LETTERS

The medaka draft genome and insights into vertebrate genome evolution

Masahiro Kasahara[1]*, Kiyoshi Naruse[2]*, Shin Sasaki[1]*, Yoichiro Nakatani[1]*, Wei Qu[1], Budrul Ahsan[1], Tomoyuki Yamada[1], Yukinobu Nagayasu[1], Koichiro Doi[1], Yasuhiro Kasai[1], Tomoko Jindo[2], Daisuke Kobayashi[2], Atsuko Shimada[2], Atsushi Toyoda[3], Yoko Kuroki[3], Asao Fujiyama[3,4], Takashi Sasaki[5], Atsushi Shimizu[5], Shuichi Asakawa[5], Nobuyoshi Shimizu[5], Shin-ichi Hashimoto[6], Jun Yang[6], Yongjun Lee[6], Kouji Matsushima[6], Sumio Sugano[7], Mitsuru Sakaizumi[8], Takanori Narita[2,9], Kazuko Ohishi[9], Shinobu Haga[9], Fumiko Ohta[9], Hisayo Nomoto[9], Keiko Nogata[9], Tomomi Morishita[9], Tomoko Endo[9], Tadasu Shin-I[9], Hiroyuki Takeda[2], Shinichi Morishita[1] & Yuji Kohara[9]

Teleosts comprise more than half of all vertebrate species and have adapted to a variety of marine and freshwater habitats[1]. Their genome evolution and diversification are important subjects for the understanding of vertebrate evolution. Although draft genome sequences of two pufferfishes have been published[2,3], analysis of more fish genomes is desirable. Here we report a high-quality draft genome sequence of a small egg-laying freshwater teleost, medaka (*Oryzias latipes*). Medaka is native to East Asia and an excellent model system for a wide range of biology, including ecotoxicology, carcinogenesis, sex determination[4–6] and developmental genetics[7]. In the assembled medaka genome (700 megabases), which is less than half of the zebrafish genome, we predicted 20,141 genes, including ~2,900 new genes, using 5′-end serial analysis of gene expression tag information. We found single nucleotide polymorphisms (SNPs) at an average rate of 3.42% between the two inbred strains derived from two regional populations; this is the highest SNP rate seen in any vertebrate species. Analyses based on the dense SNP information show a strict genetic separation of 4 million years (Myr) between the two populations, and suggest that differential selective pressures acted on specific gene categories. Four-way comparisons with the human, pufferfish (*Tetraodon*), zebrafish and medaka genomes revealed that eight major interchromosomal rearrangements took place in a remarkably short period of ~50 Myr after the whole-genome duplication event in the teleost ancestor and afterwards, intriguingly, the medaka genome preserved its ancestral karyotype for more than 300 Myr.

We applied the whole-genome shotgun approach to an inbred strain, Hd-rR (ref. 8), derived from the southern Japanese population, as the main target. A total of 13.8 million reads amounting to approximately 10.6-fold genome coverage were obtained from the shotgun plasmid, fosmid and bacterial artificial chromosome (BAC) libraries. A newly developed RAMEN assembler was used to process the shotgun reads to generate contigs and scaffolds. The N50 values (50% of nucleotides in an assembly are in scaffolds—or contigs—longer than or equal to the N50 value) are ~1.41 megabases (Mb) for scaffolds and ~9.8 kilobases (Kb) for contigs. The total length of the contigs reached 700.4 Mb, which, from now on, we refer to as the medaka genome size.

To construct ultracontigs, the scaffolds were integrated with the medaka genetic map by using SNP markers. For this purpose, we further obtained about 2.8-fold coverage of shotgun reads from another inbred strain HNI (refs 9, 10), which is derived from the northern Japanese population. The reads were assembled by RAMEN to scaffolds covering 648 Mb. Aligning the HNI contigs with the Hd-rR genome using BLASTZ[11], we identified 16.4 million SNPs as well as 1.40 million insertions and 1.45 million deletions in non-repetitive regions (Supplementary Table 2). We selected 2,401 SNPs and genetically mapped them onto medaka chromosomes using a backcross panel between the two strains. Where possible, at least one SNP marker was selected in each Hd-rR scaffold of greater than 60 Kb. As a result, the N50 ultracontig size became ~5.1 Mb (excluding gaps), and 89.7% of the assembled nucleotides were anchored to the chromosomes. Aligning the Hd-rR assembly with reference BACs totalling 2.3 Mb showed that the overall nucleotide level accuracy was 99.96% when 100 base pairs of contig ends were excluded (Supplementary Table 4). Details of genome assembly and analysis of basic features such as CpG islands and repeat elements are described in Supplementary Information.

We first focus on polymorphisms between the two inbred strains, Hd-rR and HNI. The genome-wide SNP rate is 3.42%, which is, to our knowledge, the highest SNP rate seen in any vertebrate species. The SNP rate is not constant among chromosomes (Kruskal–Wallis test, $P < 10^{-5}$) (Fig. 1a), like the divergence between chimpanzee and human[12]. This variation can not be accounted for by the difference in gene density alone, because the SNP rates in exons and introns are correlated in most regions across chromosomes (Fig. 1a). The substitution rate in CpG dinucleotides and the frequency of repetitive sequences might cause the variation because these factors loosely correlate with local nucleotide divergence rates ($R^2 = 0.378$ and 0.455, respectively, as illustrated in Supplementary Fig. 5b and 5c).

In human–chimpanzee analysis, the sex chromosomes exhibit the highest (Y) and lowest (X) mutation rates[12]. Medaka also has an XX–XY sex-determining system[4], but the differentiation of its sex chromosomes seems primitive; chromosome 1 (Chr 1; 33.7 Mb) serves as the X chromosome, whereas a duplicate Chr 1 with a 250-kb Y-specific

[1]Department of Computational Biology, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa 277-0882, Japan. [2]Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Tokyo 113-0033, Japan. [3]RIKEN Genomic Sciences Center, Yokohama 230-0045, Japan. [4]National Institute of Informatics, Tokyo 101-8430, Japan. [5]Department of Molecular Biology, Keio University School of Medicine, Tokyo 160-8582, Japan. [6]Department of Molecular Preventive Medicine, School of Medicine, The University of Tokyo, Tokyo 113-0033, Japan. [7]Department of Medical Genome Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo 108-8639, Japan. [8]Department of Environmental Science, Faculty of Science, Niigata University, Niigata 950-2181, Japan. [9]Center for Genetic Resource Information, National Institute of Genetics, Mishima 411-8540, Japan.
*These authors contributed equally to this work.

region that contains the male-determining gene, *DMY* (also known as *dmrt1b*)[6], serves as the Y chromosome. The Y-specific region is thought to have jumped to Chr 1 about 10 Myr ago, before the separation of the medaka lineage[13]. Thus, although we sequenced male genomes, it is difficult to distinguish whether the sequence reads of Chr 1 are derived from the X- or Y-chromosome. Indeed, the overall divergence rate in the medaka sex chromosome (Chr 1) does not differ much from autosomes (Fig. 1a). In the Hd-rR draft genome, the assembly of the Y-specific region is incomplete presumably owing to repetitive elements; instead, we detected polymorphisms in the reads from the region spanning at least 3.5 Mb around the Y-specific region, whereas the other region is highly homozygous. These polymorphisms in the inbred strain demonstrate the local suppression of crossing-over near the male-determining region, which has long been known in medaka genetic recombination tests[14]. It seems that the male-determining region causes restriction of recombination but its effect is limited to 10% of the length of Chr 1. The medaka Y chromosome might be at an early stage of sex differentiation.

The two medaka inbred strains Hd-rR and HNI are estimated to have diverged about 4 Myr ago[15]. Recent study using the mitochondrial cytochrome *b* gene has elucidated the detailed phylogenetic relations among four genetically different wild populations

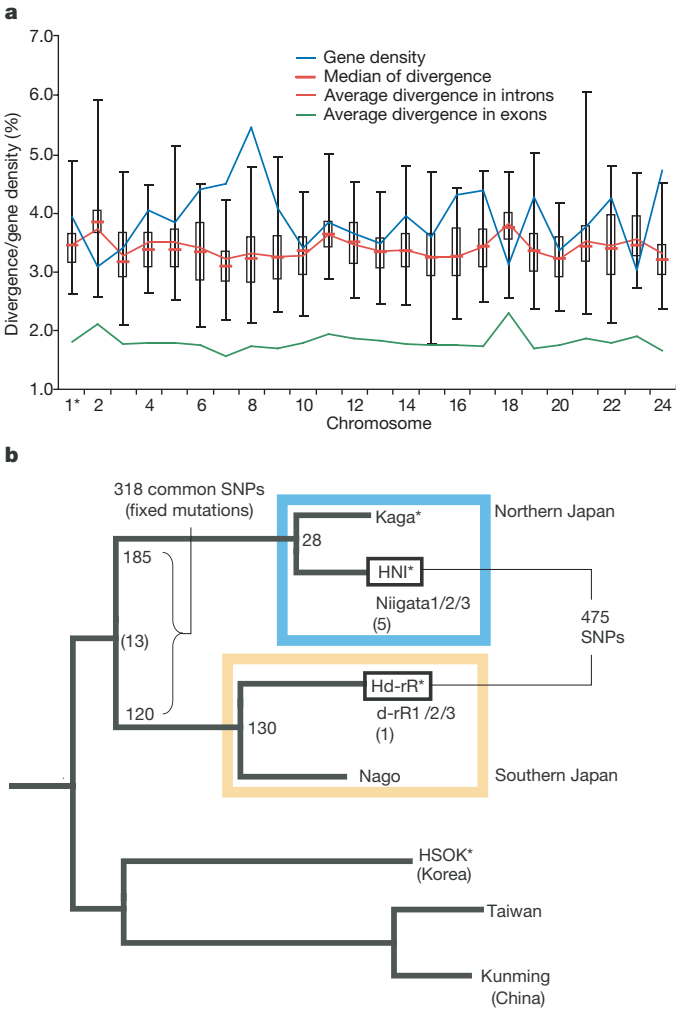

Figure 1 | Genetic variation between two medaka strains. a, Sequence diversity in 200 Kb segments of chromosomes. Box edges, the quartiles; vertical bars, the range; 1*, sex chromosome. **b**, Phylogenetic analysis of the SNPs identified between Hd-rR and HNI during regional diversification. From the parental populations of HNI and Hd-rR, six individuals named Niigata 1/2/3 and d-rR 1/2/3 were analysed to find 5 and 1 additional SNPs, respectively. * indicates an inbred strain.

(Northern Japanese, Southern Japanese, East Korean and Chinese–West Korean)[15] (Fig. 1b). Despite the indubitable accumulation of genetic variation, they all can mate and produce healthy and fertile offspring. The massive SNP resources and living stock of medaka regional strains enabled us to perform a genome-wide SNP analysis for the history of wild populations during regional diversification. We analysed genetic variation in 47 PCR-amplified regions (approximately 24.9 Kb in total) as shown in Fig. 1b. Nago and Kaga strains were chosen because they are most distantly related to Hd-rR and HNI within each population, respectively[15]. We focused on 475 SNP sites identified between Hd-rR and HNI. The comparative analysis first revealed that 130 (27%) and 28 (5.8%) of 475 SNP sites were polymorphic in the southern and northern populations, respectively, and one mutation happened to be shared by both populations (Fig. 1b). The remaining 318 (475−(130+28−1)) Hd-rR/HNI SNPs were thus preserved or fixed between the southern and northern populations (common SNPs in Fig. 1b). In comparison with the consensus sequence of the outgroup, Korea–Taiwan–China medaka (a regional medaka species with genetic variations compared with the Japanese medaka) strains, 120 and 185 mutation events were assigned to the southern and northern lineages, respectively, whereas 13 events were unclear. Overall, the total number of mutations introduced over 4 Myr is 250 (120+130) for Hd-rR and 213 (185+28) for HNI, indicating that almost the same level of mutations accumulated in the history of the two strains (chi-squared test, $\chi^2 = 3.11$, d.f. = 1, $P < 0.05$). On the other hand, the lower levels of polymorphism in the northern lineage support their rapid and recent expansion from a small size population (that is, a bottleneck effect), which has been suggested in a study with mitochondrial DNA[15]. More importantly, the high ratio of common SNPs (>65%) as well as few shared polymorphic sites indicates a strict genetic separation between the two populations for 4 Myr without major species differentiation (or speciation). Further analysis will shed light on the genetic variations and speciation in vertebrates.

To generate the medaka gene catalogue, we obtained over one million 5′-end serial analysis of gene expression (5′SAGE) tags[16], which correspond to the transcription start sites. The tags were grouped and, on the basis of transcription start site information, we predicted 20,141 non-redundant gene structures by a newly developed algorithm that uses Genscan. These predicted genes are really transcribed, however, the predicted exons may not be entirely accurate as an unavoidable consequence of the *ab initio* method. Thus, we compared the predicted gene structures with 85 known full-length complementary DNA sequences. They matched by BLASTP at 83/85 (97.6%) (expected value (*E*-value) $< 10^{-10}$) or at 72/85 (84.7%) (*E*-value $< 10^{-50}$). Furthermore, 407 (58.6%) out of 694 exons in the 85 cDNA sequences were perfectly predicted by our algorithm in which Genscan was used with the default setting for vertebrates. Full details are in Supplementary Information.

To characterize the 20,141 predicted medaka genes, we used TBLASTX[17] to compare them with the genes of six other vertebrates—human, *Tetraodon nigroviridis*, zebrafish, *Takifugu rubripes*, chicken and mouse—in the RefSeq database, and also with the gene clusters of aves, amphibia, ray-finned fish and ascidiacea in the UniGene database. We found that 3,727 have no homologues even with loose criterion (*E*-value $< 10^{-4}$) in a TBLASTX search (Fig. 2a). Then, we examined whether these novel gene candidates have any unique protein domains according to a PROSITE scan (http://ca.expasy.org/prosite/). With a stringent search criteria, by which unique domains of more than 20 amino acids were detected for 35.1% of non-novel predicted genes, only 30/3,727 (0.8%) of the new gene candidates were recognized as having known domains, suggesting that most of them are structurally unique. Interestingly, 64.4% of these candidates have CpG islands on their upstream regions and the ratio is higher than the average (50.5%) of all the predicted genes.

Among the 3,727 new gene candidates, 2,078 had no similarity even to medaka expressed sequence tags. Because 1,443 of 2,078 had open reading frames shorter than 100 amino acids, many of them might be non-coding; therefore we tried to validate the new gene candidates by testing the expression of the predicted transcripts by PCR with reverse transcription (RT–PCR). As a result, we estimated the number of 'true novel genes' to be 1,287 out of 2,078 (see Supplementary Information for details). Taking into account the remaining 1,649 expressed-sequence-tag-supported ones, at least 2,936 medaka-specific novel genes are estimated in the draft genome. This large number of new genes will provide a valuable genetic resource for medaka biology.

Of the 20,141 predicted medaka genes, we further analysed 11,617 (57.7%) that had human orthologues. Four-thousand three-hundred and forty-two (21.6%) genes constituted medaka–human reciprocally best 1:1 orthologue pairs such that TBLASTX *E*-values are less than 10^{-4} and the ratios of reciprocally aligned portions shared by two orthologues are at least 30%. Of these orthologue pairs, 2,292 were assigned the gene ontology (GO) 'biological process' annotations of their corresponding human genes (Fig. 2b). Orthologues involved in carbohydrate metabolism, alcohol metabolism and catabolism were found to be more conserved than genes implicated in the immune response, transcription, reproduction, apoptosis and stress response. Furthermore, 925 of the 1,395 human disease genes in the Online Mendelian Inheritance in Man (OMIM) database have strong orthologues among the medaka genes, such as *A2M* (alpha-2-macroglobulin) and *PSEN1* (presenilin 1), which is implicated in Alzheimer's disease, and *TP53* (tumour suppressor protein p53) and *DLEC1* (deleted in lung cancer-1; also known as *CLEC4C*) which are both involved in carcinogenesis.

To gain insight into gene evolution and species differentiation, we examined rapidly and slowly evolving gene categories between the two medaka inbred strains Hd-rR and HNI. The average K_A/K_S ratio of 8,889 qualified medaka predicted genes between the two strains is 0.413—significantly higher than that for the human–chimpanzee lineage (0.23 for K_A/K_S)[12], which has experienced major speciation for 5 Myr. We used the median K_A/K_S ratio of each GO-based functional category of medaka genes, and plotted it against that of human–chimpanzee (Fig. 2c). We focus here on the specific categories referred to by previous analyses among mammalian species, which included immunity, host defence, reproduction and olfaction as rapidly evolving categories, and intracellular signalling, neurogenesis

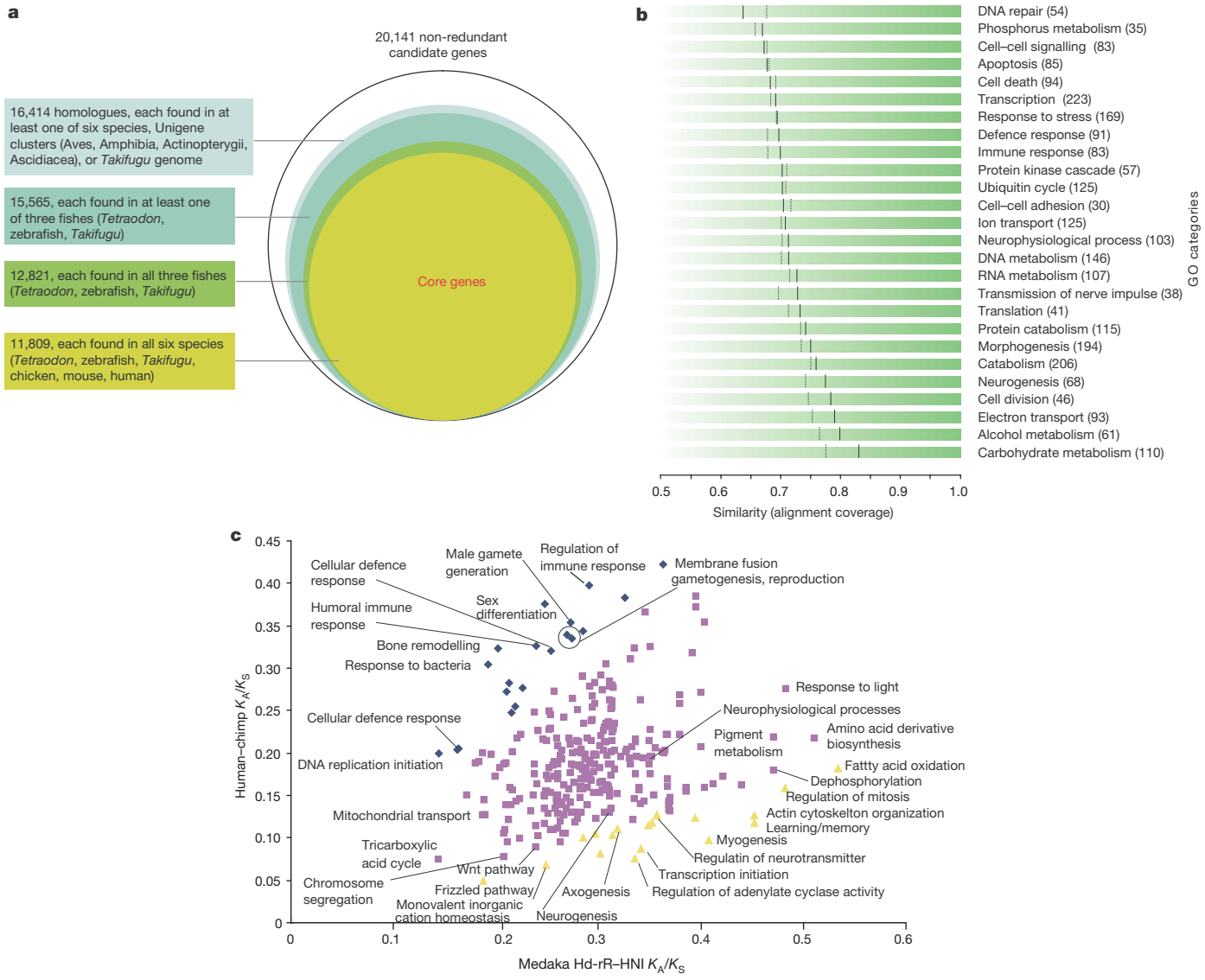

Figure 2 | Medaka genes. **a**, Breakdown of medaka gene homologues in other species. **b**, Similarity of medaka–human 1:1 orthologous pairs is arranged by GO category. Number of orthologues is in parentheses. Bars show median (thick) and mean (dashed) similarity of orthologues. **c**, Dots represent pairs of medians of K_A/K_S ratios in the medaka strains and the human–chimpanzee lineage for GO categories with ≥10 genes. Rapidly and slowly evolving GO categories in medaka relative to the hominid lineage are coloured yellow and blue, respectively.

and neurophysiology as slowly evolving ones[12]. The rapidly evolving categories are thought to be involved in adaptation to environment and sexual separation, both of which are essential processes during and/or after speciation. Intriguingly, these rapidly evolving categories are not evident in the medaka lineage, whereas mammalian slowly evolving neural-related categories exhibit relaxed constraint. Thus the reduced rate of evolution in the reproduction- and sex-related categories might explain why the two medaka strains can mate and produce fertile offspring even after a long period of geographical and genetic separation. One example is the ubiquitin specific protease 9Y (*USP9Y*) gene for male gametogenesis on hominid Y chromosome ($K_A/K_S = 1.0$) or on medaka LG21 autosome ($K_A/K_S = 0.27$). In general, the extent of phenotypic variation between organisms is not strictly related to the degree of sequence variation, and this is also the case for species differentiation of medaka. Our comparative analysis thus suggests that differential selective pressures act on specific categories in a lineage-specific manner, and this may contribute to a pattern of evolution, for example, adaptation with or without speciation.

Whole-genome duplication (WGD) and subsequent asymmetric changes in duplicated genes are thought to have an important role in genome evolution[18]. Recently, several studies have examined WGD in the teleost lineage[3,19–23] and have reconstructed ancestral karyotypes using the available genomic data[3,19,20]. Although these previous studies attempted to estimate the number of proto-chromosomes before the WGD event, interchromosomal genome rearrangements during evolution, and the correspondence between the proto-chromosomes and present chromosomes, there have been no clear answers concerning the timing of major interchromosomal rearrangements. Thus, we have conducted large-scale four-way comparisons of the medaka, human, zebrafish and *Tetraodon* genomes (see Supplementary Information for full details of scenario construction).

Here we summarize our scenario of genome evolution from the ancestral karyotype to the three teleost genomes. Figure 3 illustrates one example of how we inferred the ancestral chromosomes, and Fig. 4 depicts our scenario. The date we adopt for the WGD and lineage divergence is based on molecular clock estimates[24,25]. The key events we propose are as follows.

- In a relatively short period of ~50 Myr after the WGD event (336–404 Myr ago), the MTZ-ancestor (the last common ancestor of medaka, *Tetraodon*, and zebrafish) had 24 chromosomes and had undergone 8 major interchromosomal rearrangements (2 fissions, 4 fusions and 2 translocations).
- In contrast, since zebrafish diverged about 314–332 Myr ago, the medaka genome has preserved its ancestral genomic structure without undergoing major interchromosomal rearrangements for more than 300 Myr. The *Tetraodon* genome underwent fusion events on three occasions after separating from the medaka lineage about 184–198 Myr ago.

The zebrafish genome seems to have experienced many interchromosomal rearrangements during evolution by extensive translocations, but the precise scenario remains to be solved because of a relatively small number of zebrafish genome markers that we used in the present study. Nevertheless, these zebrafish genome markers were useful in revealing the eight major interchromosomal rearrangements

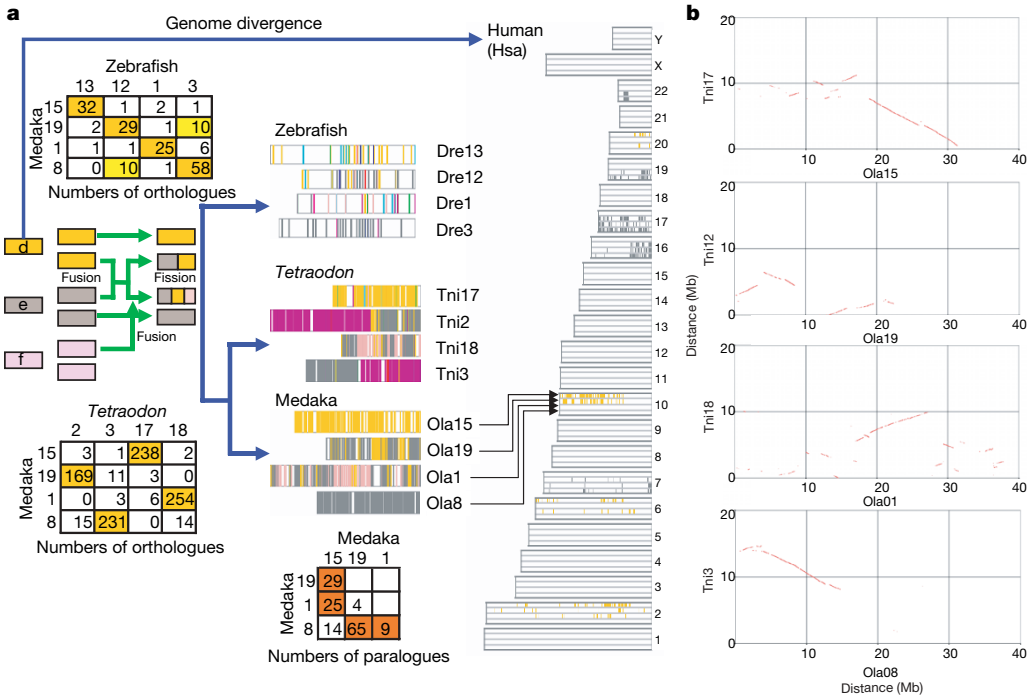

Figure 3 | Reconstruction of proto-chromosomes. **a**, Doubly conserved synteny (DCS) blocks[3,30] between one human and two medaka chromosomes were searched to identify duplicated medaka chromosomes. Human–medaka orthologues are plotted by orange or grey bars in the rows of human chromosomes (Hsa), corresponding to counterpart medaka chromosomes; in this case, four rows corresponding to Ola15, Ola19, Ola1, and Ola8. The series of co-occurrences of orthologues (orange plots) in the rows of Ola15 and Ola19 in Hsa10 shows a DCS block, and similarly more DCS blocks are found in other chromosomes. Medaka chromosomes that share a DCS block are thought to share a proto-chromosome. Ola8 and Ola15 share no DCS blocks, implying that they were from distinct proto-chromosomes. On the basis of this logic, we deduced that parts of Ola15, Ola19 and Ola1 are derived from a proto-chromosome (d, orange), whereas parts of Ola19, Ola1 and Ola8 are from another proto-chromosome (e, grey). Purple, traces of teleost proto-chromosome c (see Fig. 4 and Supplementary Fig. 13). Next, to analyse chromosome fission events, we acknowledge that two chromosomes generated by a fission event are unlikely to have common paralogues derived from the WGD in teleost. The table (centre, bottom) shows the number of paralogues between medaka chromosomes. The fairly small number between Ola1 and Ola19 imply that they were the results of chromosome fission. Finally, the correspondence among chromosomes of the three fishes was determined using the orthologue information in the tables (left, top and bottom). See Supplementary Fig. 13 for information on the other proto-chromosomes. **b**, The dot plots exhibit focal synteny blocks between medaka and *Tetraodon* chromosomes, which presents more accurate synteny than does the table of orthologue numbers. Ola, *Oryzias latipes* chromosome; Dre, *Danio rerio* chromosome; Tni, *Tetraodon nigroviridis* chromosome; and Hsa, *Homo sapiens* chromosome.

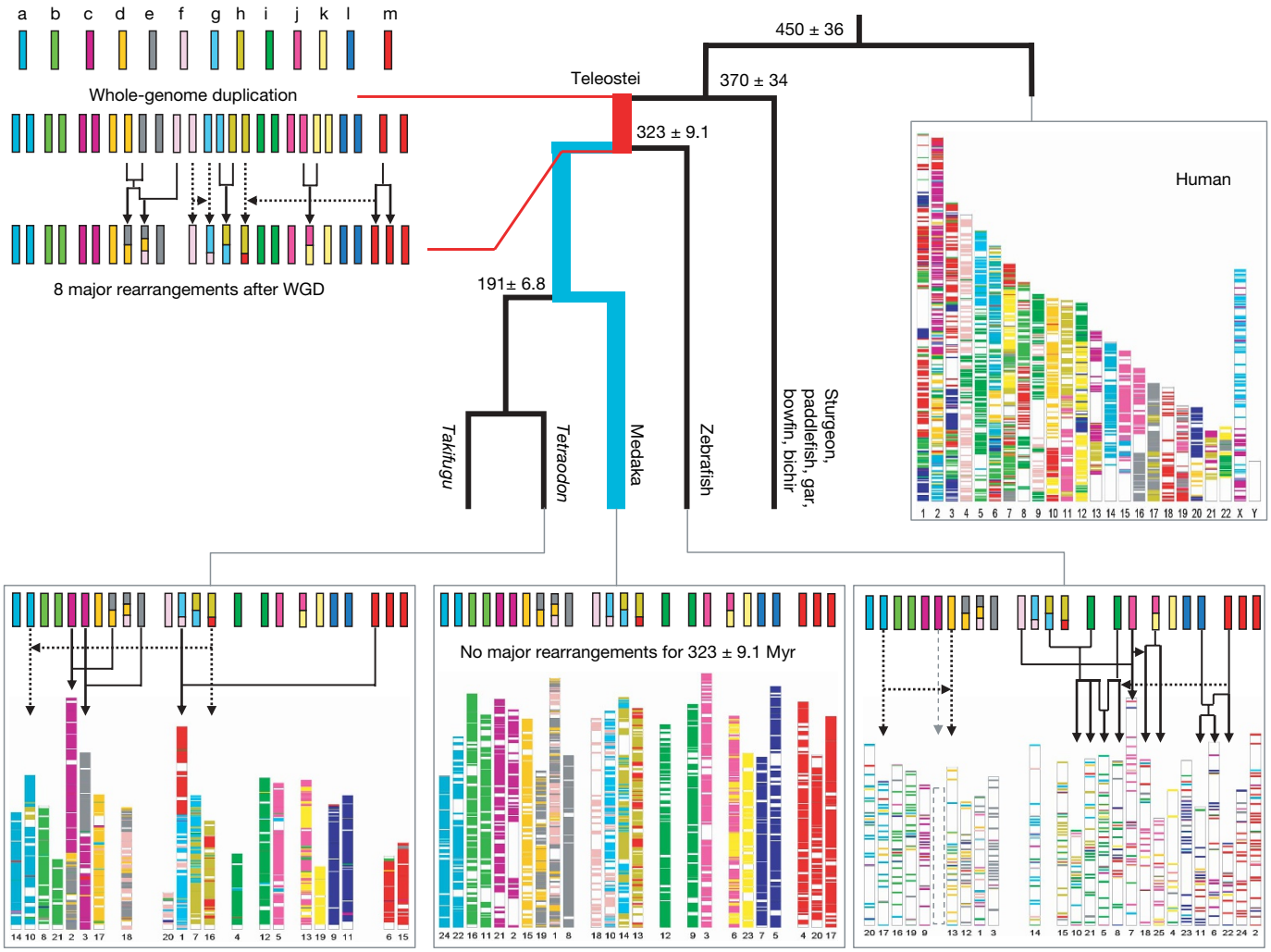

Figure 4 | Teleost genome evolution. The figure depicts a model for the distribution of ancestral chromosome segments in the human, zebrafish, medaka and *Tetraodon* genomes. Thirteen reconstructed ancestral chromosomes are represented by the coloured bars, and the genomic regions originating in the ancestral chromosomes (a–m) have the same colour coding. Major rearrangements are represented by arrows and lineage-specific small-scale translocations by dotted arrows. A dashed box in the zebrafish genome indicates that most parts of the corresponding ancestral chromosome were lost by extensive rearrangements. Individual ancestral chromosomes are labelled a–m, which corresponds to the evolutionary scenarios in Supplementary Information.

before the divergence of the three fishes. Because more than half of all the teleost species examined have either 24 or 25 chromosomes, it has been speculated that the teleost ancestor also had 24 or 25 chromosomes[19]. This is consistent with our current findings that the reconstructed MTZ-ancestor, which is the common ancestor of most teleosts[1,26,27], had 24 chromosomes. Our scenario is the most likely one, but there is an alternative: that Ola10, 13, and 14 are derived from one ancestral chromosome instead of two; and one of the duplicated ancestral chromosomes became Ola14, whereas the other underwent a fission event, yielding Ola10 and Ola13 (12 ancestral chromosome model). The alternative scenario seems unlikely because it assumes that the fission occurred in the same position in human and fish chromosomes derived from the common ancestor (Supplementary Fig. 15).

The high-quality medaka draft genome will provide a key resource for developmental genetics and serve as an important reference for various ray-finned fishes that have yet to be sequenced. Cichlids and stickleback, which are newly emerging model systems for understanding the genetic basis of vertebrate speciation, are evolutionarily closer to medaka than zebrafish. The same is true for many commercially important fish species, which include tuna, flounder, sea bream and fugu as well[28]. Indeed, the medaka and *Tetraodon* genomes have long synteny blocks in the genome (Supplementary Fig. 12). Furthermore, there are many close relatives of medaka that are indigenous from East to Southeast Asia[29], many of which are now maintained in laboratories. Together with the medaka draft genome, low-coverage shotgun genome sequences of these fish species will shed further light on the mechanisms underlying speciation and diversity, which have yet to be fully addressed at the genome-sequence level.

METHODS

Full details of methods are described in Supplementary Information. The medaka strain Hd-rR was provided by Y. Ishikawa. Other strains (HNI, d-rR, Kaga, Nago, HSOK, Taiwan and Kunming) were from our (H.T. and M.S.) laboratory stocks, except for Niigata from H. Mitani. These strains are available from the National BioResource Project (http://shigen.lab.nig.ac.jp/medaka/). Sperm DNA of Hd-rR was provided by M. Matsuda and used for whole-genome shotgun. Genomic DNA was also prepared from male adult bodies. Messenger RNA for 5′SAGE analysis was obtained from 0–7-day-old embryos and adult body tissue. Whole-genome shotgun assembly was made with the RAMEN assembler (to be published elsewhere) and using adaptations to various problems, including PCR slippage. Ultracontigs were produced by anchoring scaffolds to the chromosomes by using genetic markers, including a large number of SNP markers. The descriptions of sequence assembly refer to the latest assembly version 1.0, whereas other analyses were based on version 0.9. The two assemblies have almost the same contigs and scaffolds, but the former assembly has longer ultracontigs because more genetic markers were integrated. The

medaka gene catalogue was generated by our transcription-start-site-based gene prediction algorithm. The point of this algorithm is to enable us to predict the first exon and 5′ UTR, which were difficult to predict solely by a conventional gene prediction tool like Genscan. A formula was developed to define CpG islands for medaka. Novel repeats that occupied 9.2% of the medaka genome were found by our *de novo* repeat detection algorithm. The synonymous (K_S) and the non-synonymous (K_A) substitution rates of individual genes were calculated using the PAML package. The method for reconstruction of the ancestral karyotype is described briefly in the legend of Fig. 3. The method used orthologous chromosome correspondence among the three teleost genomes, paralogous chromosome correspondence between the medaka and *Tetraodon* genomes, and doubly conserved synteny blocks between the medaka and human genomes.

Received 10 November 2006; accepted 11 April 2007.

1. Nelson, J. S. *Fishes of the World* (John Wiley & Sons, New York, 1994).
2. Aparicio, S. *et al.* Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297,** 1301–1310 (2002).
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. Aida, T. On the inheritance of color in a fresh-water fish, *Aplocheilus latipes* Temminck and Schlegel, with special reference to sex-linked inheritance. *Genetics* **6,** 554–573 (1921).
5. Yamamoto, T. Artificially induced sex-reversal in genotypic males of the Medaka (*Oryzias latipes*). *J. Exp. Zool.* **123,** 571–594 (1953).
6. Matsuda, M. *et al. DMY* is a Y-specific DM-domain gene required for male development in the medaka fish. *Nature* **417,** 559–563 (2002).
7. Furutani-Seiki, M. & Wittbrodt, J. Medaka and zebrafish, an evolutionary twin study. *Mech. Dev.* **121,** 629–637 (2004).
8. Hyodo-Taguchi, Y. Inbred strains of the medaka, *Oryzias latipes*. *Fish Biol. J. Medaka* **8,** 11–14 (1990).
9. Wittbrodt, J., Shima, A. & Schartl, M. Medaka—a model organism from the Far East. *Nature Rev. Genet.* **3,** 53–64 (2002).
10. 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).
11. Schwartz, S. *et al.* Human–mouse alignments with BLASTZ. *Genome Res.* **13,** 103–107 (2003).
12. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437,** 69–87 (2005).
13. Zhang, J. Evolution of *DMY*, a newly emergent male sex-determination gene of medaka fish. *Genetics* **166,** 1887–1895 (2004).
14. Kondo, M., Nagao, E., Mitani, H. & Shima, A. Differences in recombination frequencies during female and male meioses of the sex chromosomes of the medaka, *Oryzias latipes*. *Genet. Res.* **78,** 23–30 (2001).
15. Takehana, Y., Nagai, N., Matsuda, M., Tsuchiya, K. & Sakaizumi, M. Geographic variation and diversity of the cytochrome *b* gene in Japanese wild populations of medaka, *Oryzias latipes*. *Zoolog. Sci.* **20,** 1279–1291 (2003).
16. Hashimoto, S. *et al.* 5′-end SAGE for the analysis of transcriptional start sites. *Nature Biotechnol.* **22,** 1146–1149 (2004).
17. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
18. Ohno, S. *Evolution by Gene Duplication* (Springer, New York, 1970).
19. Naruse, K. *et al.* A medaka gene map: the trace of ancestral vertebrate proto-chromosomes revealed by comparative gene mapping. *Genome Res.* **14,** 820–828 (2004).
20. Woods, I. G. *et al.* The zebrafish gene map defines ancestral vertebrate chromosomes. *Genome Res.* **15,** 1307–1314 (2005).
21. Taylor, J. S., Braasch, I., Frickey, T., Meyer, A. & Van de Peer, Y. Genome duplication, a trait shared by 22,000 species of ray-finned fish. *Genome Res.* **13,** 382–390 (2003).
22. Christoffels, A. *et al.* Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. *Mol. Biol. Evol.* **21,** 1146–1151 (2004).
23. Vandepoele, K., De Vos, W., Taylor, J. S., Meyer, A. & Van de Peer, Y. Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc. Natl Acad. Sci. USA* **101,** 1638–1643 (2004).
24. Crow, K. D., Stadler, P. F., Lynch, V. J., Amemiya, C. & Wagner, G. P. The "fish-specific" hox cluster duplication is coincident with the origin of teleosts. *Mol. Biol. Evol.* **23,** 121–136 (2006).
25. Yamanoue, Y., Miya, M., Inoue, J. G., Matsuura, K. & Nishida, M. The mitochondrial genome of spotted green pufferfish *Tetraodon nigroviridis* (Teleostei: Tetraodontiformes) and divergence time estimation among model organisms in fishes. *Genes Genet. Syst.* **81,** 29–39 (2006).
26. Hedges, S. B. & Kumar, S. Genomics. Vertebrate genomes compared. *Science* **297,** 1283–1285 (2002).
27. Miya, M. *et al.* Major patterns of higher teleostean phylogenies: a new perspective based on 100 complete mitochondrial DNA sequences. *Mol. Phylogenet. Evol.* **26,** 121–138 (2003).
28. Inoue, J. G., Miya, M., Tsukamoto, K. & Nishida, M. Basal actinopterygian relationships: a mitogenomic perspective on the phylogeny of the "ancient fish". *Mol. Phylogenet. Evol.* **26,** 110–120 (2003).
29. Takehana, Y., Naruse, K. & Sakaizumi, M. Molecular phylogeny of the medaka fishes genus *Oryzias* (Beloniformes: Adrianichthyidae) based on nuclear and mitochondrial DNA sequences. *Mol. Phylogenet. Evol.* **36,** 417–428 (2005).
30. Kellis, M., Birren, B. W. & Lander, E. S. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428,** 617–624 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by a Grant-in-Aid for Scientific Research on Priority Areas "Genome" from the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT), and the Japan Science and Technology Corporation (JST). We thank the Human Genome Center, University of Tokyo for computational time, and the National BioResource Project of MEXT for medaka strain supply and other support. We thank T. Koh for his assistance in genome assembly, A. Mori, T. Shishiki and T. Furudate for their assistance in genome browser development, Y. Ishikawa, H. Mitani and M. Matsuda for medaka resources, M. Shinya and T. Kimura for the Hd-rR/HNI F_2 panel, N. Shibata for *O. curvinotus*, and A. Hase, N. Hasegawa, S. Iiyama, N. Ishihara, K. Kawaguchi, Y. Minakuchi, S. Miura, J. Miyamoto, H. Miyauchi, M. Mizukoshi, Y. Mochizuki, Y. Sugiyama, H. Takizawa-Hayashi, M. Tamiya, T. Tandoh and E. Yokoyama for their technical assistance.

Author Contributions Project planning and supervision by Y. Kohara, S.M. and H.T.; whole-genome shotgun assembly and primer design for SNP markers by M.K. and S. Sasaki.; Analysis of SNPs and genomic alternations around the transcription start sites by S. Sasaki.; reconstruction of the teleost genome evolution by Y. Nakatani; homologue, orthologue and paralogue analysis by W.Q.; Evolutionary rate analysis of protein coding genes by S. Sasaki and W.Q.; gene prediction using 5′SAGE tags by B.A.; elucidation of novel repetitive elements by T.Y.; validation of primer specificity and efficiency by M.K., W.Q., B.A. and T.Y.; medaka genome browser development by Y. Nagayasu, Y. Kasai and K.D.; supervision of bioinformatics analysis by S.M.; whole-genome shotgun sequencing and data management by K.O., S. Haga, F.O., H.N., K. Nogata, T.M., T.E., T. Shin-I and Y. Kohara; construction of fosmid and 7.5 Kb plasmid libraries by K. Naruse, T.N. and H.T.; BAC library construction and end sequencing by A.T., Y. Kuroki, A.F., T. Sasaki, A. Shimizu, S.A. and N.S.; high-density SNP map construction by K. Naruse, T.J., A. Shimada and H.T.; genetic mapping using the panel of Kumming and Hd-rR by M.S.; 5′SAGE library construction by S-i.H., J.Y., Y.L., S. Sugano and K.M.; identification of non-coding genes and microRNA (miRNA) candidates by B.A.; expression analysis of medaka novel genes and miRNA by D.K. and H.T.; and paper writing by S.M. jointly with H.T., Y. Kohara, M.K., S. Sasaki, Y. Nakatani, W.Q., B.A., K. Naruse and T.Y.

Author Information The *Oryzias latipes* whole-genome shotgun project data have been deposited at DDBJ/EMBL/GenBank under the project accessions BAAF03000000 (Hd-rR, version 0.9), BAAF04000000 (Hd-rR, version 1.0), BAAE01000000 (HNI), and ACAAA0000001–ACAAA0356693 (5′SAGE tags). The assembly and annotations are also available from UT Genome Browser (http://medaka.utgenome.org/), NIG (http://dolphin.lab.nig.ac.jp/medaka/) and Ensembl (http://www.ensembl.org/). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to Y.Kohara (ykohara@lab.nig.ac.jp), S.M. (moris@cb.k.u-tokyo.ac.jp) and H.T. (htakeda@biol.s.u-tokyo.ac.jp).