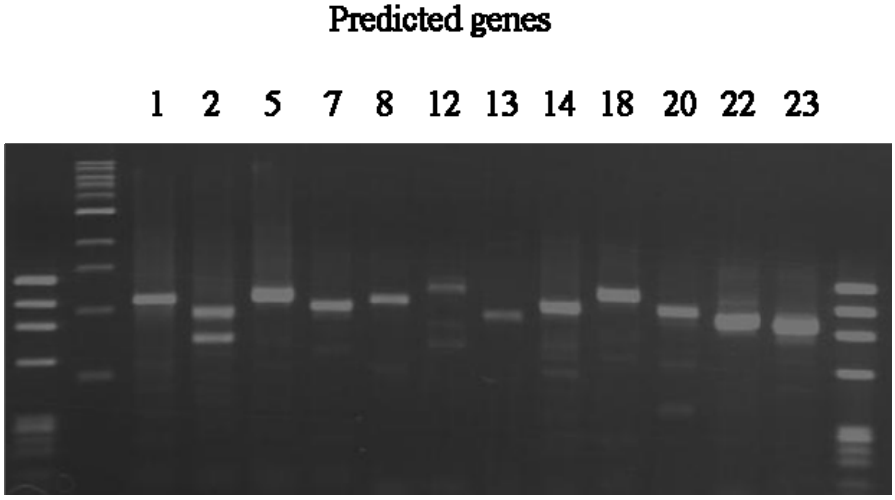


Figure 4. Electrophoretic photograph of RT-PCR.

3. Medaka browser “UTGB”

Medaka gene prediction was completed mainly by Nagayasu, Y., Kasai, Y. & Doi, K.
The medaka genome browser is available at http://medaka.utgenome.org/ (Fig. 5).

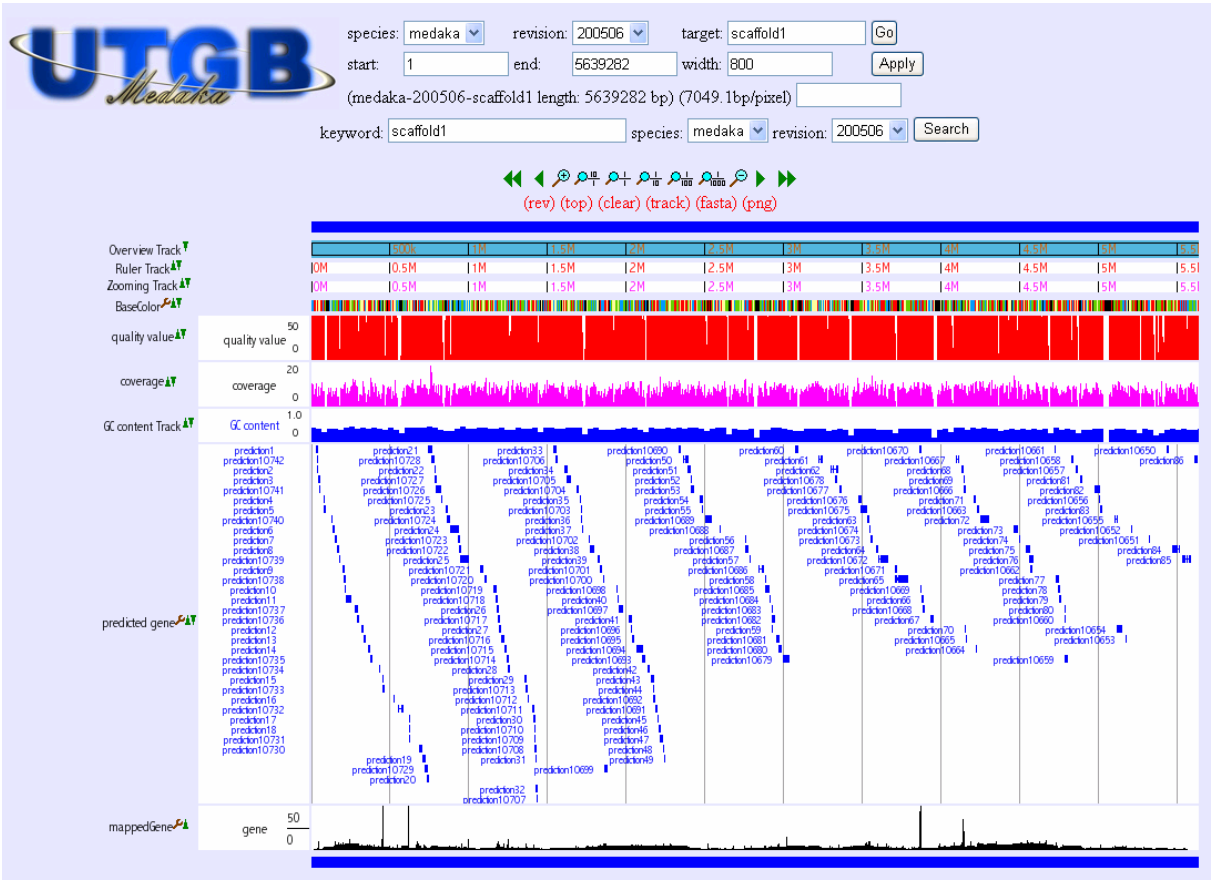


Figure 5. A snapshot of UT Genome Browser (Medaka)

Methods

Novel genes detection

I used the TBLASTX program to compare each gene in the predicted medaka gene catalogue to human Refseq genes (download at National Center for Biotechnology Information (NCBI) database, http://ncbi.nlm.nih.gov), *Tetraodon*, zebrafish, *Takifugu*, chicken, mouse genes (cDNA sequences were download from Ensembl database (http://www.ensembl.org) and I picked the longer cDNA sequence of a same gene as the representative gene.). To detect the novel medaka genes, I aligned the predicted genes to Unigene clusters of Aves (*Gallus gallus*), Amphibia (*Xenopus laevis, Xenopus tropicalis*), Actinopterygii (*Danio rerio*, *Oncorhynchus mykiss*, *Fundulus heteroclitus* ,*Gasterosteus aculeatus, Salmo salar, Takifugu rubripes*) and Ascidiacea(*Ciona intestinalis, Ciona savignyi, Molgula tectiformis*) classes as well. Then, I aligned the predicted genes to the *Takifugu* genome to further confirm *Takifugu* genes. At last, I aligned them to known medaka ESTs with BLASTN program to get the number of novel medaka genes without EST support.

Paralogue and orthologue detection

I performed an all-against-all alignment among the predicted medaka genes using TBLASTX ($E < 10^{-4}$). Because of the possibility of mistake in predicting amino acid frameworks, I used the DNA sequences of genes in replace of amino acid sequences. I then calculated the coverage of the aligned portion of both query and target genes. Since the ratio of the alignment to the total length exhibits stronger evidence of similarity than the E-value does, reciprocal alignment coverage (RAC), ratio of reciprocally aligned portions shared between two paralogues, was used as the criterion and 0.3 as the minimum cut-off value. Paralogous pairs were then defined as 1:1 reciprocal best matches with RAC>0.3. Since Ka/Ks calculations are of limited use between highly homologous sequences, they were not used in our paralogue detection stream owing to low homology ratio between paralogues. I used the same method to detect paralogues in *Tetraodon* (downloaded from the *Tetraodon* Project http://www.genoscope.cns.fr/externe/tetranew/) and *Takifugu* (downloaded from the FUGU Genome Project http://www.fugu-sg.org/) genes.

Core orthologous relationships were detected using TBLASTX ($E < 10^{-4}$) with a query's alignment coverage (AC) >0.3. 1:1 orthologues were 1:1 reciprocal best matches among the species with RAC>0.3.

Synteny detection

I used medaka-human 1:1 orthologous pairs that were identified as to their respective positions on medaka or human chromosomes. I grouped two medaka-human 1:1 orthologues into a synteny block if they were neighboring on the human genome and also located on the same medaka chromosome, and I iterated this process until no more updates were made on all

synteny blocks. I identified 572 human-medaka synteny blocks.

Conservation of Gene Ontology

In the calculation of which GO was significantly conserved or non-conserved between medaka genes and medaka pair-to-pair paralogous genes, I used Fisher's exact probability test. The number of medaka genes and medaka pair-to-pair paralogous genes with one or more biological process GO is 6,929 of 20,141 and 887 of 1,458, respectively. If the number of genes of a GO in medaka genes is x, in medaka pair-to-pair paralogous genes is y, I would use the follow equation to get z-value and find the corresponding P-value in normal probability distribution.

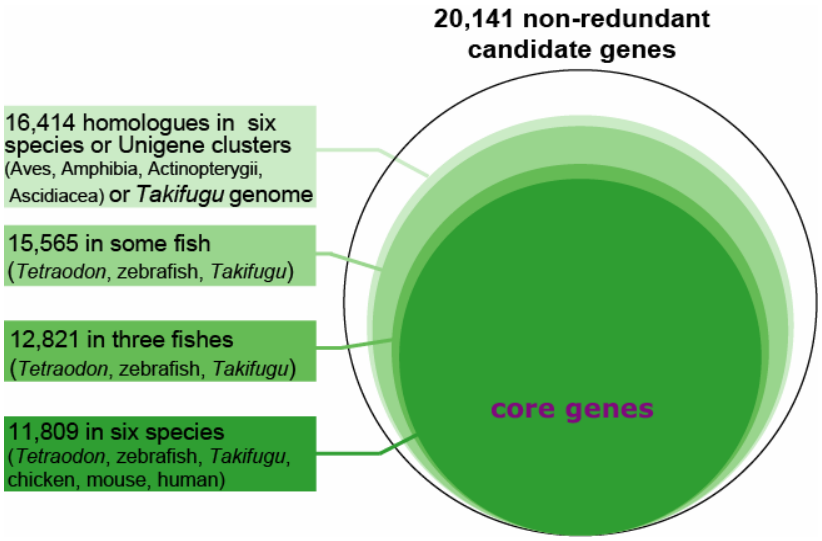
$$z = 1/\mathrm{SQRT}((x + y)/7816*(1-(y+V2)/7816)*(1/6929+1/887))*(y/887- x/6929)$$

Results

1. Novel genes

I used the TBLASTX program to compare each gene in the predicted medaka gene catalogue to human Refseq genes, *Tetraodon*, zebrafish, *Takifugu*, chicken, mouse genes and Unigene clusters of Aves, Amphibia, Actinopterygii and Ascidiacea classes (see Methods). To detect the novel medaka genes, I set a loose similarity search criterion ($E < 10^{-4}$) to ensure that a non-homologous gene was indeed novel. The predicted genes were aligned to the *Takifugu* genome to further confirm *Takifugu* genes. About 78% of the 20,141 predicted genes were conserved in another species (*Tetraodon*, zebrafish, *Takifugu*, chicken, mouse, human): 64% were common in fish (zebrafish, *Takifugu*, *Tetraodon*), and 59% (11,809 genes) were core genes found in any of the fore mentioned 6 species (Fig. 6a). Of the predicted genes, 3,727 had no homologues to fore mentioned genes of six species and Unigene clusters, and 2,078 even had no similarity to known medaka ESTs.

a



b

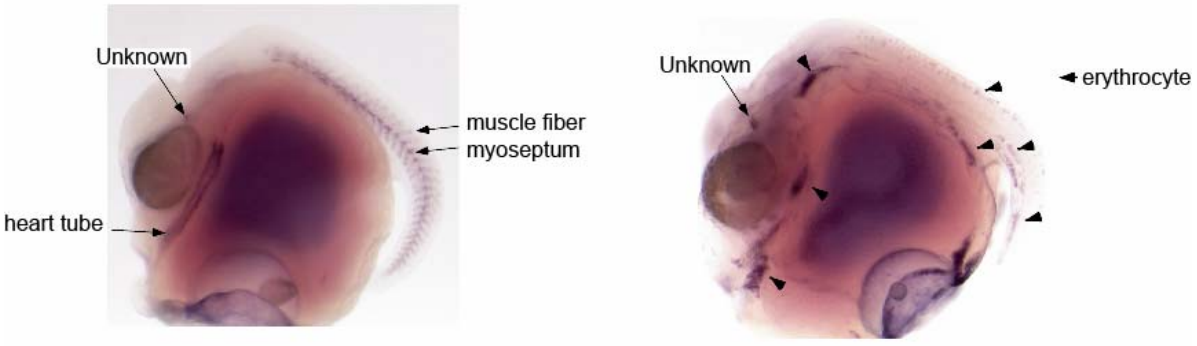


Figure 6. Novel genes in medaka. a, Breakdown of homologues of medaka genes in other species. b, Two medaka embryo clones with the expression region of a novel gene highlighted by *in situ* hybridization.

A surprisingly large number of the predicted genes were considered to be novel, with unknown functions. This result may be due to our novel gene identification method, which was based on 5' expression tags, while other gene identification methods primarily rely on the detection of homologies with known genes of other species. The discovery of a large number of novel genes indicates a medaka-specific physiology. *In situ* hybridization experiment has confirmed the expression region on the embryo of several of these genes (Fig. 6ｂ). A pioneer morpholino-based gene knockdown experiment to study the function of the gene expressed in the left clone shown in Fig. 6b, was performed by our collaborator of Takeda Laboratory. Fusion of the somite was observed in 5% (10/197) individuals (Fig. 7) and 53% (103/197) embryos were cell-aggregated or terrible deformity or dead, whereas the corresponding ratios in control were 4% (1/25) and 0% (0/25). Although it can not be judged that the function of this novel gene is somite abnormal, this gene was suggested to have certain lethality.

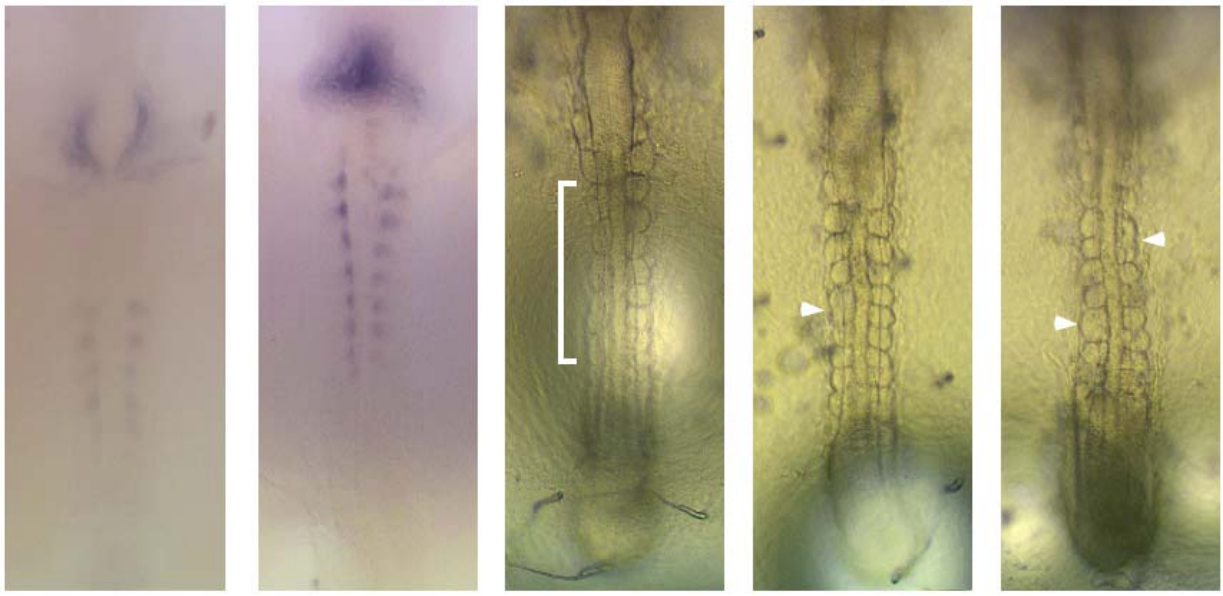


Figure 7. Fusion of the somite found in morpholino-based gene knockdown experiment. The two photos on the left show the dorsal dyeing of medaka embryos (stage22-23 and stage23-24, respectively) of *in situ* hybridization. The three photos on the right are embryos with morpholino injection. Arrows and key parenthesis show fusion of the somite.

2. Evolution of orthologues

11,617 of 20,141 predicted medaka genes had strong human orthologues. 4,342 constituted medaka－human reciprocally best 1:1 orthologue pairs (see Methods). I temporarily labeled medaka genes with the gene ontology (GO) annotations of the corresponding human orthologues. 2,292 of the 4,342 medaka-human 1:1 orthologue pairs were identified with one or more gene ontology (GO) "biological process" IDs. Orthologues involved in carbohydrate metabolism, alcohol metabolism, and catabolism were more conserved than those implicated in immune response, transcription, apoptosis, DNA repair and response to stress (Fig. 8). This result was similar to that from the comparison of 1:1 chicken-human orthologues[8]. Genes related to adaptation to the environment seem to be less conserved in their protein-coding sequences. However, a recent report indicated that the upstream sequences which represent transcriptional regulatory of these genes showed higher conservation[12]. The respective evolution of promoter region and protein-coding region is considered to be uncorrelated or have more complicated relationship in the purifying selection process.

Furthermore, 925 of 1,395 human disease genes (download from OMIM database http://www.ncbi.nlm.nih.gov/omim/, there are 1,395 Entrez genes which possess at least one refseq sequence in 1,843 "#Phenotype description, molecular basis known" OMIM entries.) have strong orthologues in medaka genes, such as A2M (alpha-2-macroglobulin) and PSEN1 (presenilin 1) implicating in Alzheimer disease, TP53(tumor protein p53) and DLEC1 (deleted in lung cancer-1) involved in carcinogenesis *etc.*, which emphasizes the importance of the medaka as a model organism in experimental medical science.

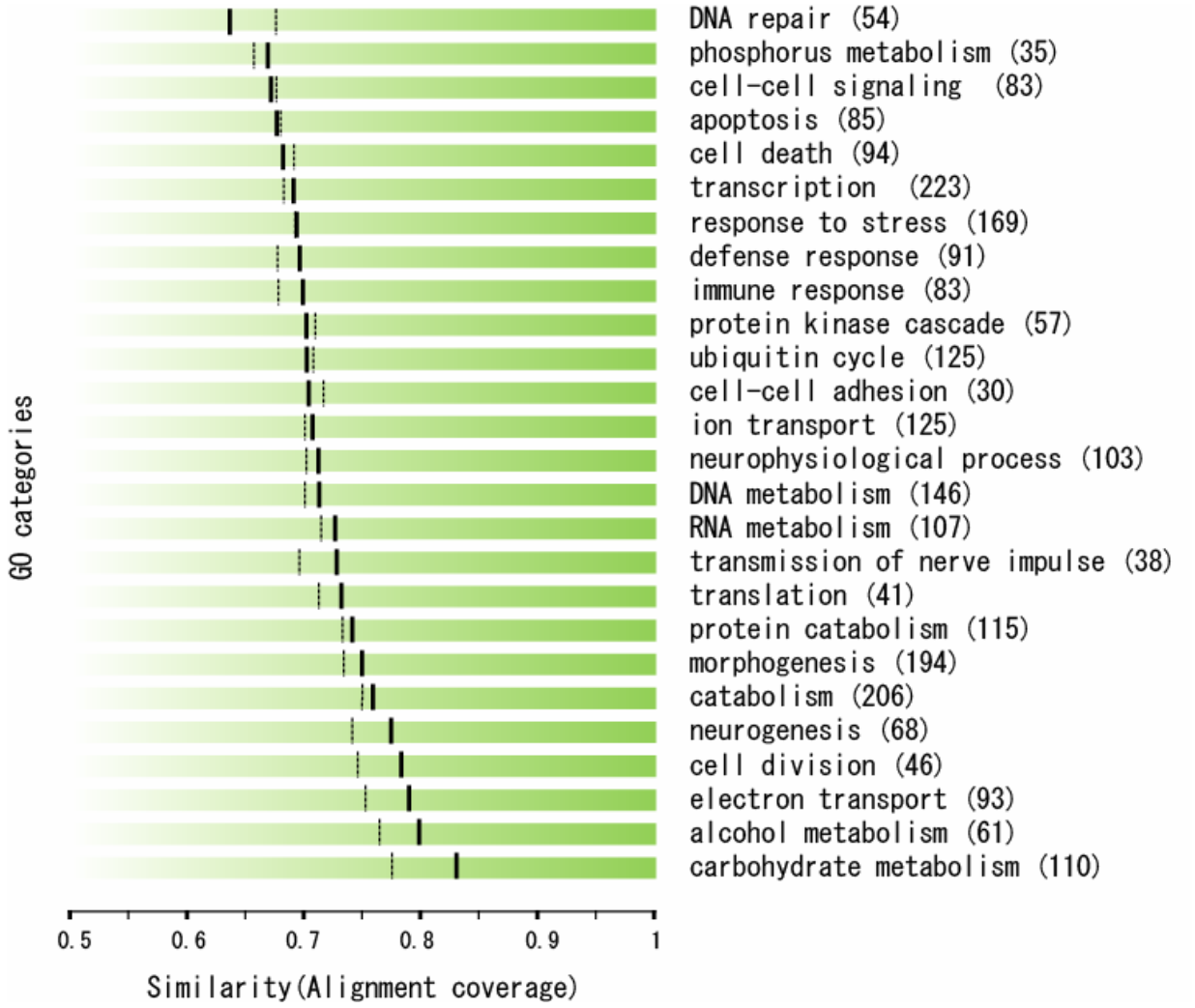
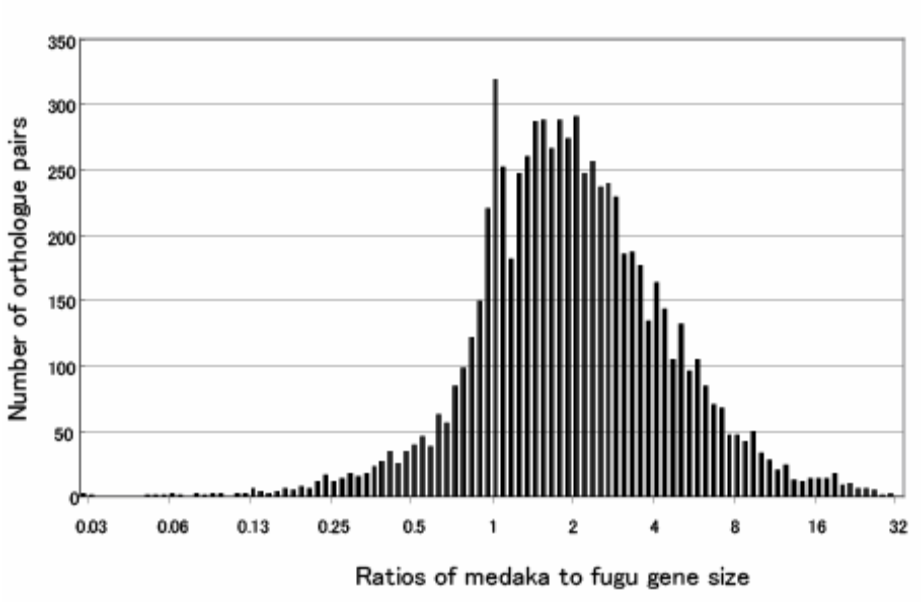


Figure 8. Similarity of medaka-human 1:1 orthologous pairs. I calculated the coverage of the aligned portion to identify their similarity (see Methods). The lateral coordinates of the bold and dashed shot bars show the median and mean similarity, respectively, of orthologues, according to GO "biological process" categories. The numbers in parentheses indicate the number of genes involved in each category.

3. Genome size difference in Fish.

Although the number of *Takifugu* genes approximates to the number of medaka genes, the *Takifugu* genome (390Mb) is about half the size of the medaka genome (700Mb). I compared 7,476 medaka-*Takifugu* 1:1 orthologues in terms of their gene sizes, which were the intron-exon region sizes of genes on the genome. Fig. 9a and Fig. 9b show the distribution of the gene size ratios and cDNA size ratios of the medaka-*Takifugu* orthologues, respectively. Two major peaks in the distribution of gene size ratios are observed; one is about 2, and another 1 is due to single-exon genes. The median cDNA size ratio of medaka to *Takifugu* is 1.12 (the median is 1.26), whereas the median gene size (exon+intron) ratio of medaka to *Takifugu* is 1.93 (the average is 2.87). This gene size difference affects the genome size increased from *Takifugu* to medaka. About 28% of the medaka genome consisted of exon and intron regions, while the remaining intergenic regions were another significant factor contributing to the genome size difference in the fishes. The question of whether special motifs or repetitive sequences have evolved in the intronic regions is of great interest.

a



b

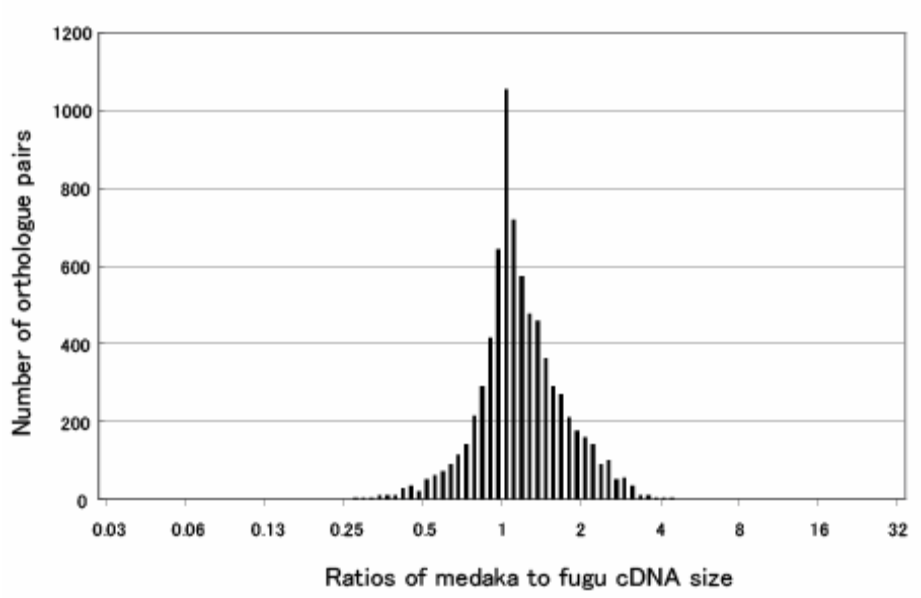


Figure 9. a, Comparison of the gene size of 7,476 medaka-Takifugu 1:1 orthologues. b, Comparison of the cDNA size of 7,476 medaka-Takifugu 1:1 orthologues.

4. Evolution of paralogues

A great number of paralogous pairs are found in fish. In medaka, 1,730 pairs of duplicated genes were identified; while *Tetraodon* had 2,212 (see Methods). Paralogues are also known as WGD fossils. They are thought to participate primarily in genetic robustness and in a complementary "subfunctionalization" function[13]. I investigated the evolutionary relationships of medaka genes to *Tetraodon* genes and how many paralogous pairs in the fishes were conserved (Fig. 10).

In the evolutionary relationships of the total genes in the two fishes, 4,817 genes are 1:1 orthologues across the two species (Fig. 10a). 65.6% (13,222/20,141) of medaka predicted genes have common strong homology (1:1 or n:n, see Methods) with *Tetraodon*. 5,094 medaka genes have no weak homology with *Tetraodon* genes. A more interesting result is the evolutionary relationships of paralogous pairs in this two fishes (Fig. 10b). The length of a colored bar represents the count of paralogous pairs with at least one of the counterparts in a pair conserved in the relationships. White colored numbers displayed in colored bars present numbers of paralogous pairs which are also conserved in two fishes as paralogous pairs. So, 100/454, 729/929, 1,028/1,259 of medaka paralogous pairs are conserved in *Tetraodon* as paralogous pairs in the relationship 1:1, n:n and weak homology, respectively. The corresponding numbers in *Tetraodon* are 100/465, 673/994, 1,185/1,469.

Intuitively, the ratios of “pair-to-pair” paralogous pairs seem quite large, and Fig. 10c shows the breakdown of 1,730 medaka paralogous pairs in the n:n evolutionary relationships to *Tetraodon* genes. The detailed numbers in cases 1-6 are 729, 41, 159, 625, 60 and 116, respectively. 42% of medaka paralogous pairs are pair-to-pair type (case 1 in Fig. 10c). Over half have orthologues in *Tetraodon* paralogous genes.

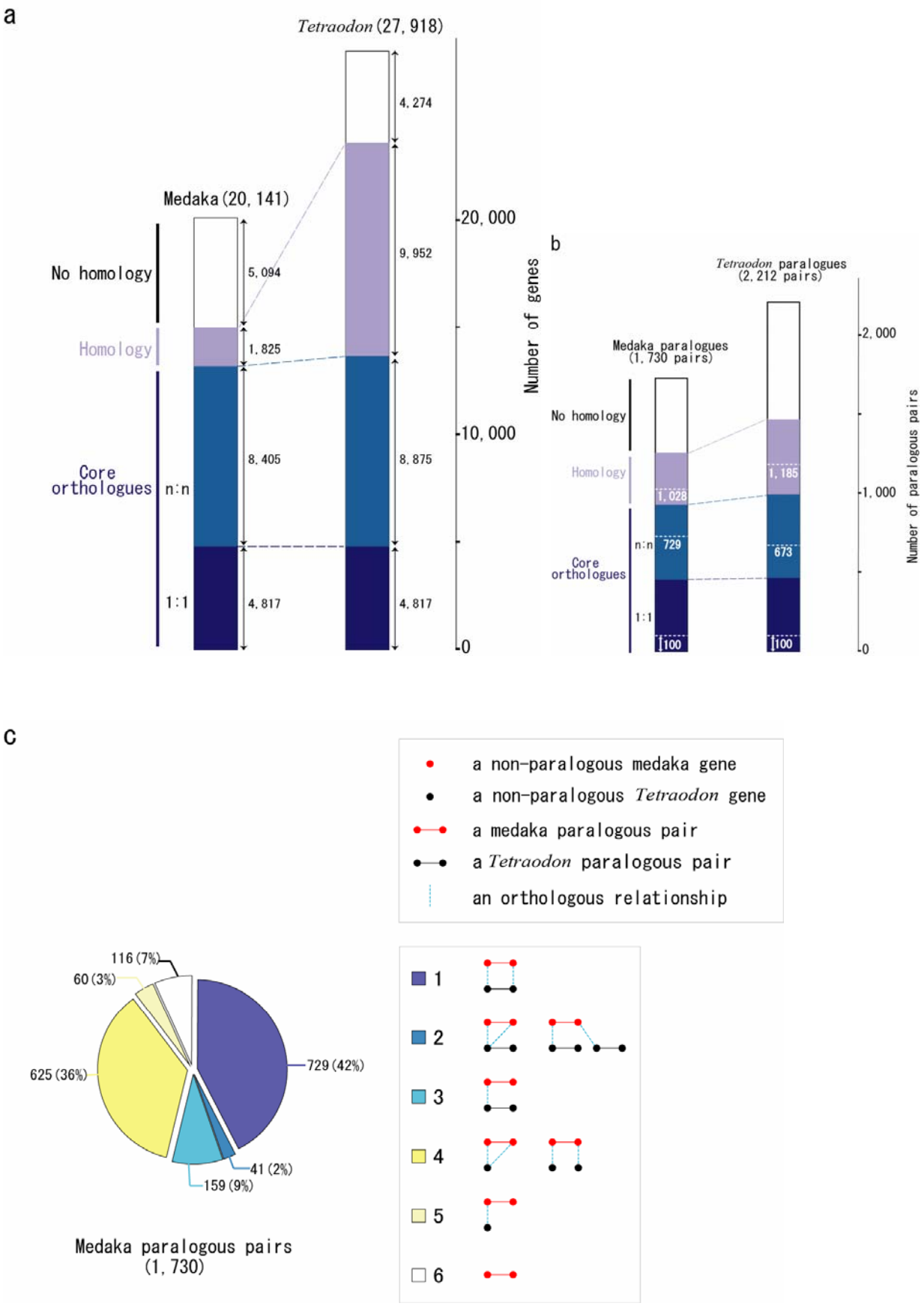


Figure 10. Comparison of gene sets in medaka and *Tetraodon*. The length of a colored bar represents numbers of genes or paralogous pairs in the relationship. a, Evolutionary relationships of the

total genes in the two fishes. b, Evolutionary relationships of paralogous pairs in two fishes. Write colored numbers displayed in colored bars present numbers of paralogous pairs which are also conserved in the other fish as paralogous pairs in the relationship 1:1, n:n, or weak homology, which is presented by the color of the bar. c, A breakdown of 1,730 medaka paralogous pairs in the evolutionary relationships to *Tetraodon* genes. There are six cases: 1, a medaka paralogous pair is also conserved as a paralogous pair in *Tetraodon*; 2, a medaka paralogous pair is conserved in paralogous pairs of *Tetraodon* but not as a whole paralogous pair; 3, only one gene in a medaka paralogous pair is conserved in paralogous genes of *Tetraodon*; 4, both genes of a medaka paralogous pair are non-paralogous genes of *Tetraodon*; 5, only one gene in a medaka paralogous pair is conserved in non-paralogous genes of *Tetraodon*; 6 neither of the two genes in a medaka paralogous pair is conserved in genes of *Tetraodon*.

At last, I calculated Gene Ontology clusters which were especially conserved or disappeared in medaka pair-to-pair paralogues. Comparing 729 medaka pair-to-pair paralogous pairs with 20,141 medaka genes, pair-to-pair paralogues involved in transport, catabolism, alcohol metabolism, transcription factor NF-*kappa*B, carbohydrate metabolism *etc*. are significantly conserved (Table 3). And pair-to-pair paralogues involved in RNA metabolism, DNA metabolism, transcription, response to DNA damage stimulus, phosphorus metabolism *etc*. are significantly non-conserved (Table 4). This result is extremely associated with the fore mentioned analysis of evolution of medaka-human orthologues. Orthologues involved in carbohydrate metabolism, alcohol metabolism, and catabolism were more conserved than those implicated in immune response, transcription, apoptosis, DNA repair and response to stress (Fig. 8). So the interesting finding is that genes related to adaptation to the environment seem to be less conserved in pair-to-pair paralogues as well. A similar analysis was performed between 1,001 medaka not-pair-to-pair paralogous pairs and 20,141 medaka genes. To be quite a contrast to the comparison mentioned above, not-pair-to-pair paralogues involved in cell adhesion, apoptosis, reproduction *etc*. are significantly conserved (Table 5).

Table 3. Contents of significantly conserved GO between medaka genes and medaka pair-to-pair paralogous genes.

GO	Medaka genes (tatal:20,141)	Medaka pair-in-pair genes (total:1458)	P-value
organic acid transport	67	20	6.0E-04
catabolism	542	97	1.0E-03
localization	1474	229	2.0E-03
transport	1467	228	2.0E-03
amine transport	55	15	7.0E-03
amino acid biosynthesis	36	11	9.0E-03
regulation of body fluids	83	20	9.0E-03
macromolecule catabolism	384	68	1.0E-02
regulation of biosynthesis	37	11	1.0E-02
positive regulation of signal transduction	64	16	1.0E-02
regulation of I-kappaB kinase/NF-kappaB cascade	64	16	1.0E-02
alcohol metabolism	167	33	2.0E-02
hemostasis	73	17	2.0E-02
steroid metabolism	88	19	4.0E-02
protein catabolism	308	53	4.0E-02
carbohydrate metabolism	289	50	4.0E-02

Table 4. Contents of significantly non-conserved GO between medaka genes and medaka pair-to-pair paralogous genes.

GO	Medaka genes (tatal:20,141)	Medaka pair-in-pair genes (total:1458)	P-value
RNA metabolism	225	13	3.0E-03
DNA metabolism	364	27	4.0E-03
transcription	763	72	8.0E-03
response to endogenous stimulus	148	9	3.0E-02
response to DNA damage stimulus	137	8	3.0E-02
phosphorus metabolism	80	3	3.0E-02
DNA repair	127	8	5.0E-02

Table 5. Contents of significantly conserved GO between medaka genes and medaka paralogous genes without pair-to-pair paralogous genes.

GO	Medaka genes (tatal:20,141)	Medaka paralogous genes without pair-in-pair genes (total:2,002)	P-value
cell adhesion	465	94	1.0E-04
reproduction	99	27	6.0E-04
cell death	313	62	3.0E-03
glycoprotein metabolism	49	15	4.0E-03
apoptosis	287	57	4.0E-03
lipid catabolism	55	16	5.0E-03
male gamete generation	52	15	6.0E-03
morphogenesis	803	134	9.0E-03
Golgi vesicle transport	51	13	3.0E-02
organogenesis	626	103	3.0E-02
regulation of translation	27	8	4.0E-02
protein kinase cascade	177	34	4.0E-02

5. Similarity of expression pattern

Genes with identical functions usually diversify easily or die under relaxed selection, and hence the majority of paralogues are thought to have complementary expression patterns with each other[13]. However, duplicated genes tend to share similar cis-regulatory motifs, a significant co-expression relationship is expected in paralogues, as compared to that between two random genes, motivating us to analyze the expression divergence between medaka paralogues. The result suggests a positive correlation of expression levels as Fig. 11 illustrates the distribution of expression levels between paralogues. I observed 61 pairs of abundant paralogues that were associated with more than 38 tags (the mean value of 20,141 medaka genes), whereas the expected number of abundant pairs, according to the random model, was only 12 (Table 6). However, more or less expression divergences have happened in quite a large portion of medaka duplicated genes, which enables tissue or developmental specialization to evolve.

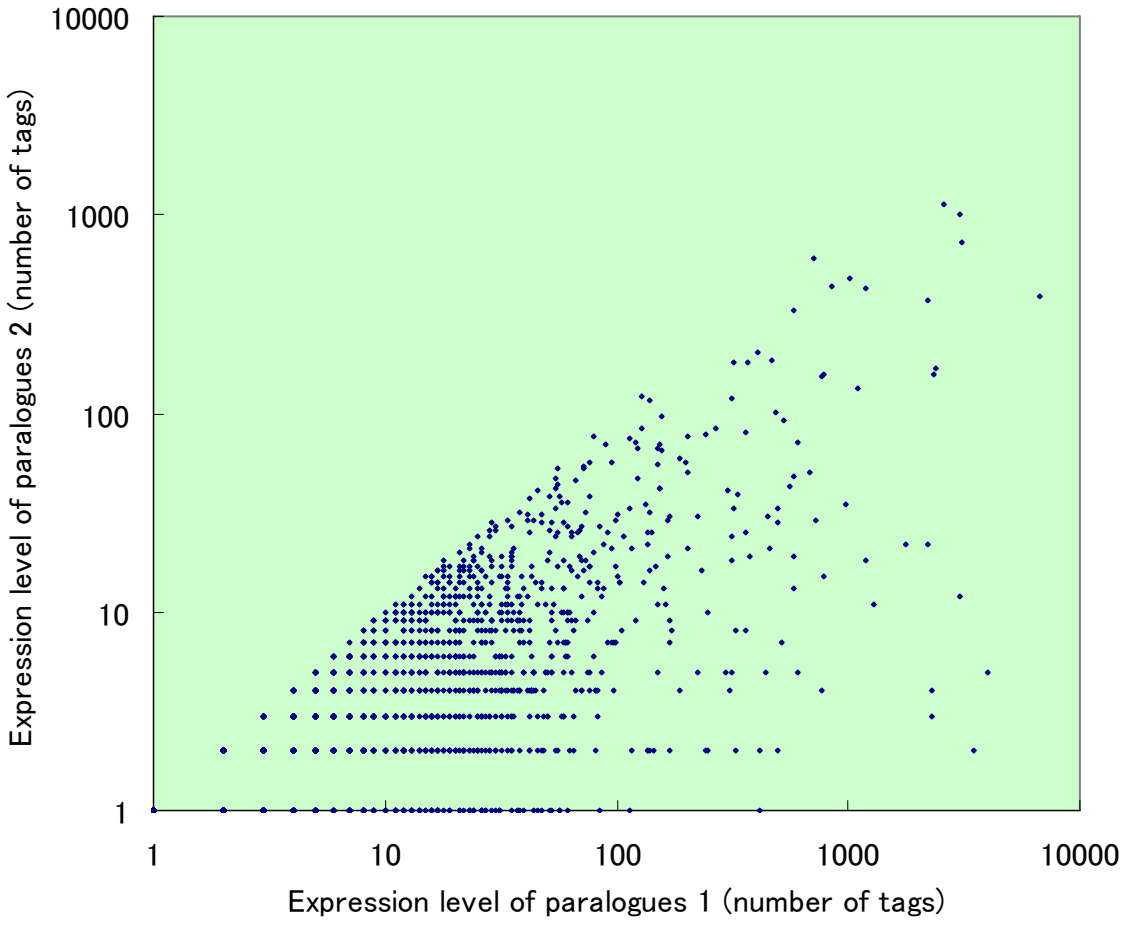


Figure 11. Expression levels between 1,730 medaka paralogues.

Table 6. Breakdown of medaka gene expression (number of tags). The number of expected paralogous pairs was calculated by assuming a random model. When the number of tags of both genes in a paralogous pair are larger than the median value of 20,141 medaka genes, the expectation value is 368*(9368/20141)*(3460/20141)/2=374 pairs; When the number of tags of both genes in a paralogous pair are larger than the mean value of 20,141 medaka genes, the expectation value is 1648*(1648/20141)*(3460/20141)/2=12 pairs.

	Total genes/20141	Paralogous genes/3,460
median	4	6
mean	38.07	36.36
tags > median (4)	9,368	2,026
tags > mean (38)	1,648	364

Expression level	Expectation pairs	Observation pairs
P1&P2 > median (4)	374	690
P1&P2 > mean (38)	12	61

6. Local gene duplications

In mammalian genomes, local gene duplication is undoubtedly an important element in the evolution of genetic diversity. I examined local gene duplication clusters in medaka to measure their contribution to fish-specific biology. Each gene is compared with the preceding and subsequent ten genes on the chromosome, and constant homologous genes in the sliding window are treated as locally duplicated gene clusters. I found 710 local gene clusters in medaka, containing 1,638 genes. These numbers are considerably lower than those in the mouse and rat genomes because in the mouse (rat, respectively) 910 (784) local gene clusters having 3,784 (3,089) genes were identified according to a similar method and measurement[6,7]. I did not observe large clusters of immunoglobulin or olfactory receptors unlike those in the mouse and rat genomes, but keratin genes, which code for fish scales, accounted for a large proportion (Table 7).

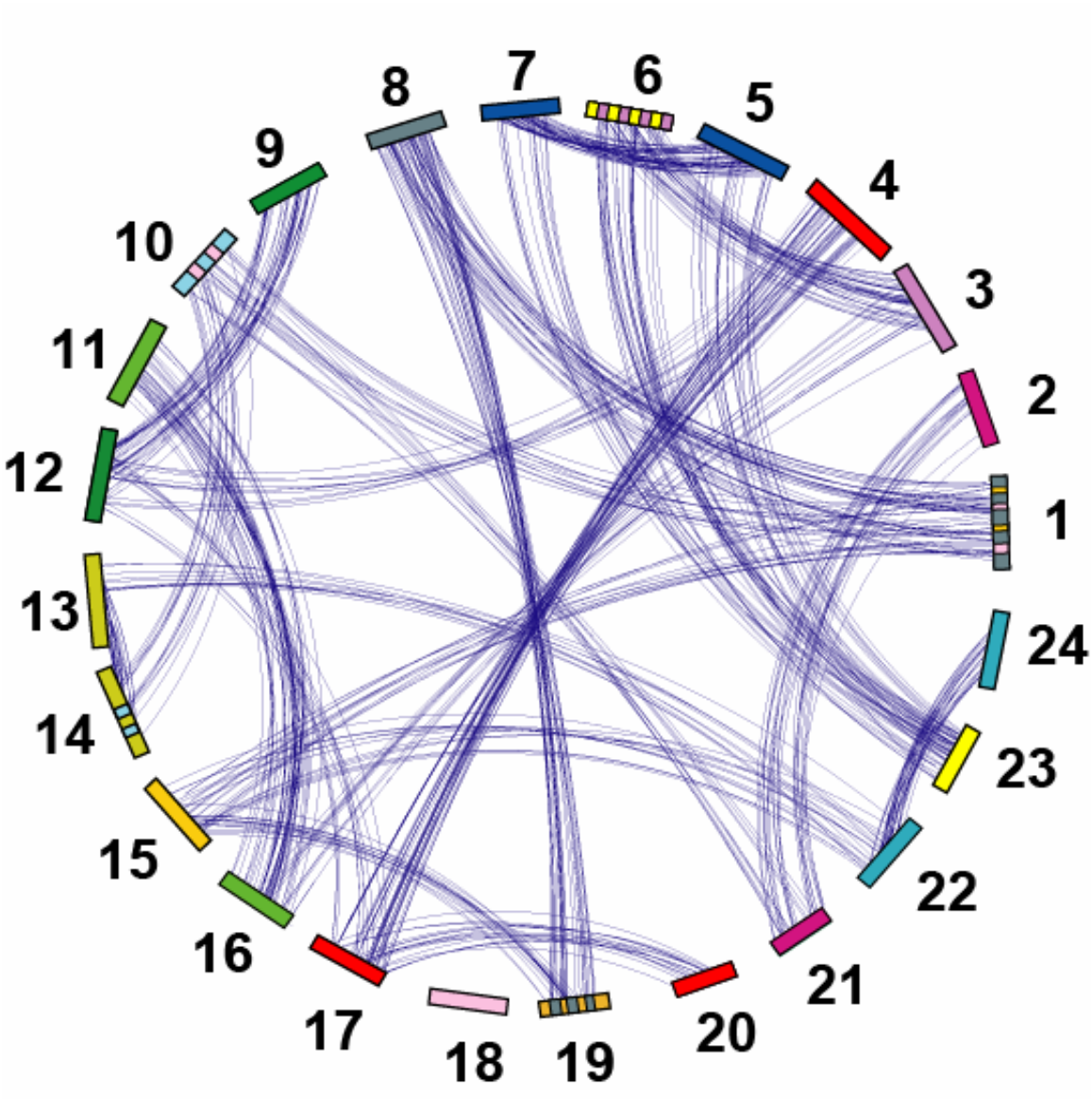
Table 7. Local gene duplication in the medaka draft genome.

Numbers of gene involved	Chromosome	Annotation
15	17	zinc finger protein
10	16	zinc finger protein
9	14	claudin
9	11	homeo box
7	7	homeo box
7	16	unknown
7	2	unknown
6	8	hemoglobin
6	17	tripartite motif-containing
6	11	unknown
6	8	homeo box
5	8	keratin
5	6	keratin
5	7	keratin
5	1	ubiquitin specific protease
5	18	protocadherin
5	3	meprin
5	23	aldose reductase

7. Genome duplication

Fish genomes underwent whole genome duplication after the divergence from the mammalian linkage. In order to confirm genome duplication, I made pairwise comparison among predicted genes, and identified 1,730 pairs of duplicated genes. Figure 12a shows a global distribution of paralogues on medaka chromosomes. Most of the chromosomes have strong correspondences with one or two other chromosomes. 95% (4,117 of 4,342) of medaka-human 1:1 orthologues could be identified to their respective positions on medaka or human chromosomes. A total of 572 human-medaka synteny blocks (see Methods) were identified (Fig. 12b). The global distribution of paralogues in the medaka genome and the existence of synteny blocks with an interleaving pattern covering a large part of the human genome are powerful evidence of WGD[3].

a



b

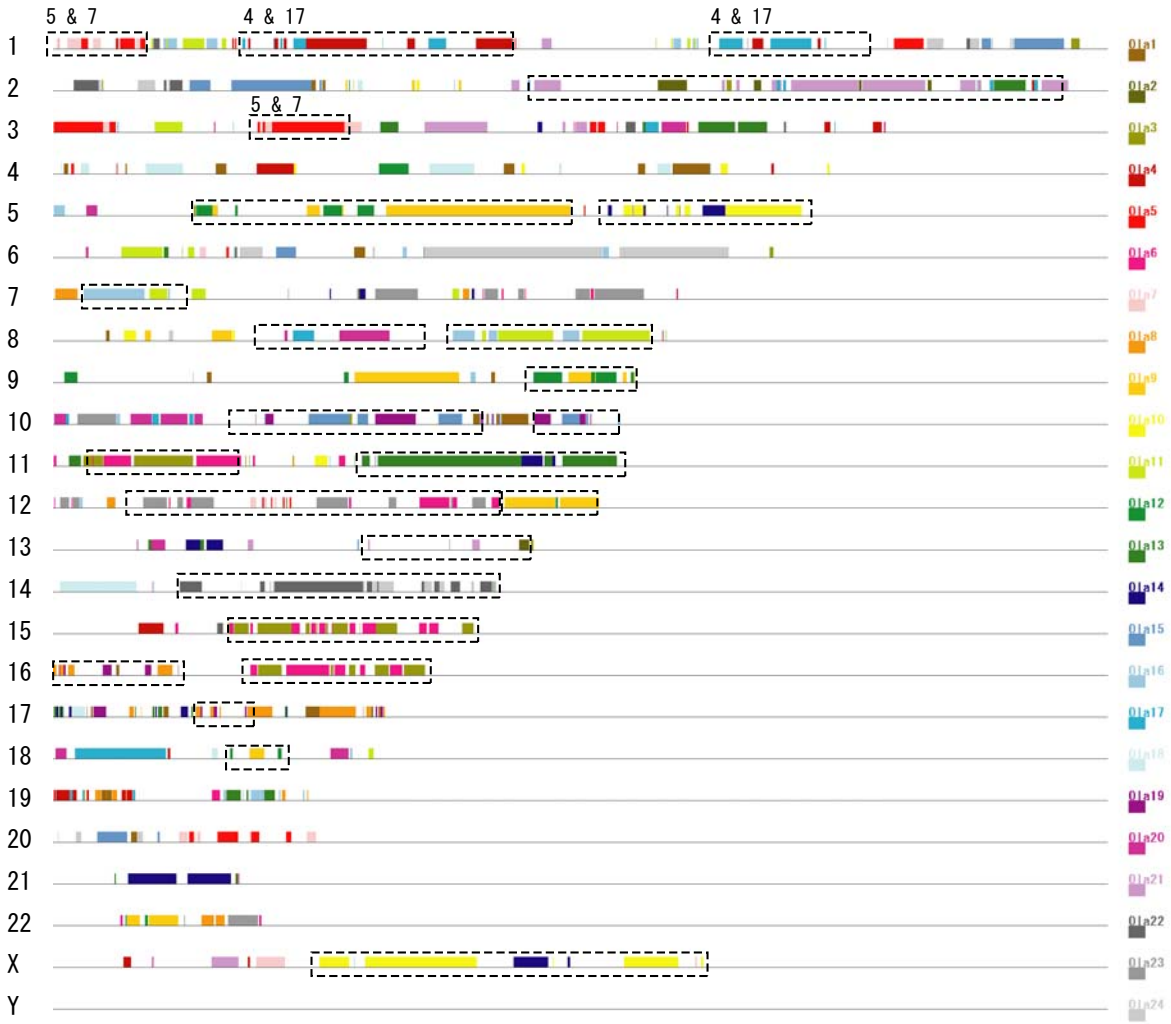


Figure 12. Genome duplication. a, Global distribution of duplicated genes in the medaka genome. Paralogue pairs are 1:1 reciprocal best matches with an aligned portion greater than 30%. b, Blocks of ancestral chromosomes on human chromosomes. Synteny blocks along human chromosomes are mapped to two or three medaka chromosomes in an interleaving pattern. The black boxes configured by dashed lines represent examples of blocks of ancestral chromosomes. For example, ancestral blocks mapped to medaka Ola5 and Ola7 always occur together. On the other hand, medaka Ola5 and Ola7 have a strong correspondence with paralogues.

8. Transcriptome map

Over one million-plus 5'SAGE tags provided both TSS information, which is used to predict genes, and a genome-wide mRNA expression profile. I used 18,484 predicted medaka genes that were mapped to medaka chromosomes using 711,385 SAGE tags. The whole chromosome viewpoint revealed a higher-order structure of the genome (Fig. 13), which was similar to the human transcriptome map[14] in that domains with highly or weakly expressed genes are scattered on the chromosomes. The expression level landscape agreed quite well with that of gene density.

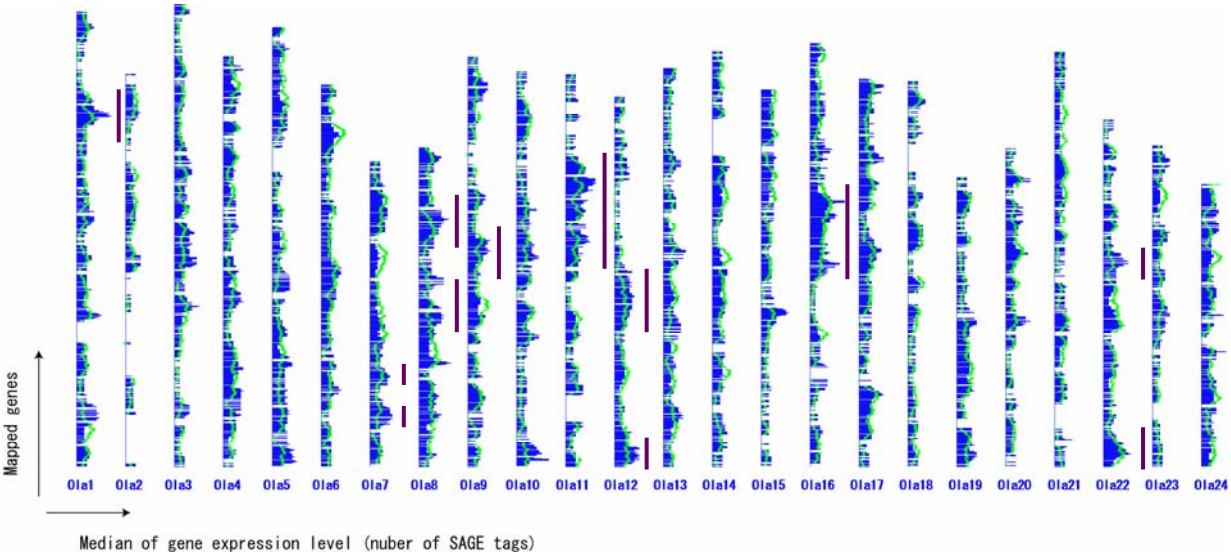


Figure 13. Medaka transcriptome map. The blue section to the right represents a moving median of expression tags occurring in each slide window of 39 genes among the 18,484 predicted medaka genes that are mapped to chromosomes. Since approximately 90% of ultra-contigs are located on the medaka genome, 18,484 of a total of 20,141 genes are mapped to the genome. The green curved line represents the gene density with a slide window size of 1Mb. The purple bar indicates regions with highly expressed genes.

References

1. Wittbrodt, J., Shima, A. & Schartl, M. Medaka--a model organism from the far East. *Nature Rev. Genet.* **3**, 53-64 (2002).
2. Nelson, J. S. *Fishes of the World.* (John Wiley & Sons, New York, 1994).
3. Jaillon, O. *et al.* Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* **431**, 946-957 (2004).
4. Myers, E. W. *et al.* A whole-genome assembly of Drosophila. *Science* **287**, 2196-2204 (2000).
5. International Human Genome Sequencing Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69-87 (2005).
6. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-562 (2002).
7. Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493-521 (2004).
8. International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695-716 (2004).
9. Naruse, K., Hori, H., Shimizu, N., Kohara, Y. & Takeda, H. Medaka genomics: a bridge between mutant phenotype and gene function. *Mech. Dev.* **121**, 619-628 (2004).
10. Hashimoto, S. *et al.* 5'-end SAGE for the analysis of transcriptional start sites. *Nature Biotechnol.* **22**, 1146-1149 (2004).
11. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78-94 (1997).
12. Lee, S., Kohane, I. & Kasif, S. Genes involved in complex adaptive processes tend to have highly conserved upstream regions. *BMC Genomics* 6:168 (2005).
13. Force, A. *et al.* Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531-1545 (1999).
14 Versteeg, R. *et al.* The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res.* **13**, 1998-2004 (2003).

Acknowledgements

First of all, I would like to express my gratitude to my supervisor, Pro. Shinichi Morishita, for his excellent guidance, a lot of encouragement and continuous support to my study. Also, I would like to thank all the members involved in the MEDAKA GENOME SEQUENCING PROJECT for giving me many insightful comments. This project has been conducted by the collaboration with these five laboratories. They are Morishita Lab. (U. Tokyo) for genome assembler, genome browser "UTGE" and bioinformatics analysis; Kohara Lab. (NIG) for plasmid and fosmid sequencing; Takeda Lab. (U. Tokyo) for material preparation and mapping; Fujiyama Lab. (NII) for BAC sequencing; Ph. D Shinichi Hashimoto (U. Tokyo) for 5'SAGE tag collection. My research is totally due to this successful teamwork.

Special thanks to these members of the bioinformatics group for their significantly associated works and valuable advices. They are Mr. Masahiro Kasahara and Shin Sasaki who assembled WGS fragments into a medaka draft genome using their in-house WGS assembler Ramen; Mr. Ahsan Budrul, who made gene predictions using TSS information of 5'SAGE and identified non-coding genes including novel miRNA candidates; Mr. Yoichiro Nakatani, who performed the computational analysis to estimate the teleost genome evolution; Mr. Tomoyuki Yamada, who developed a de novo repeat finding tool to elucidate novel repetitive elements in the medaka genome; Mr. Yukinobu Nagayasu, Mr. Koichiro Doi and Mr. Yasuhiro Kasai, who developed the medaka genome browser; Again Pro. Shinichi Morishita, who supervised the bioinformatics group.

平成十七年度修士論文

メダカ遺伝子の比較解析

曲　薇