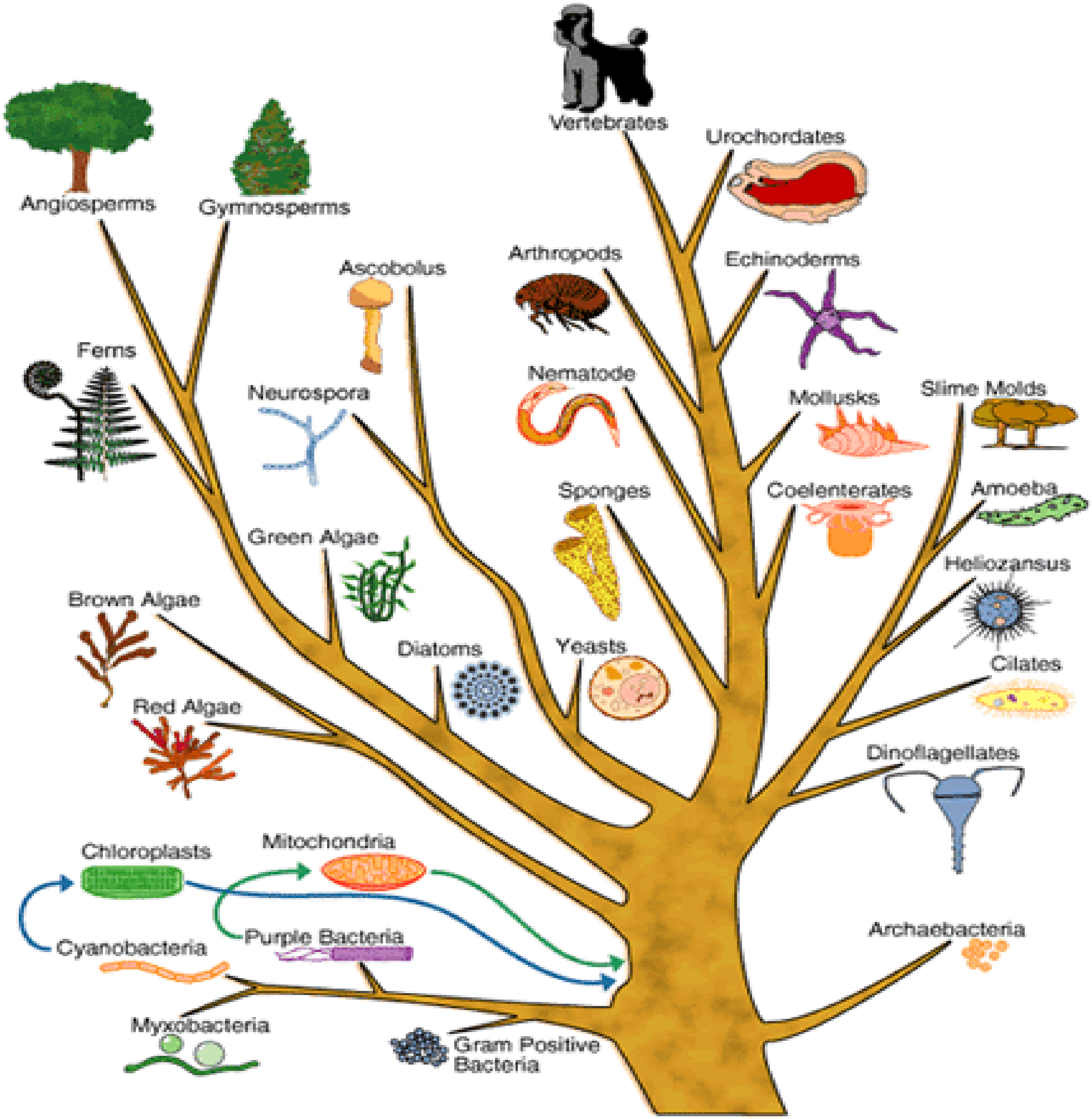
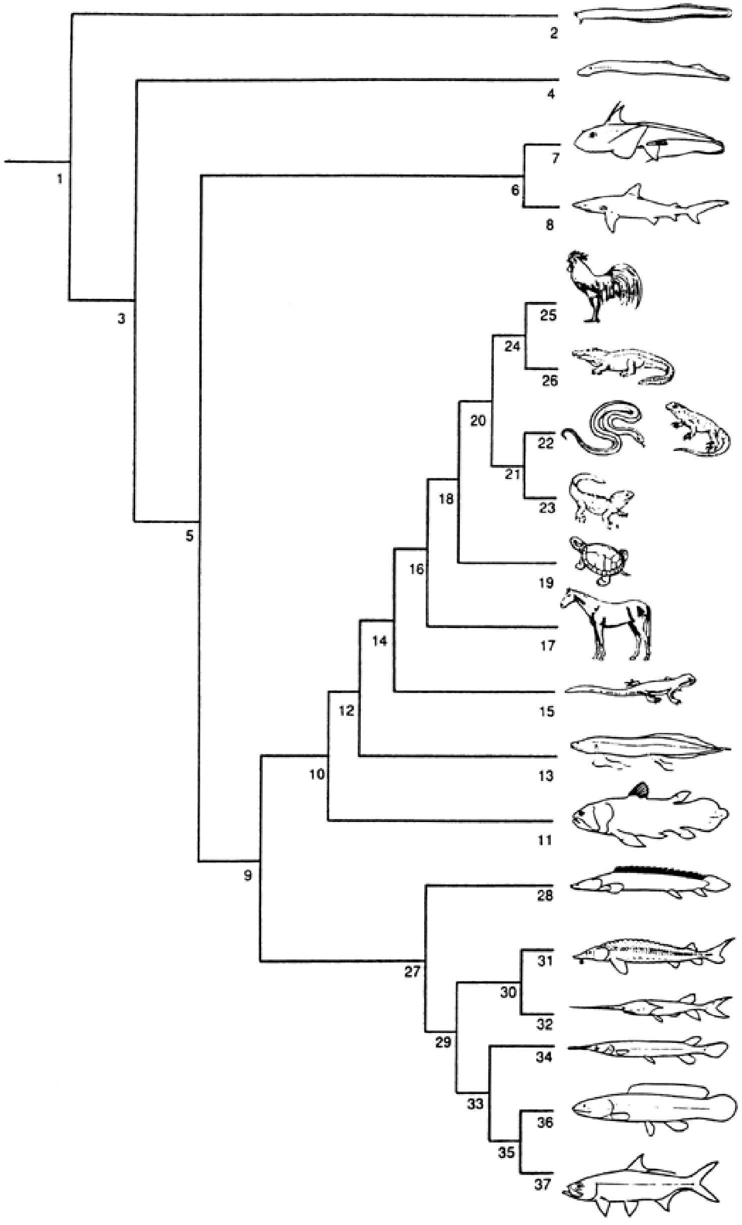
The medaka draft genome and insights into vertebrate evolution

Nature 447: 714-719

Jclub 20.11.2007

Tarmo Puurand


Vertebrates
Urochordates
Angiosperms
Gymnosperms
Arthropods
Echinoderms
Ascobolus
Ferns
Nematode
Neurospora
Mollusks
Slime Molds
Sponges
Coelenterates
Amoeba
Green Algae
Heliozansus
Brown Algae
Diatoms
Yeasts
Cilates
Red Algae
Dinoflagellates
Chloroplasts
Mitochondria
Purple Bacteria
Cyanobacteria
Archaebacteria
Myxobacteria
Gram Positive
Bacteria



1. A phylogeny of Craniata showing the position of the so-called "fishes" (nodes **2, 4, 6, 11, 13, 27**). Node number in bold: Scientific name (*Vernacular names*, total number of species in the group). Note that for "fishes", species numbers are calculated from the *Catalog of Fishes*, Eschmeyer, Version November 2000. **1**: Craniata (53,721 spp.); **2**:Myxini (Myxiniformes = Hyperotreti: *Hagfishes*, 61 spp.); **3**: Vertebrata; **4**:Petromyzontiformes = Hyperoartii (*Lampreys*, 43 spp.); **5**: Gnathostomata; **6**: Chondrichthyes (907 spp.); 7: Holocephali (*Chimaeras*, 34 spp.); **8**: Elasmobranchii (*Sharks, Guitarfishes, Sawfishes, Saw sharks, Rays, Skates, Electric rays*, 763 spp.); **9**: Osteichthyes; **10**: Sarcopterygii; **11**: Actinistia (*Coelacanths*, 2); **12**: Choanata; **13**: Dipnoi (*Lungfishes*, 6 spp.); **14**: Tetrapoda (27,541 spp.); **15**: Amphibia (Lissamphibia: *Frogs, Toads, Newts, Salamanders, Caecilians*); **16**: Amniota; **17**: Synapsida (Mammalia: *Mammals*); **18**: Sauropsida; **19**: Testudines (*Tortoises, Turtles*); **20**: Diapsida; **21** Lepidosauromorpha (Lepidosauria); **22**: Squamata (*Amphisbaenas, Lizards, Snakes*); **23**: Sphenodontida = Rhynchocephalia (*Tuatara*); **24**: Archosauromorpha; **25**: Aves (*Birds*); **26**: Crocodylia (*Alligators, Caimans, Crocodiles, Gavials*); **27**: Actinopterygii; **28**: Cladistia (*Bichirs, Reedfish*, 11); **29**: Actinopterygii; **30**: Chondrostei; **31**: Acipenseroidei (*Sturgeons*, 24 spp.); **32**: Polyodontoidei (*Paddlefishes*, 2 spp.); **33**: Neopterygii; **34**: Ginglymodi (*Gars*, 7 spp.); **35**: Halecostomi; **36**: Halecomorpha (*Bowfin*, 1 sp.); **37**: Teleostei (25,075 spp.)

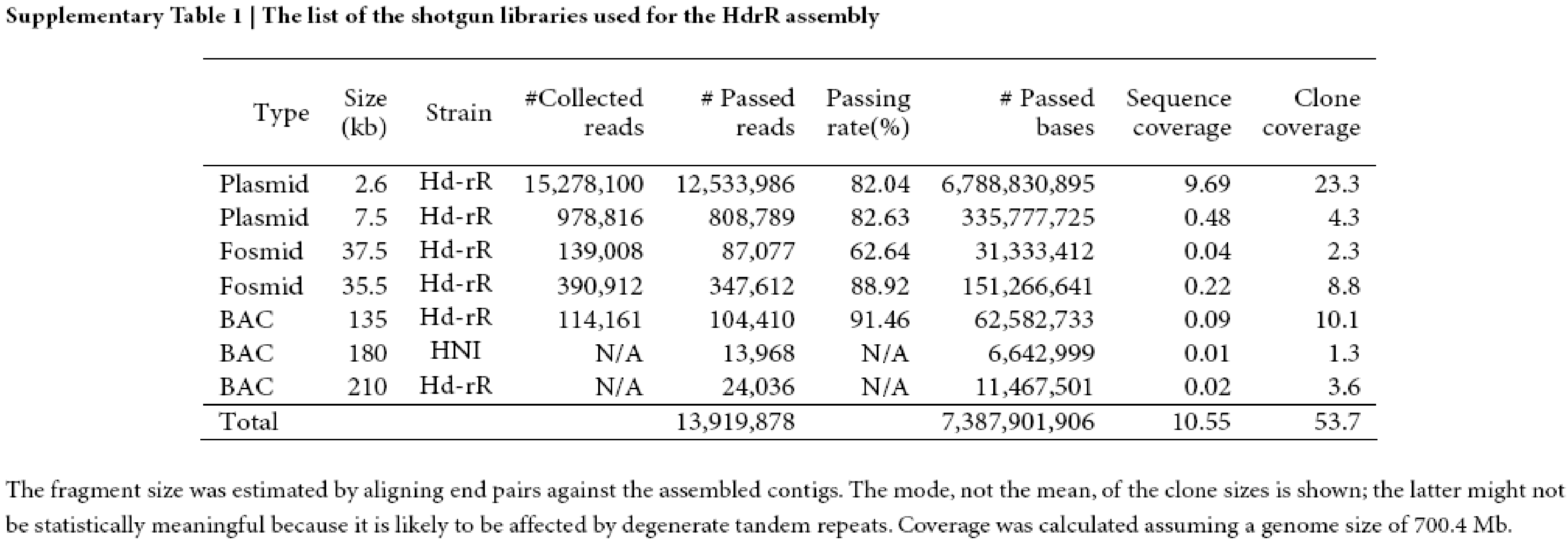
Contents

- Sequencing and assembly
- Genome landscape
- SNP analysis
- Medaka genes
- Genome evolution

Species under comparision

- Medaka- riisikala (*Oryzias latipes*)
- Zebrafish- (*Danio regio*)
- Tetraodon- (*Tetraodon nigroviridis*)

Sequencing and assembly

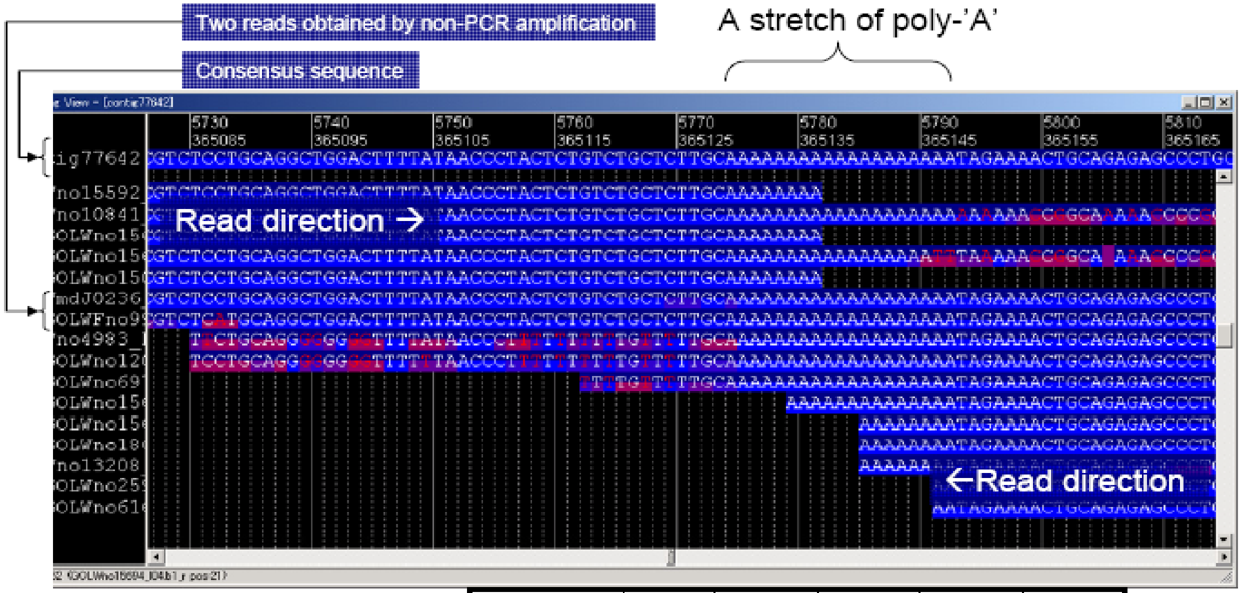


Supplementary Table 1 | The list of the shotgun libraries used for the HdrR assembly

Type	Size (kb)	Strain	#Collected reads	# Passed reads	Passing rate(%)	# Passed bases	Sequence coverage	Clone coverage
Plasmid	2.6	Hd-rR	15,278,100	12,533,986	82.04	6,788,830,895	9.69	23.3
Plasmid	7.5	Hd-rR	978,816	808,789	82.63	335,777,725	0.48	4.3
Fosmid	37.5	Hd-rR	139,008	87,077	62.64	31,333,412	0.04	2.3
Fosmid	35.5	Hd-rR	390,912	347,612	88.92	151,266,641	0.22	8.8
BAC	135	Hd-rR	114,161	104,410	91.46	62,582,733	0.09	10.1
BAC	180	HNI	N/A	13,968	N/A	6,642,999	0.01	1.3
BAC	210	Hd-rR	N/A	24,036	N/A	11,467,501	0.02	3.6
Total				13,919,878		7,387,901,906	10.55	53.7

The fragment size was estimated by aligning end pairs against the assembled contigs. The mode, not the mean, of the clone sizes is shown; the latter might not be statistically meaningful because it is likely to be affected by degenerate tandem repeats. Coverage was calculated assuming a genome size of 700.4 Mb.

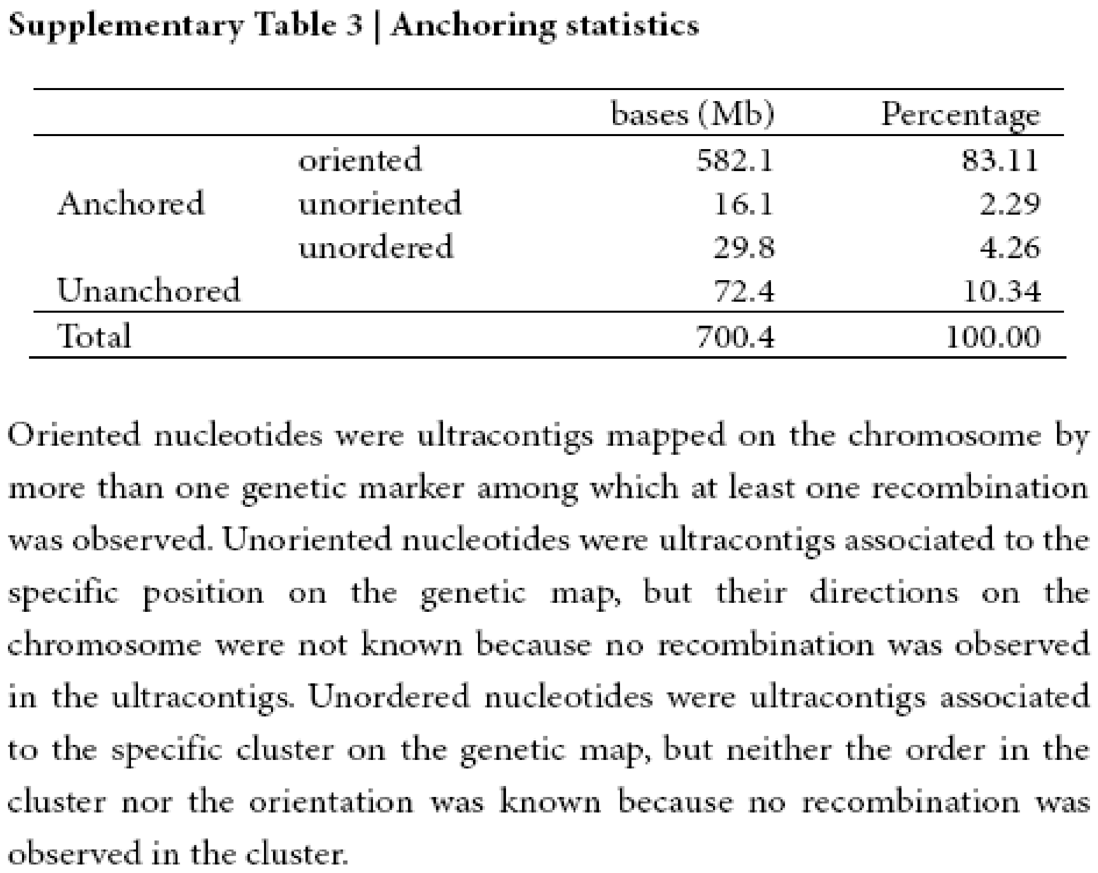
Sequencing and assembly



	G	G	G	G	G
phredQV	25 ↑	20~24	15~19	10~14	0~9

Supplementary Figure 1 | An example of PCR slippage. The five reads in the upper part of the multiple alignment go from left to right, while the nine reads at the bottom go from right to left. The two reads in between were obtained using a non-PCR based sequencing protocol, whereas the others were affected by PCR slippage. About half of the reads were terminated at a homopolymer. The other half go beyond the homopolymer, but systematic sequencing errors are observed in the subsequent regions.

Sequencing and assembly

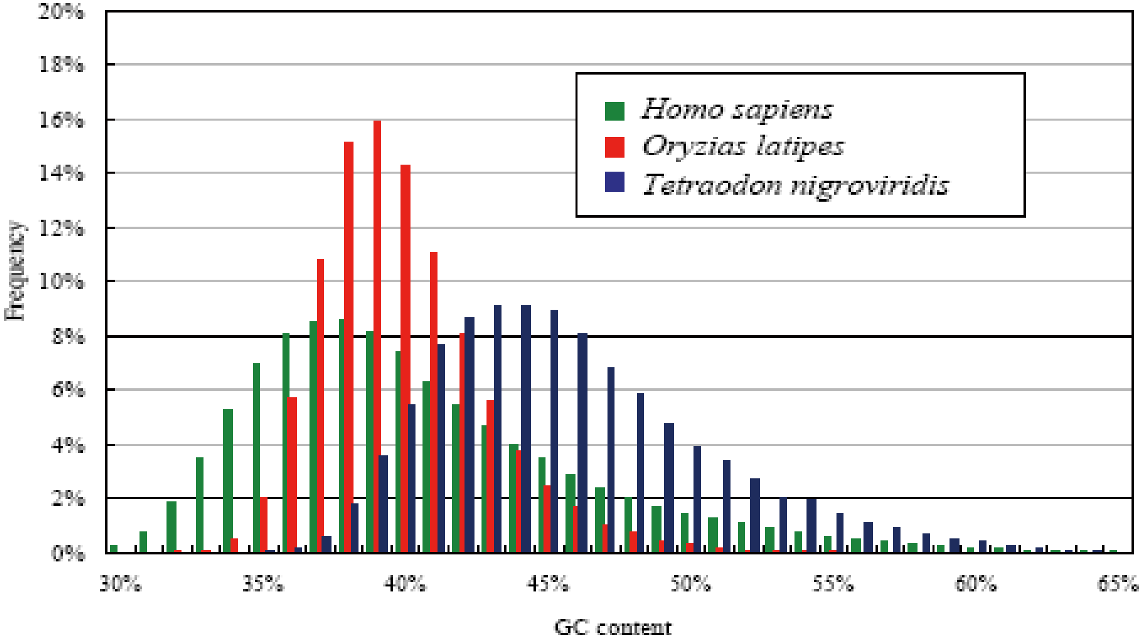


Supplementary Table 3 | Anchoring statistics

		bases (Mb)	Percentage
Anchored	oriented	582.1	83.11
	unoriented	16.1	2.29
	unordered	29.8	4.26
Unanchored		72.4	10.34
Total		700.4	100.00

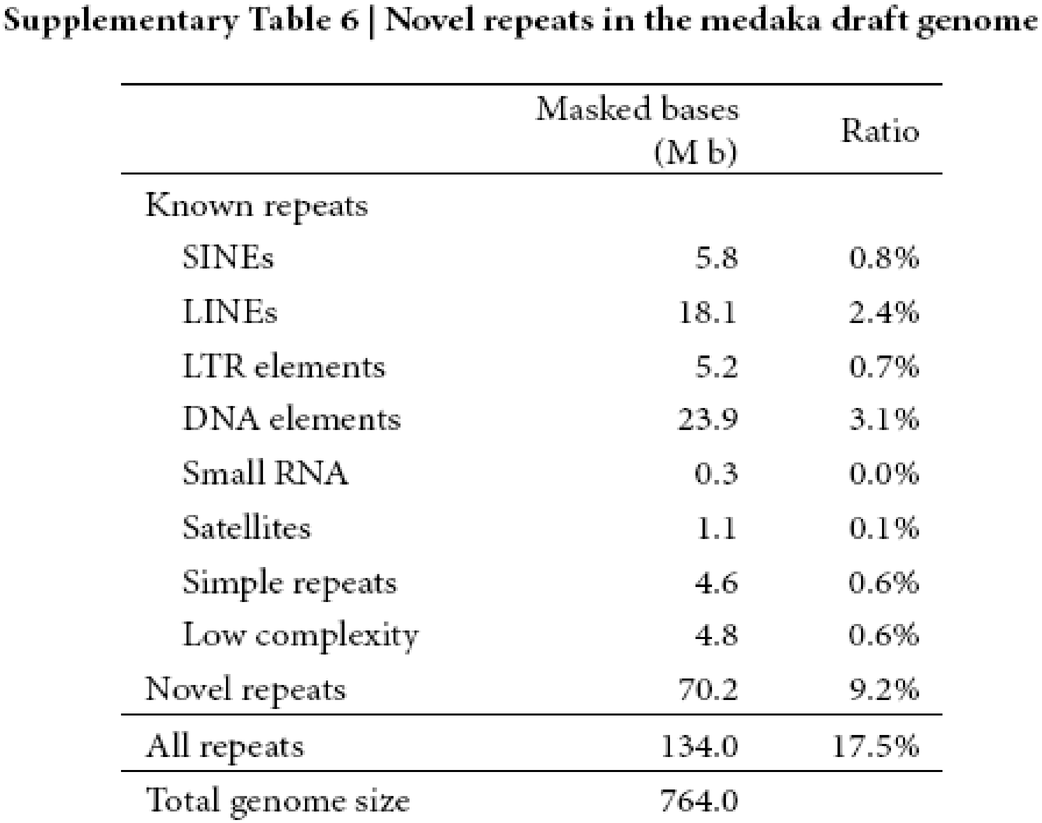
Oriented nucleotides were ultracontigs mapped on the chromosome by more than one genetic marker among which at least one recombination was observed. Unoriented nucleotides were ultracontigs associated to the specific position on the genetic map, but their directions on the chromosome were not known because no recombination was observed in the ultracontigs. Unordered nucleotides were ultracontigs associated to the specific cluster on the genetic map, but neither the order in the cluster nor the orientation was known because no recombination was observed in the cluster.

Genome landscape



Supplementary Figure 3 | Distribution of GC content calculated for 10-Kb non-overlapping windows in the medaka (red), *Tetraodon* (blue) and human (green) genomes.

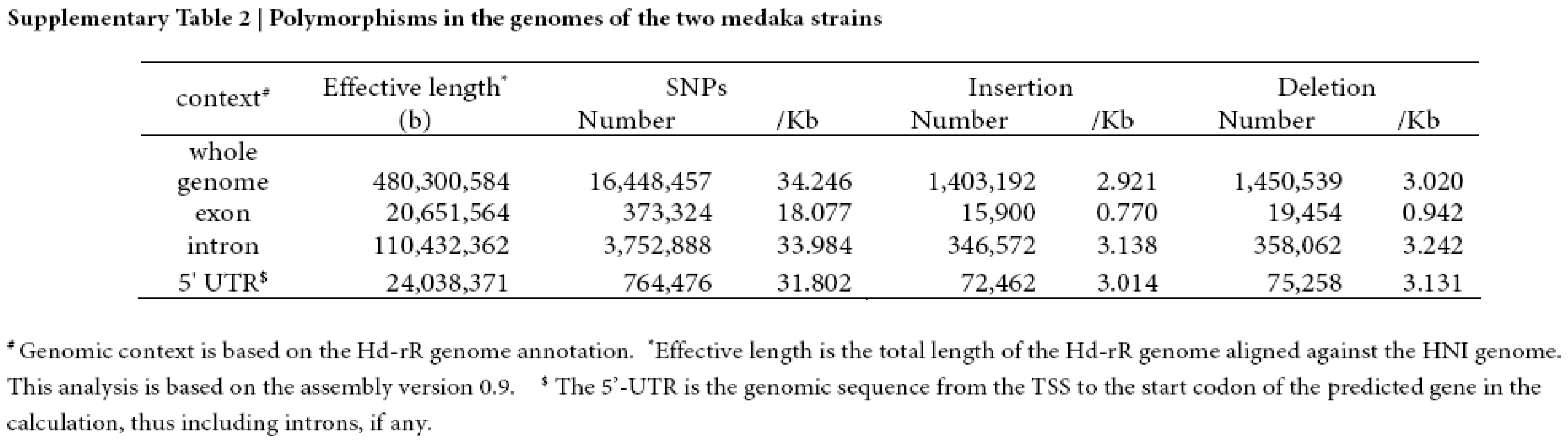
Genome landscape



Supplementary Table 6 | Novel repeats in the medaka draft genome

	Masked bases (M b)	Ratio
Known repeats		
SINEs	5.8	0.8%
LINEs	18.1	2.4%
LTR elements	5.2	0.7%
DNA elements	23.9	3.1%
Small RNA	0.3	0.0%
Satellites	1.1	0.1%
Simple repeats	4.6	0.6%
Low complexity	4.8	0.6%
Novel repeats	70.2	9.2%
All repeats	134.0	17.5%
Total genome size	764.0	

SNP analysis

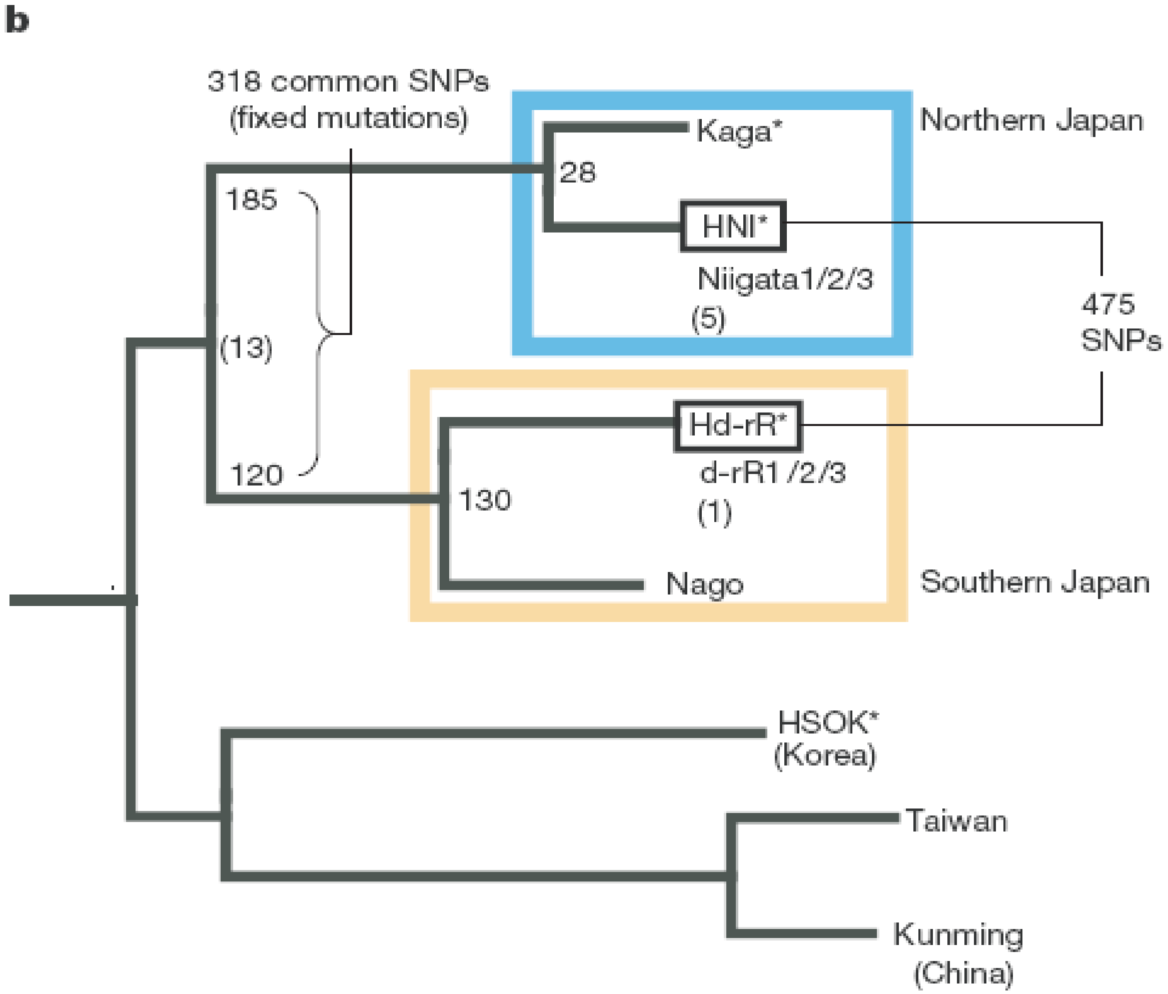


Supplementary Table 2 | Polymorphisms in the genomes of the two medaka strains

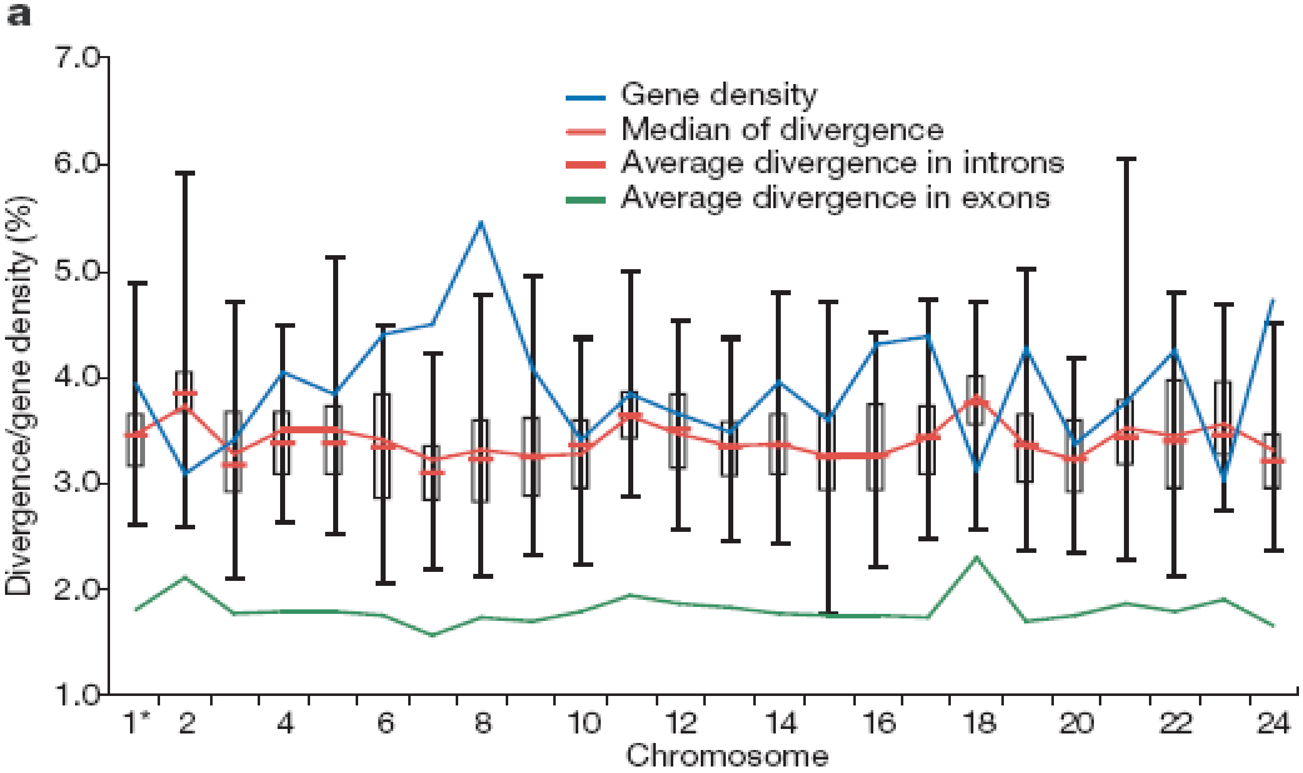
context[#]	Effective length[*]	SNPs		Insertion		Deletion	
	(b)	Number	/Kb	Number	/Kb	Number	/Kb
whole genome	480,300,584	16,448,457	34.246	1,403,192	2.921	1,450,539	3.020
exon	20,651,564	373,324	18.077	15,900	0.770	19,454	0.942
intron	110,432,362	3,752,888	33.984	346,572	3.138	358,062	3.242
5' UTR[$]	24,038,371	764,476	31.802	72,462	3.014	75,258	3.131

[#] Genomic context is based on the Hd-rR genome annotation. [*]Effective length is the total length of the Hd-rR genome aligned against the HNI genome. This analysis is based on the assembly version 0.9. [$] The 5'-UTR is the genomic sequence from the TSS to the start codon of the predicted gene in the calculation, thus including introns, if any.

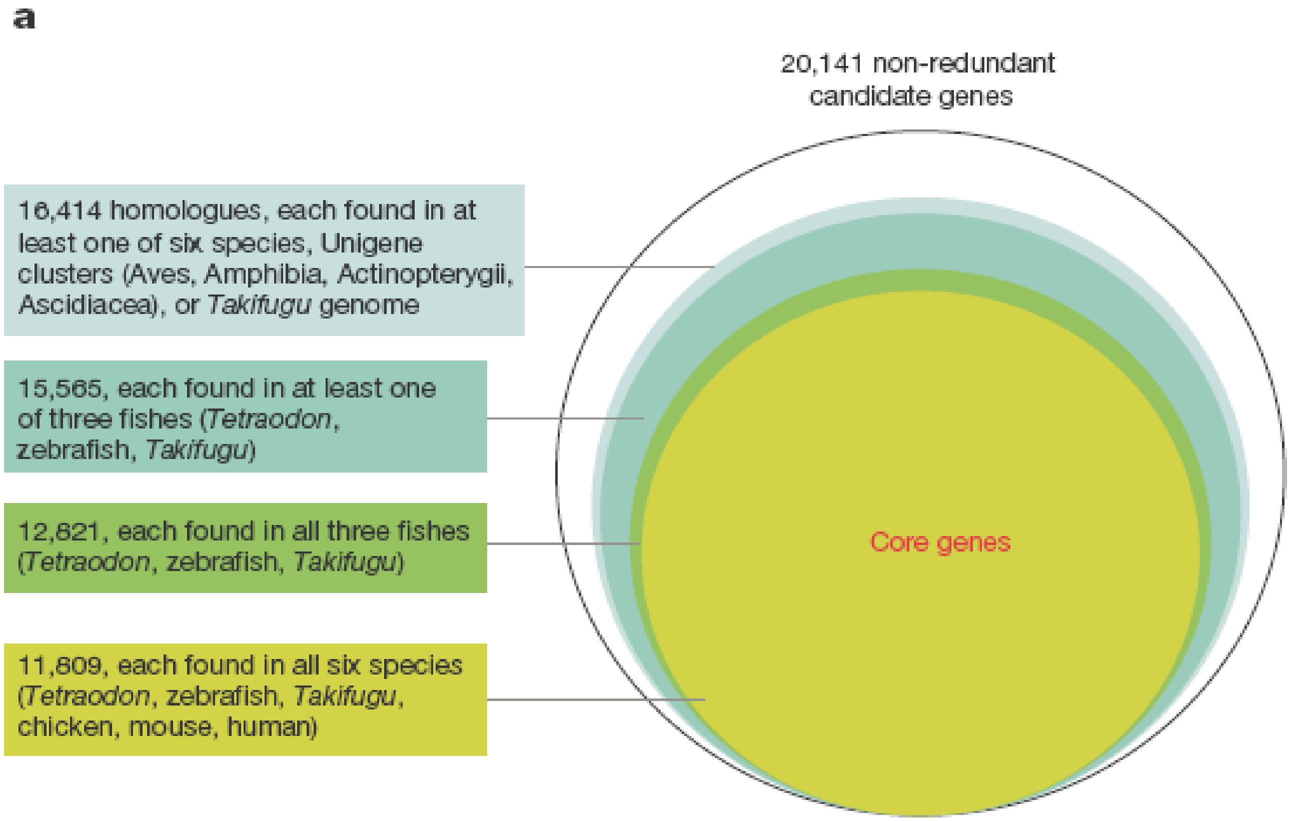
SNP analysis



Medaka genes



Medaka genes



20,141 non-redundant candidate genes

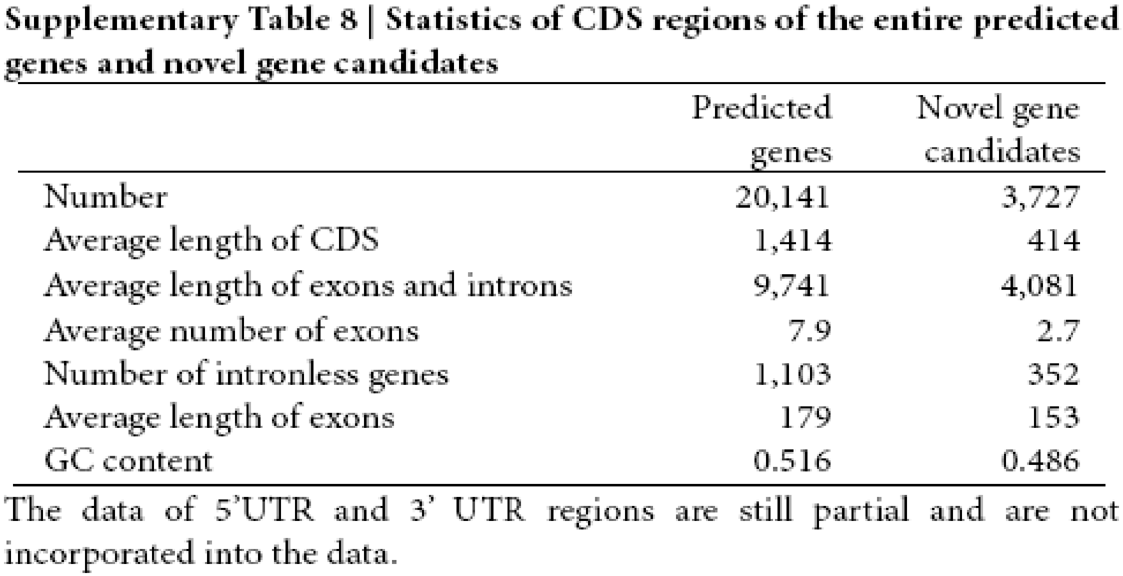
16,414 homologues, each found in at least one of six species, Unigene clusters (Aves, Amphibia, Actinopterygii, Ascidiacea), or *Takifugu* genome

15,565, each found in at least one of three fishes (*Tetraodon*, zebrafish, *Takifugu*)

12,621, each found in all three fishes (*Tetraodon*, zebrafish, *Takifugu*)

11,609, each found in all six species (*Tetraodon*, zebrafish, *Takifugu*, chicken, mouse, human)

Medaka genes

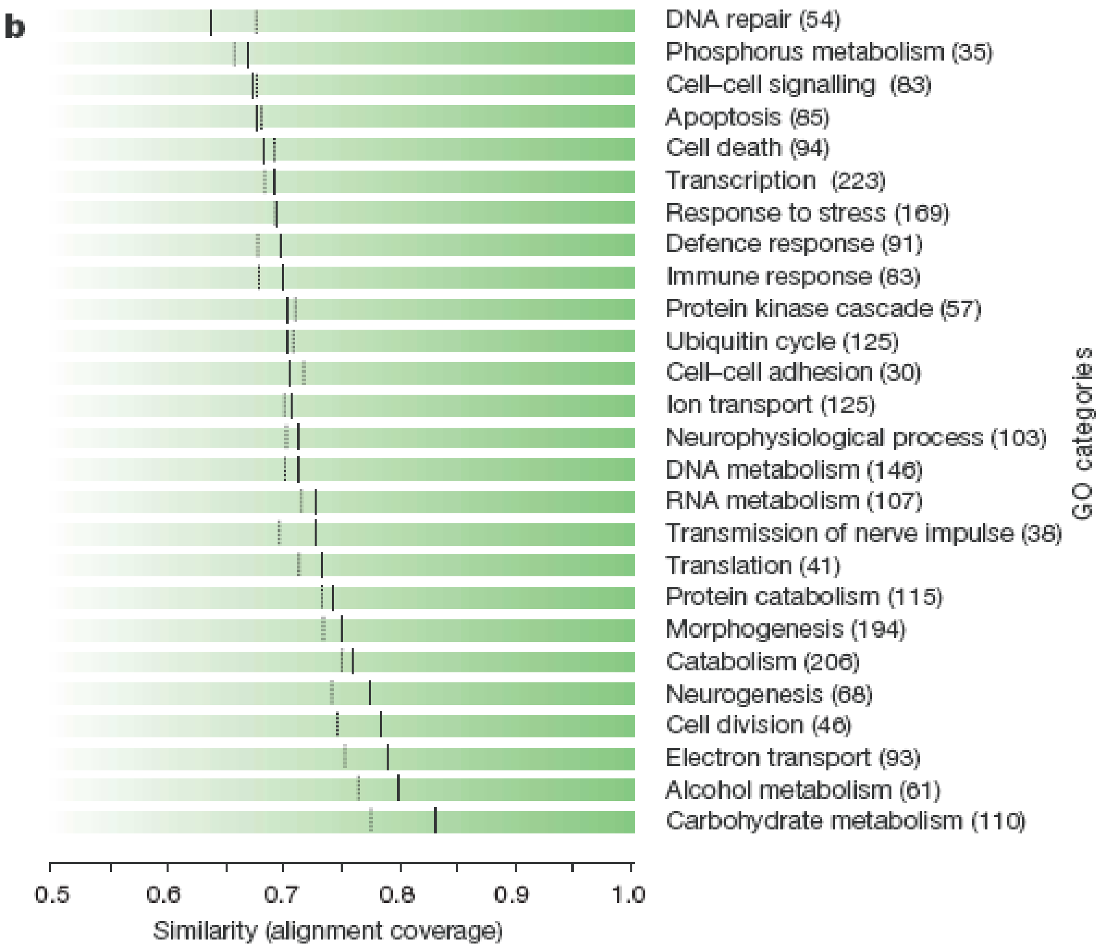


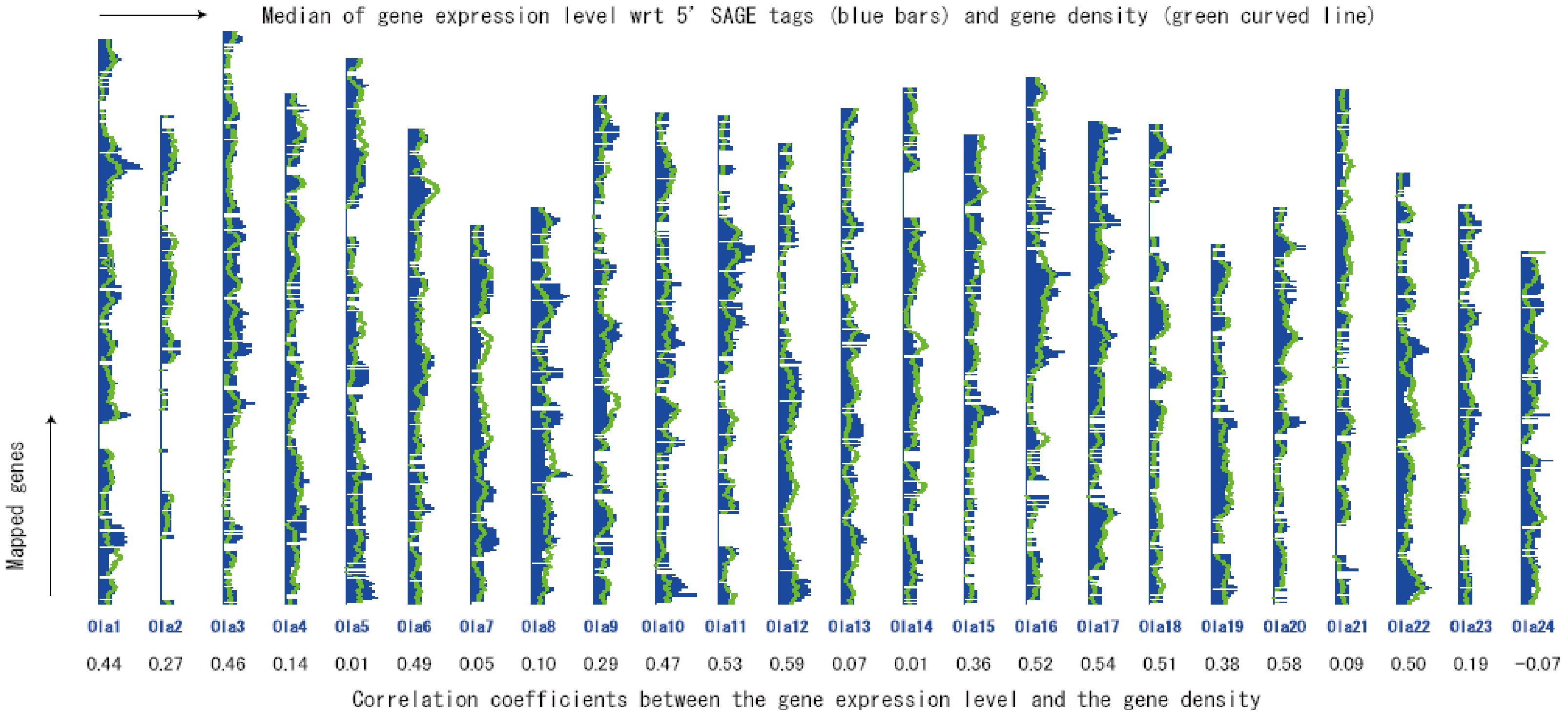
Supplementary Table 8 | Statistics of CDS regions of the entire predicted genes and novel gene candidates

	Predicted genes	Novel gene candidates
Number	20,141	3,727
Average length of CDS	1,414	414
Average length of exons and introns	9,741	4,081
Average number of exons	7.9	2.7
Number of intronless genes	1,103	352
Average length of exons	179	153
GC content	0.516	0.486

The data of 5'UTR and 3' UTR regions are still partial and are not incorporated into the data.

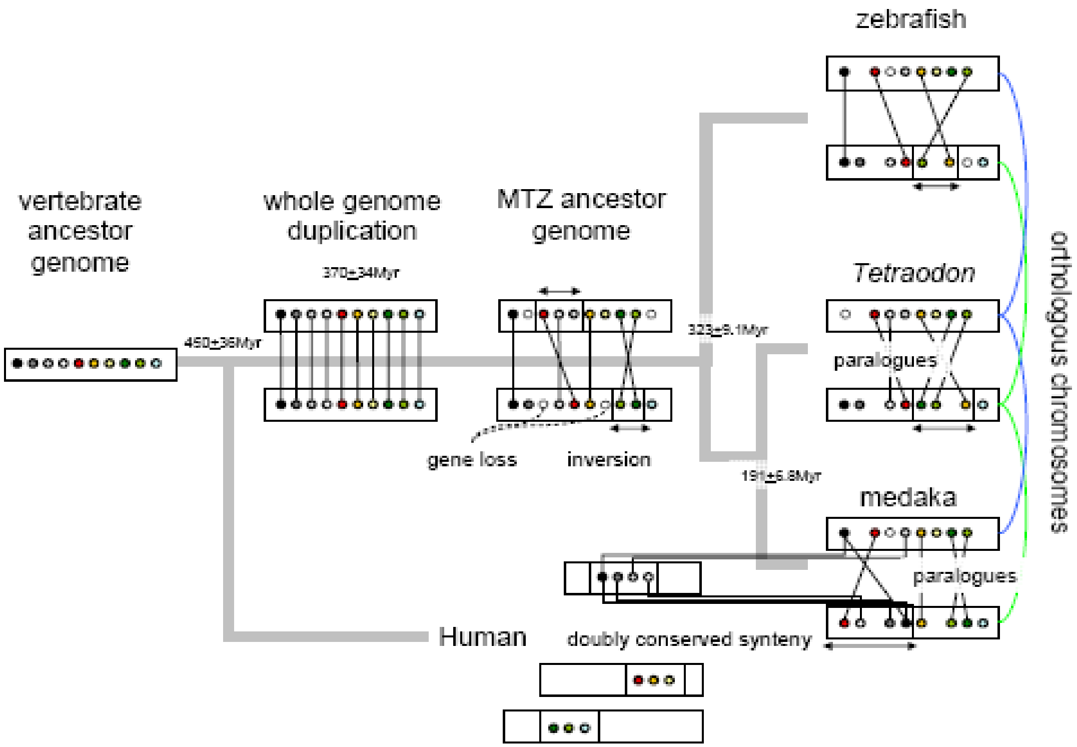
Medaka genes





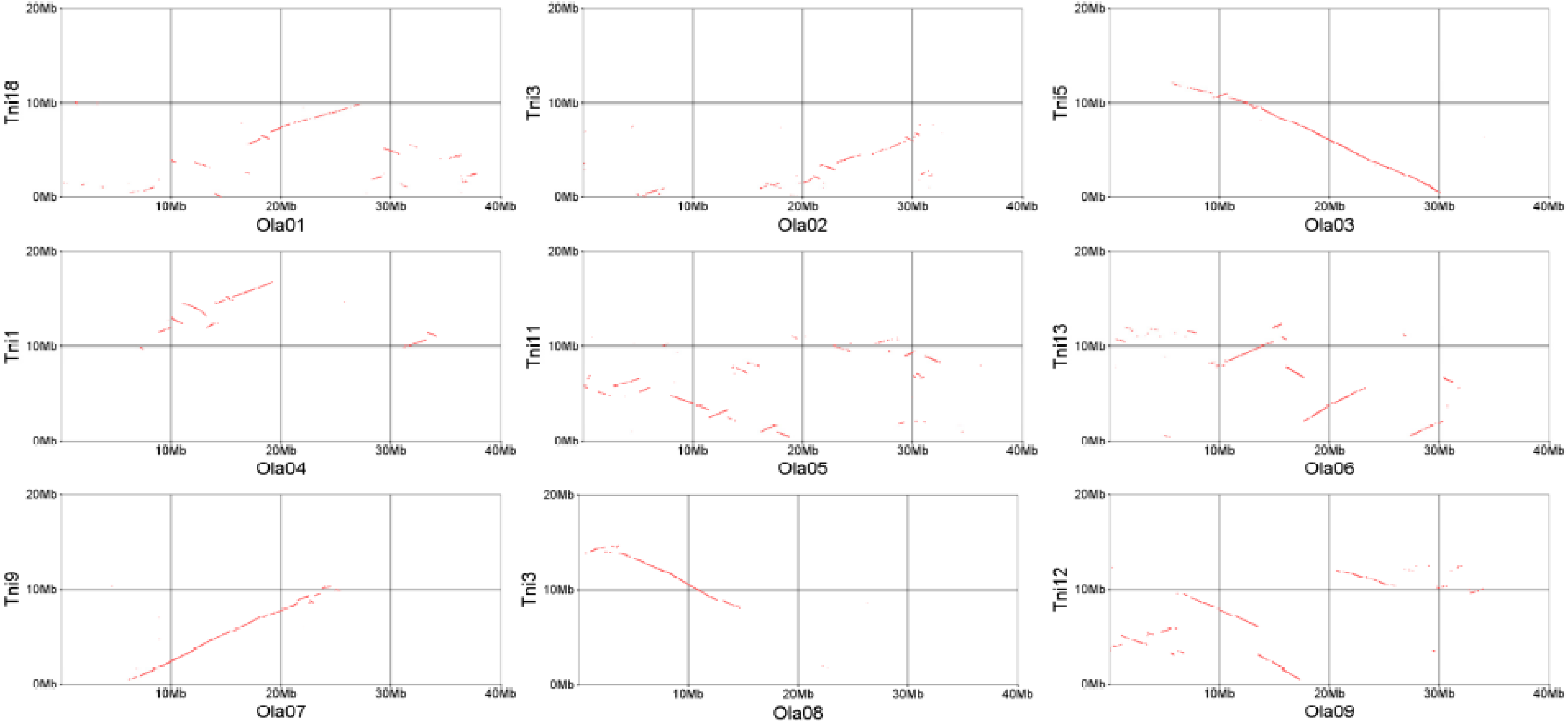
Supplementary Figure 7 | Medaka transcriptome map. We generated a transcriptome map that comprised 711,385 tags of 18,484 predicted medaka genes mapped to the medaka chromosomes. The vertical lines represent the medaka chromosomes. Each blue bar to the right represents the median of 5'SAGE tags occurring in each slide window of 39. The green curved line represents the gene density with a slide window size of 1Mb. The domains with highly or weakly expressed genes were scattered on the chromosomes.   Figures under chromosome numbers show correlation coefficients between the expression level and the gene density in individual chromosomes.

Genome evolution

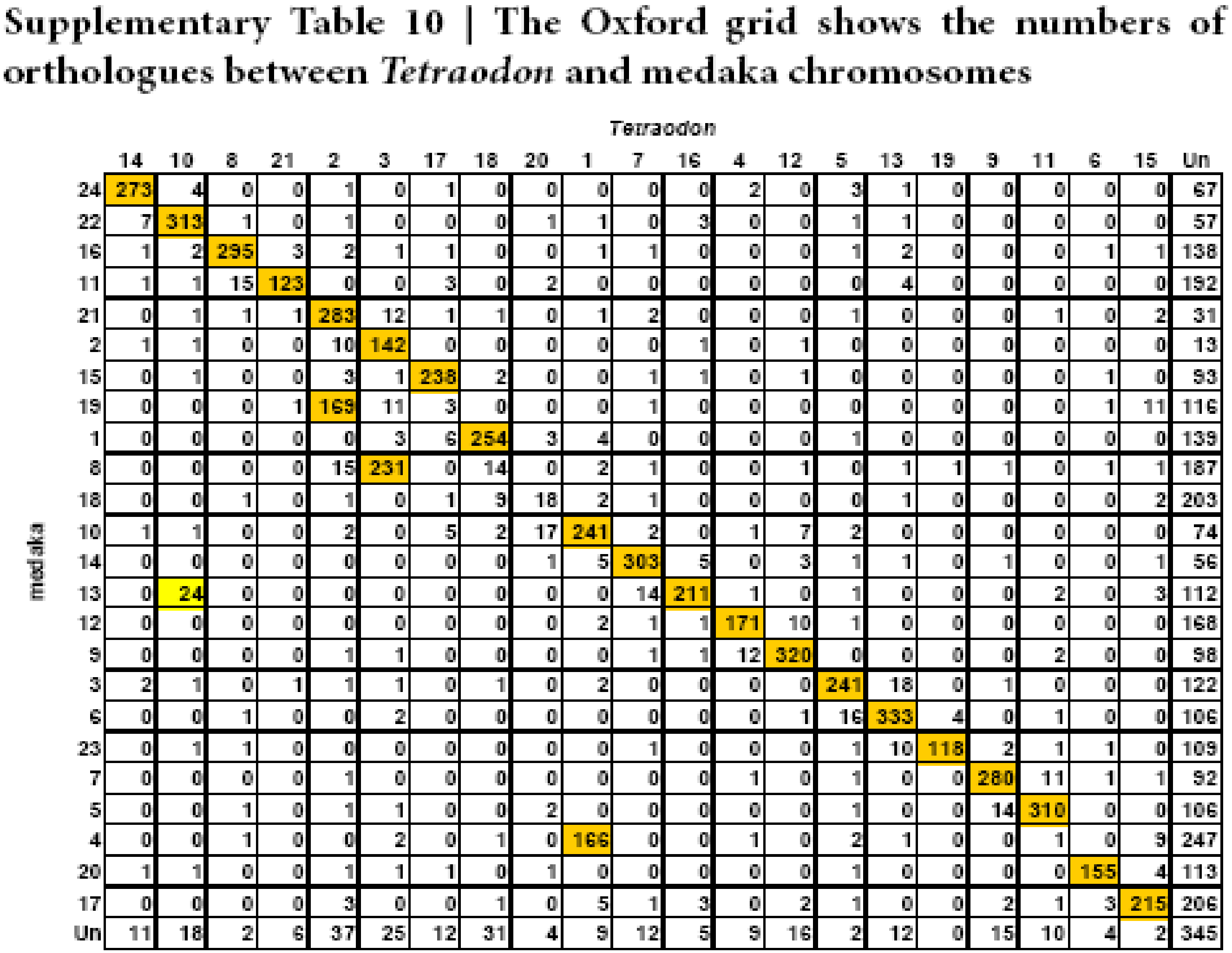


Supplementary Figure 11 | A model of the whole genome duplication and subsequent genome rearrangements in the teleost lineage.

Genome evolution



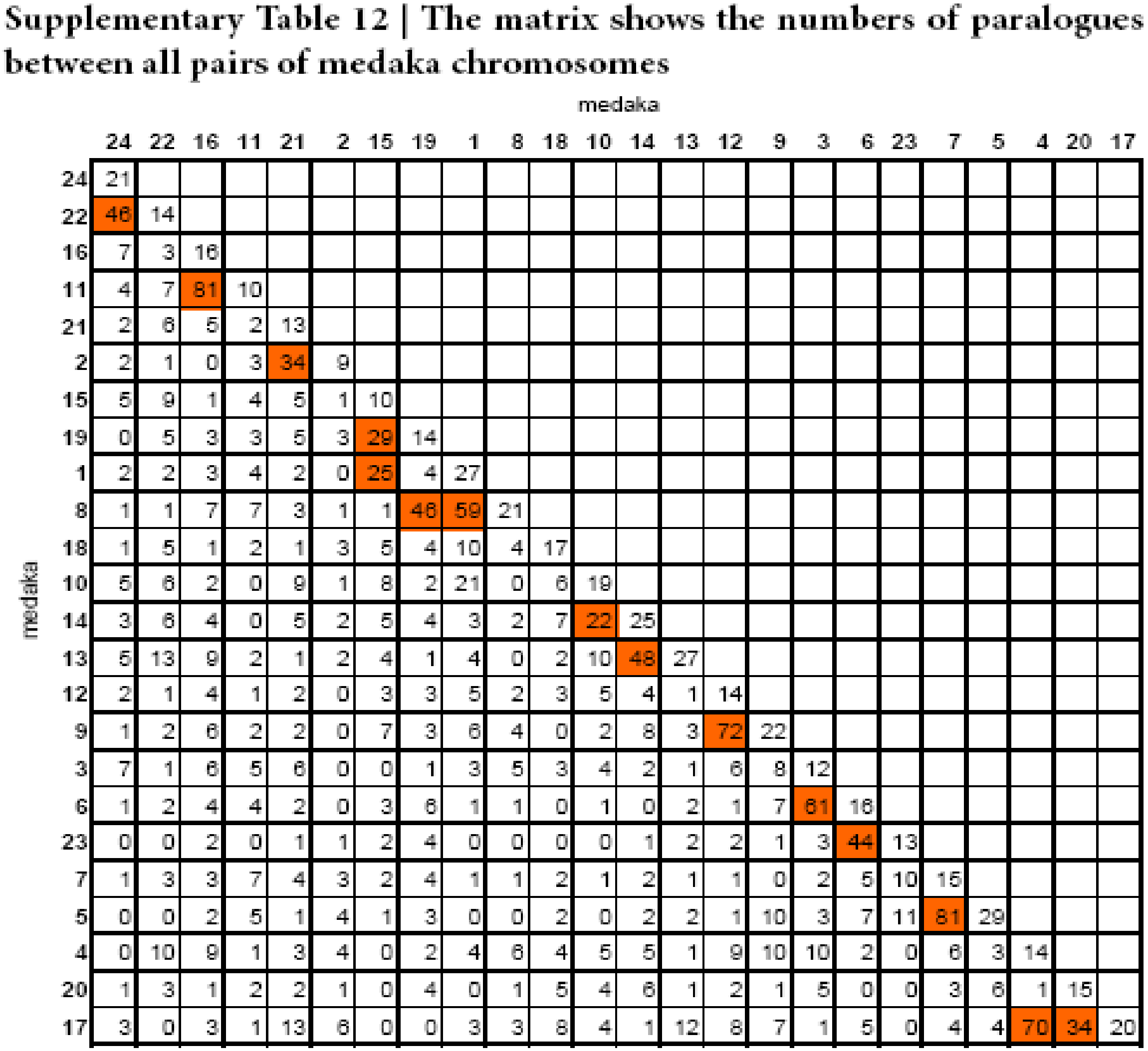
Genome evolution



Supplementary Table 10 | The Oxford grid shows the numbers of orthologues between *Tetraodon* and medaka chromosomes

												Tetraodon											
		14	10	8	21	2	3	17	18	20	1	7	16	4	12	5	13	19	9	11	6	15	Un
	24	273	4	0	0	1	0	1	0	0	0	0	0	2	0	3	1	0	0	0	0	0	67
	22	7	313	1	0	1	0	0	0	1	1	0	3	0	0	1	1	0	0	0	0	0	57
	16	1	2	295	3	2	1	1	0	0	1	1	0	0	0	1	2	0	0	0	1	1	138
	11	1	1	15	123	0	0	3	0	2	0	0	0	0	0	0	4	0	0	0	0	0	192
	21	0	1	1	1	283	12	1	1	0	1	2	0	0	0	1	0	0	0	1	0	2	31
	2	1	1	0	0	10	142	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	13
	15	0	1	0	0	3	1	238	2	0	0	1	1	0	1	0	0	0	0	0	1	0	93
	19	0	0	0	1	169	11	3	0	0	0	1	0	0	0	0	0	0	0	0	1	11	116
	1	0	0	0	0	0	3	6	254	3	4	0	0	0	0	1	0	0	0	0	0	0	139
	8	0	0	0	0	15	231	0	14	0	2	1	0	0	1	0	1	1	1	0	1	1	187
	18	0	0	1	0	1	0	1	9	18	2	1	0	0	0	0	1	0	0	0	0	2	203
medaka	10	1	1	0	0	2	0	5	2	17	241	2	0	1	7	2	0	0	0	0	0	0	74
	14	0	0	0	0	0	0	0	0	1	5	303	5	0	3	1	1	0	1	0	0	1	56
	13	0	24	0	0	0	0	0	0	0	0	14	211	1	0	1	0	0	0	2	0	3	112
	12	0	0	0	0	0	0	0	0	0	2	1	1	171	10	1	0	0	0	0	0	0	168
	9	0	0	0	0	1	1	0	0	0	0	1	1	12	320	0	0	0	0	2	0	0	98
	3	2	1	0	1	1	1	0	1	0	2	0	0	0	0	241	18	0	1	0	0	0	122
	6	0	0	1	0	0	2	0	0	0	0	0	0	0	1	16	333	4	0	1	0	0	106
	23	0	1	1	0	0	0	0	0	0	0	1	0	0	0	1	10	118	2	1	1	0	109
	7	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	280	11	1	1	92
	5	0	0	1	0	1	1	0	0	2	0	0	0	0	0	1	0	0	14	310	0	0	106
	4	0	0	1	0	0	2	0	1	0	166	0	0	1	0	2	1	0	0	1	0	9	247
	20	1	1	0	0	1	1	1	0	1	0	0	0	0	0	1	0	0	0	0	155	4	113
	17	0	0	0	0	3	0	0	1	0	5	1	3	0	2	1	0	0	2	1	3	215	206
	Un	11	18	2	6	37	25	12	31	4	9	12	5	9	16	2	12	0	15	10	4	2	345

Genome evolution

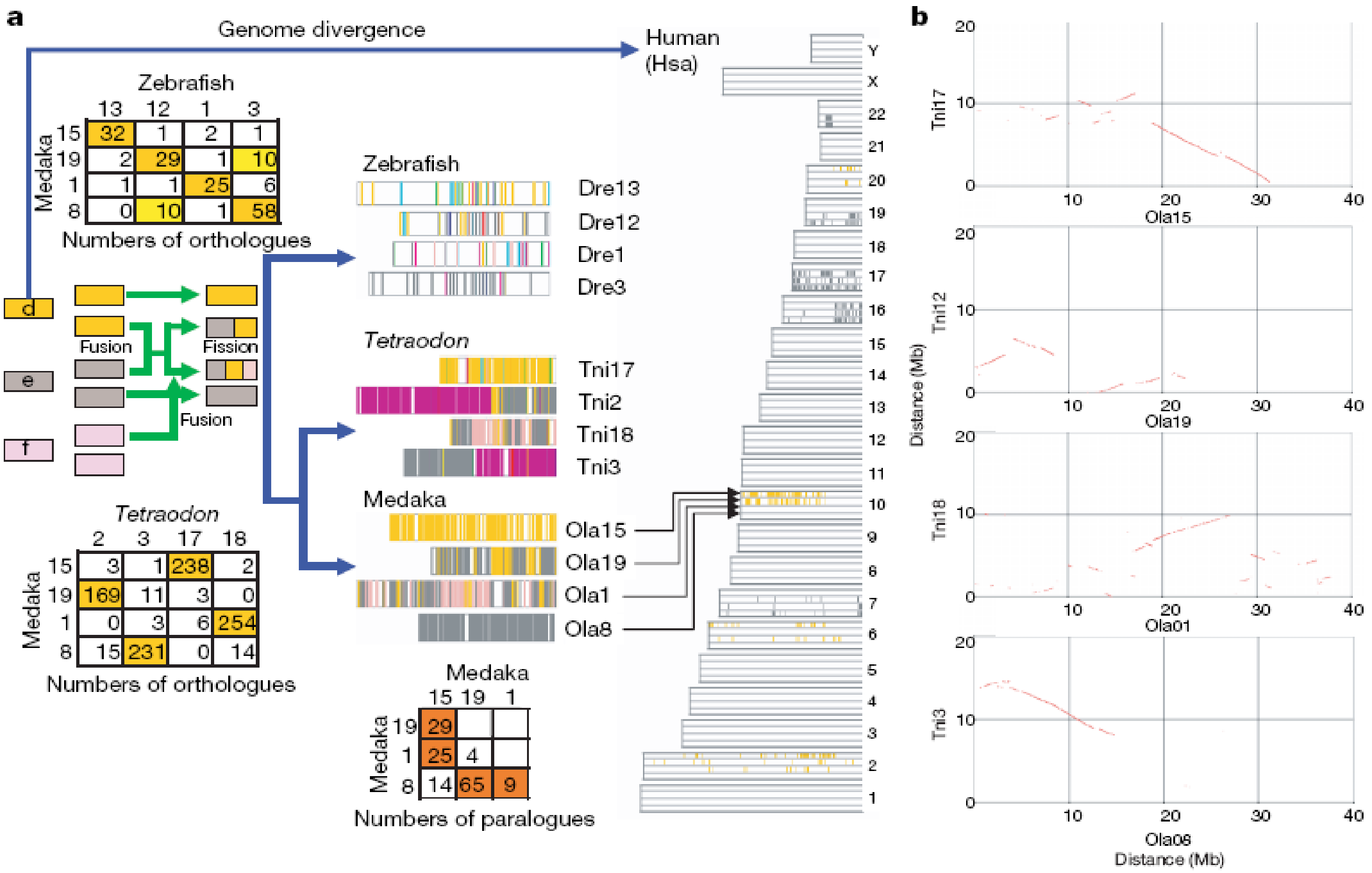


Supplementary Table 12 | The matrix shows the numbers of paralogues between all pairs of medaka chromosomes

medaka

medaka	24	22	16	11	21	2	15	19	1	8	18	10	14	13	12	9	3	6	23	7	5	4	20	17
24	21																							
22	46	14																						
16	7	3	16																					
11	4	7	81	10																				
21	2	6	5	2	13																			
2	2	1	0	3	34	9																		
15	5	9	1	4	5	1	10																	
19	0	5	3	3	5	3	29	14																
1	2	2	3	4	2	0	25	4	27															
8	1	1	7	7	3	1	1	46	59	21														
18	1	5	1	2	1	3	5	4	10	4	17													
10	5	6	2	0	9	1	8	2	21	0	6	19												
14	3	6	4	0	5	2	5	4	3	2	7	22	25											
13	5	13	9	2	1	2	4	1	4	0	2	10	48	27										
12	2	1	4	1	2	0	3	3	5	2	3	5	4	1	14									
9	1	2	6	2	2	0	7	3	6	4	0	2	8	3	72	22								
3	7	1	6	5	6	0	0	1	3	5	3	4	2	1	6	8	12							
6	1	2	4	4	2	0	3	6	1	1	0	1	0	2	1	7	61	16						
23	0	0	2	0	1	1	2	4	0	0	0	0	1	2	2	1	3	44	13					
7	1	3	3	7	4	3	2	4	1	1	2	1	2	1	1	0	2	5	10	15				
5	0	0	2	5	1	4	1	3	0	0	2	0	2	2	1	10	3	7	11	81	29			
4	0	10	9	1	3	4	0	2	4	6	4	5	5	1	9	10	10	2	0	6	3	14		
20	1	3	1	2	2	1	0	4	0	1	5	4	6	1	2	1	5	0	0	3	6	1	15	
17	3	0	3	1	13	6	0	0	3	3	8	4	1	12	8	7	1	5	0	4	4	70	34	20

Genome evolution



Zebrafish

Medaka	13	12	1	3
15	32	1	2	1
19	2	29	1	10
1	1	1	25	6
8	0	10	1	58

Numbers of orthologues

Tetraodon

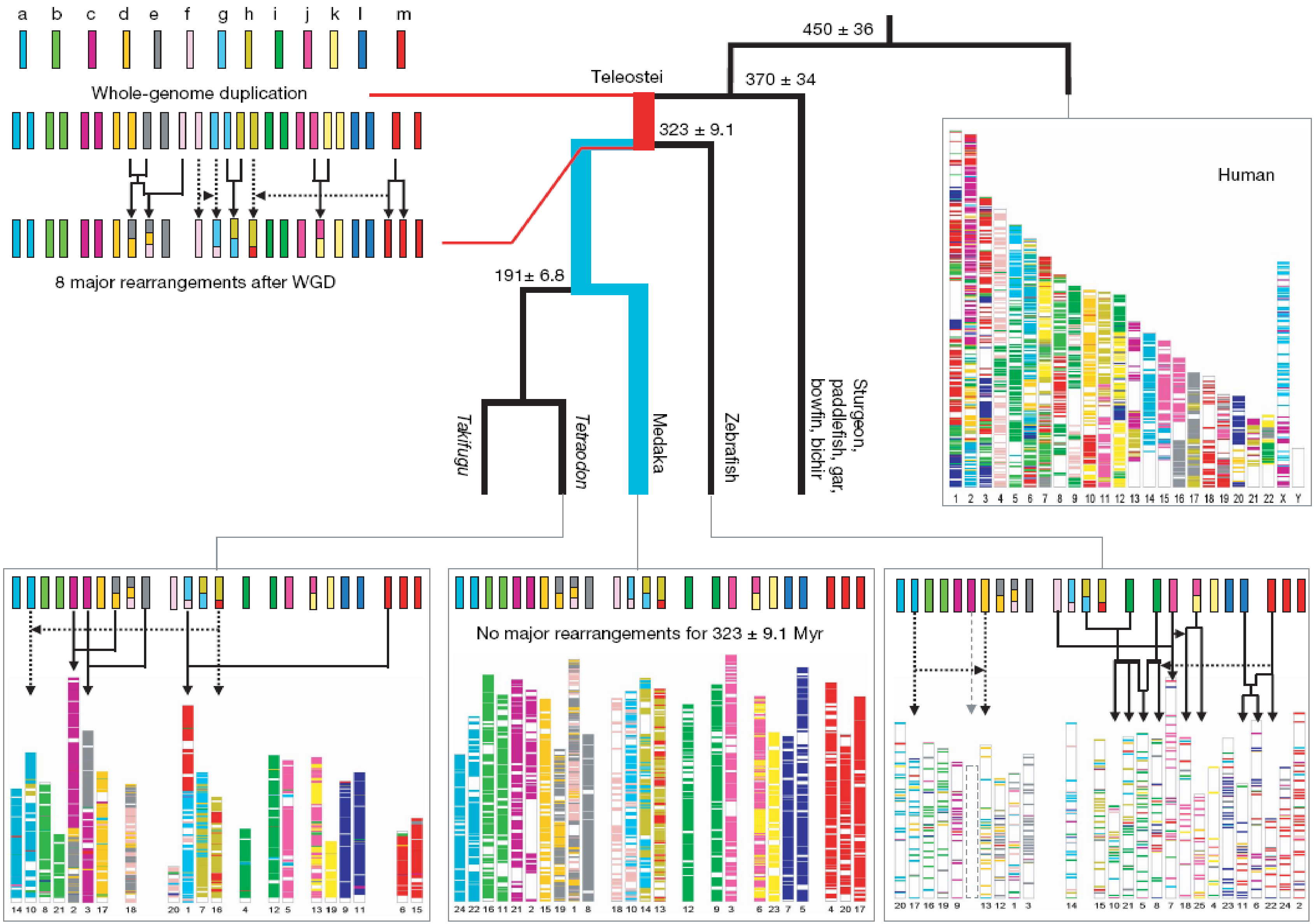
Medaka	2	3	17	18
15	3	1	238	2
19	169	11	3	0
1	0	3	6	254
8	15	231	0	14

Numbers of orthologues

Medaka

Medaka	15	19	1
19	29		
1	25	4	
8	14	65	9

Numbers of paralogues


a b c d e f g h i j k l m
Whole-genome duplication
8 major rearrangements after WGD
Teleostei
450 ± 36
370 ± 34
323 ± 9.1
191± 6.8
Takifugu
Tetraodon
Medaka
Zebrafish
Sturgeon, paddlefish, gar, bowfin, bichir
Human
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
14 10 8 21 2 3 17 18 20 1 7 16 4 12 5 13 19 9 11 6 15
No major rearrangements for 323 ± 9.1 Myr
24 22 16 11 21 2 15 19 1 8 18 10 14 13 12 9 3 6 23 7 5 4 20 17
20 17 16 19 9 13 12 1 3 14 15 10 21 5 8 7 18 25 4 23 11 6 22 24 2