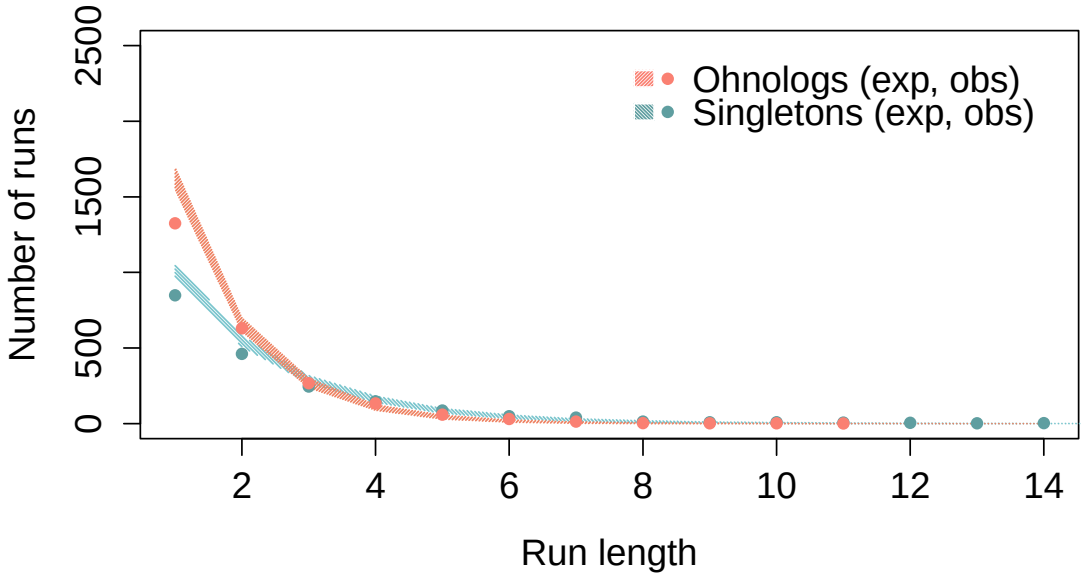
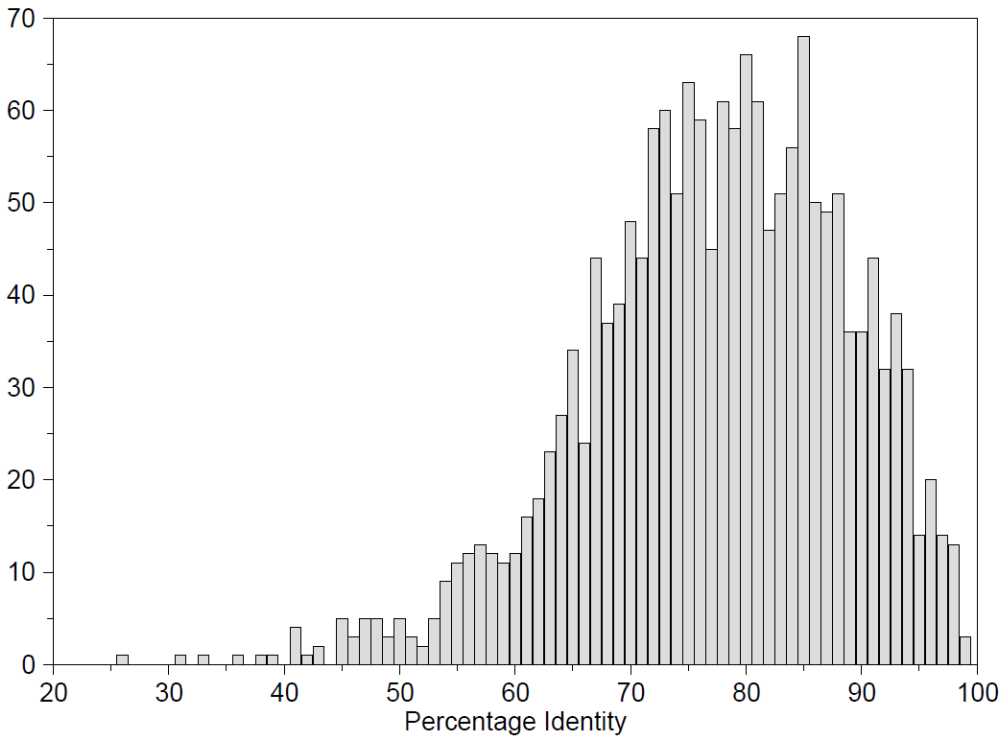
SUPPLEMENTARY INFORMATION

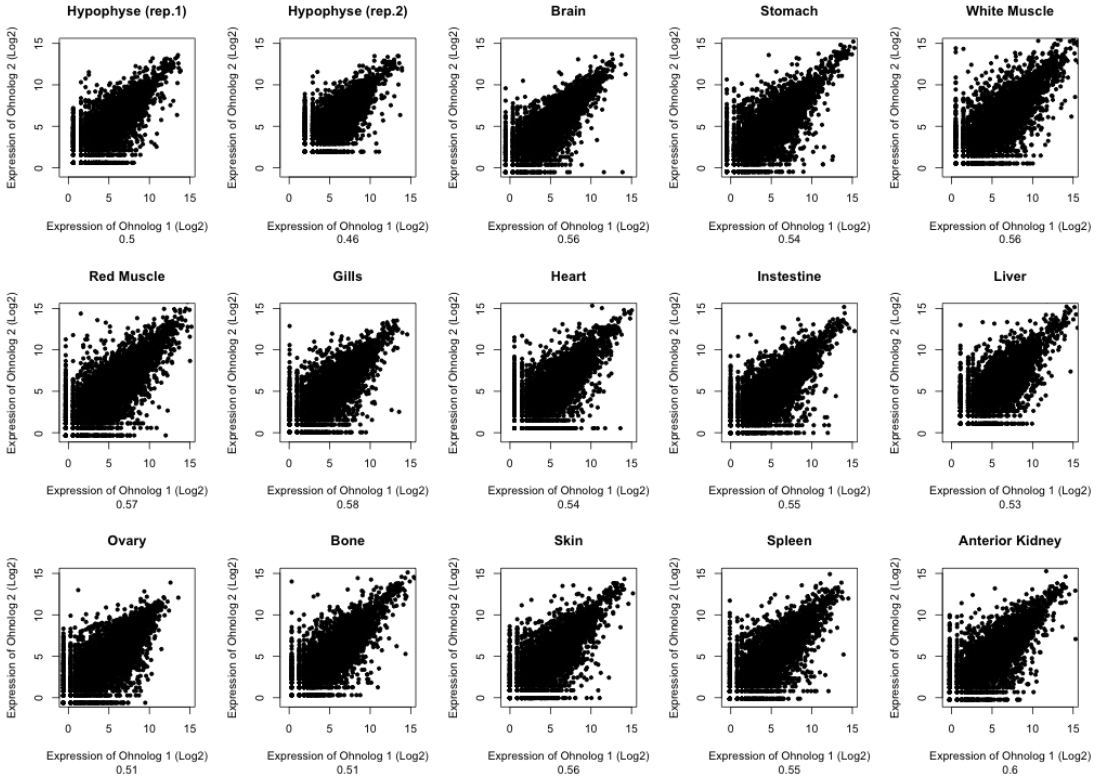
Supplementary Figures



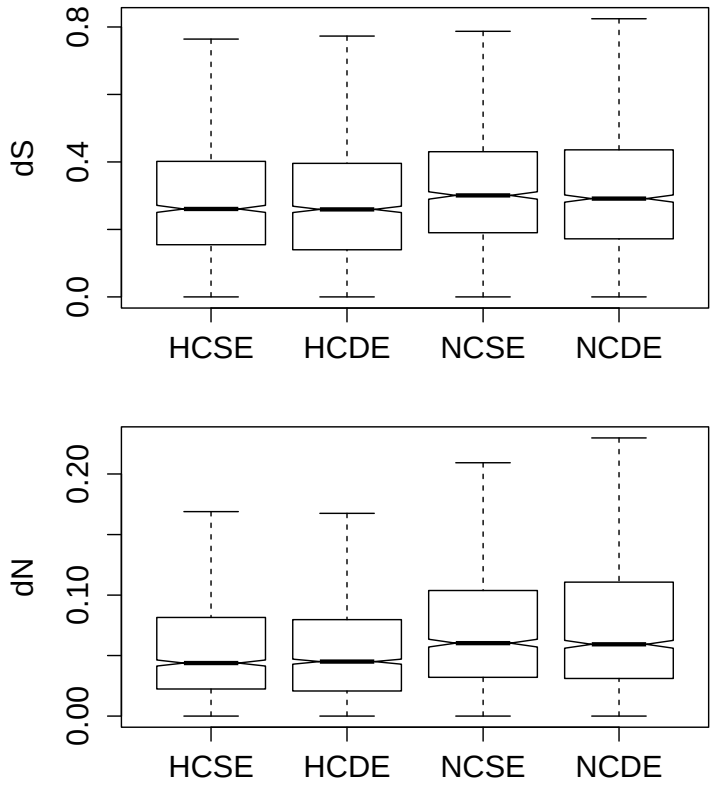
Supplementary Figure 1: Random alternation of ohnologs and singletons in the trout genome. Number of expected and observed runs of successive ohnologs and singletons in DCS regions. We found no evidence of any regular patterning in the alternation of ohnolog and singleton genes in the trout genome. The small differences between expectations and observations are most likely an artifact due to split gene structures, where different exons of the same gene are wrongly annotated as different genes, as these would appear as several successive ohnologs or singletons instead of only one gene. Together with the high retention of collinearity between duplicated scaffolds, this result suggests that gene fractionation mostly occurs through random mutations and sequence decay, rather than through large-scale deletions and rearrangements.



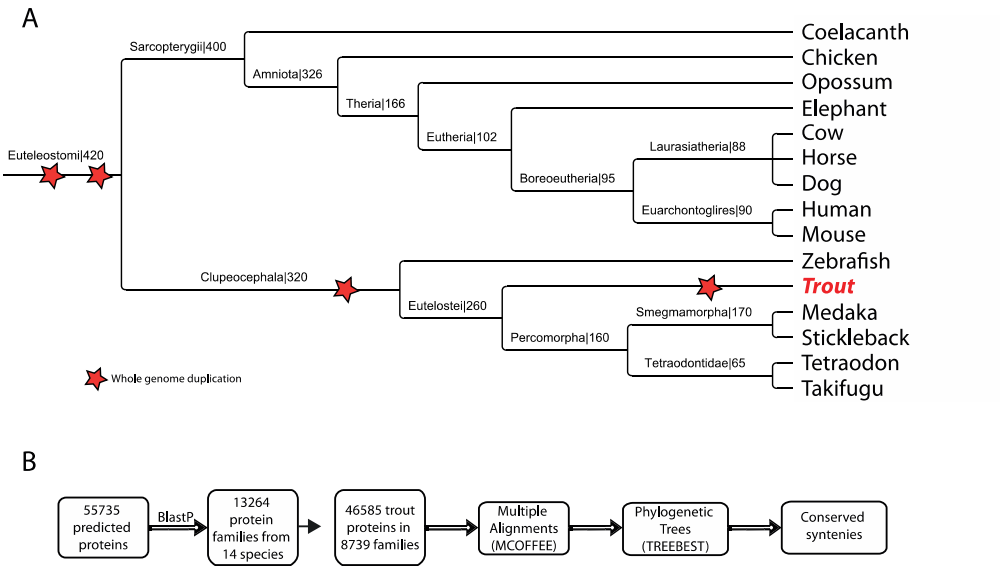
Supplementary Figure 2: Distribution of percentage of identity measures between singleton genes and their corresponding pseudogenes. To better understand the fate of inactivated gene copies, protein sequence predicted from a given gene model were aligned to their paralogous region showing that the identity between the Ss4R protein-coding singletons and their corresponding pseudogenes remains high (average amino acid identity 79.0%, SD = 5.5%).



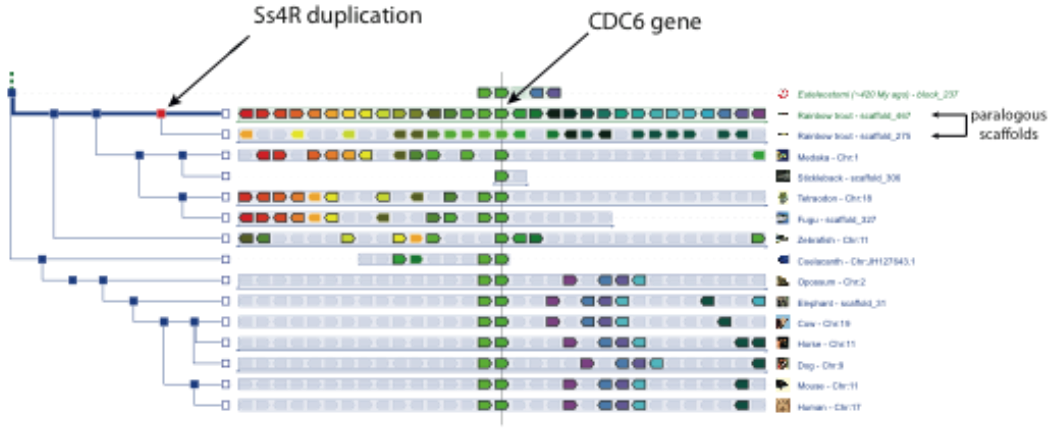
Supplementary Figure 3: Correlation of the expression of ohnolog genes across different tissues. All expressions are normalized using DESeq[1] and are represented in a log2 scale (+0.01 to avoid null values). Pearson's determination coefficients (R^2) are reported underneath for each tissue and have been calculated on 6,123 ohnolog pairs for which sequencing reads could be confidently attributed between ohnologs.



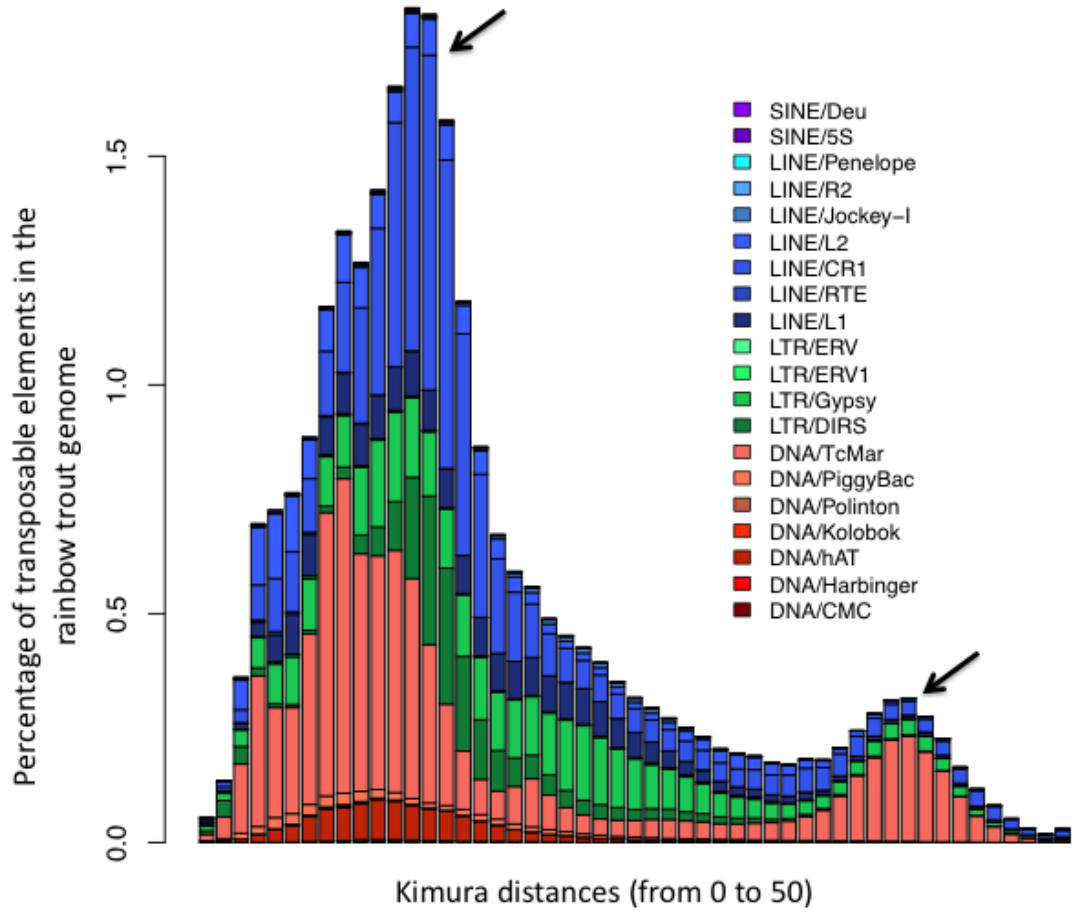
Supplementary Figure 4: Sequence evolution in each category of ohnologs. Whisker plots (with whiskers representing the range of the distribution, excluding the 5% most extremes values) showing that ohnologs with highly correlated expression levels (HC) display on average both lower dS (rates of substitutions per synonymous site) and lower dN (rates of substitutions per non-synonymous site) than ohnologs with uncorrelated expression levels (NC). This suggests that correlated ohnologs are overall subjected to lower mutational rates in addition to higher selective pressure, as evidenced by their lower dN/dS ratios (Figure 4, main text). Variations in main expression levels (same expression levels: SE; different average expression levels: DE), however, are not related to differences in mutational rates within the HC and NC groups. Numbers of ohnolog pairs in each group: HCSE=1,407; HCDE=1,895; NCSE=1,248; NCDE=1,573.



Supplementary Figure 5: Integration of Trout predicted protein in vertebrate families. (A) A cladogram of 14 selected vertebrate species used in the analysis to identify membership of trout proteins. The tree includes representative species of Amniotes (Euarchontoglires, Laurasiatheria, Metatheria, Sauria) and fish (Sarcopterygian and Actinopterygians). Stars represent whole genome duplications. Branches are not to scale. Branch labels indicate the clade name followed by the divergence time. (B) Schematics of the bioinformatics pipeline used to integrate trout predicted proteins in vertebrate gene families and to reconstruct phylogenetic trees.



Supplementary Figure 6: The Genomicus server. A dedicated Genomicus server was developed to provide access to the comparative genomics and phylogenetic tree reconstructions involving trout genes. The address (http://www.dyogen.ens.fr/genomicus-trout-01.01/) provides a search form where gene names may be entered. The figure shows an example with the Cell Division Cycle 6 homolog (CDC6) gene in the center in light green, surrounded by adjacent genes on scaffold_467. The CDC6 gene exists in all the other species except chicken, and the local gene content is preserved 5' to CDC6 in other teleost fish (red/yellow), but 3' to CDC6 in Amniotes (blue/purple).



Supplementary Figure 7: Transposable element history in the rainbow trout genome based on Kimura distances. Divergence of TE superfamily copies were evaluated using Kimura distances, highlighting bursts of transposition in the rainbow trout genome. Y-axis represents the content of a given TE superfamily in the genome and X-axis indicates its presence at a given distance (divergence from the repeat consensus). Black arrows show the two main bursts of transposition that occurred in the rainbow trout genome.

Supplementary Tables

Supplementary Table 1: Summary of the evolutionary history of the duplicated trout genome. Summary of (i) the ancestral linkage groups before the Ts3R WGD event (column 1). (ii) their descending linkage groups before the Ss4R event (column 2). (iii) the resulting ohnologous regions on modern trout chromosomes (columns 3 and 4; chromosomes in column 3 share regions of Ss4R paralogy with all the chromosomes in column 4). Pre-Ss4R linkage groups were considered as Ts3R-ohnologous if they share at least 5 pairs of Ts3R ohnologs (as deduced by orthology with medaka Ts3R ohnologs). While each pre-Ts3R linkage group resulted in two linkage groups just after the WGD event, rearrangements may have occurred between the Ts3R and Ss4R events, so that more than two pre-Ss4R groups can contain regions descending from a single pre-Ts3R group. Large ohnologous regions between trout chromosomes were defined as regions sharing at least 20 pairs of Ss4R ohnologs. Again, because of rearrangements since the Ss4R WGD, each pre-Ss4R linkage group can be split on two or more chromosomes (and most modern chromosomes are mosaics of regions originating from different pre-Ss4R linkage groups). Of note, some pre-Ss4R linkage groups possibly correspond to single ancestral chromosomes (for example, II.b and II.c, which both correspond to regions of modern chromosome 3 and additional ohnologous regions on chromosomes 22 and 2) but evidence was insufficient to regroup them in the ancestral genome reconstruction process (see Methods). Conversely, many pre-Ss4R could not be paired as descending from the same pre-Ts3R linkage group due to inconclusive evidence. These groups were reported as different ancestral linkage groups (IX to XIX, the second post-Ts3R group being shown as NA: Not Assigned), although it is likely that the ancestral genome before the Ts3R contained less than 19 chromosomes (estimated 13)[2] and several of these groups are actually Ts3R duplicates.

Before Ts3R WGD	Before Ss4R WGD	Chromosomes sharing Ss4R ohnologous regions	
I	I.a	8	28
	I.b	5	12, Sex
	I.c	4	5, 27
		5	1, 4
	I.d	6	11
	I.e	11	15
II	II.a	18	7, 14
	II.b	3	22
	II.c	3	2
III	III.a	7	17
	III.b	9	16
	III.c	9	15
IV	IV.a	2	1, 4
	IV.b	6	26
	IV.c	15	21
V	V.a	5	23
	V.b	16	1, 20
	V.c	12	13
VI	VI.a	27	8, 24
	VI.b	10	12
VII	VII.a	19	29
	VII.b	4	8

VIII	VIII.a	1	23
	VIII.b	20	2
IX	IX.a	13	17
	NA	NA	NA
X	X.a	20	23
	NA	NA	NA
XI	XI.a	9	21
	NA	NA	NA
XII	XII.a	14	29
	NA	NA	NA
XIII	XIII.a	8	6
	NA	NA	NA
XIV	XIV.a	8	24
	NA	NA	NA
XV	XV.a	10	19
	NA	NA	NA
XVI	XVI.a	8	19
	NA	NA	NA
XVII	XVII.a	10	Sex
	NA	NA	NA
XVIII	XVIII.a	4	12
	NA	NA	NA
XIX	XIX.a	11	22
	NA	NA	NA

Supplementary Table 2: Comparison of the number of miRNA loci between rainbow trout and other vertebrates. Numbers of mature miRNAs present in miRBase and their corresponding loci in metazoans. The locus per miRNA ratio (Locus/miRNA) is indicated. Interestingly, this ratio ranged between 1.22 and 1.45 in teleosts subjected to Ts3R, between 1.06 and 1.22 in tetrapods, and between 1.02 and 1.05 in non-vertebrate Metazoans.

Species	WGD	Loci	Mature miRNAs	Loci/miRNA
O. mykiss	**Ss4R**	**495**	**164**	**3.02**
D. rerio	Ts3R	344	247	1.39
O. latipes	Ts3R	168	116	1.45
F. rubipres	Ts3R	129	106	1.22
X. tropicalis	2R	189	154	1.22
G. gallus	2R	684	644	1.06
M. musculus	2R	855	781	1.09
C. intestinalis		348	333	1.05
D. melanogaster		238	231	1.03
C. elegans		223	218	1.02

Supplementary Table 3: Functional enrichments of gene families that have been systematically retained as ohnologs after vertebrate WGD events: in human after the 1R-2R WGD events, in zebrafish after the Ts3R event, and in rainbow trout after the Ss4R event. The gene content of the vertebrate ancestor (Euteleostomi) was reconstructed using Ensembl Compara gene phylogenies, and ontology annotations of human genes were transposed to their ancestral counterparts. Ancestral vertebrate genes that were retained as 1R-2R ohnologs in the human genome[3], as well as Ts3R ohnologs in zebrafish[4] and Ss4R ohnologs in trout were compared to the remainder of the ancestral gene set for functional enrichments, using a random sampling procedure (10,000 iterations, resulting in an empirical p-value corrected for multiple tests using Benjamini-Hochberg false discovery rate correction with an FDR of 0.1).

GENE ONTOLOGY TERM	P-value
Biological process	
synaptic transmission	7.0e-10
regulation of cell proliferation	4.4e-08
axon guidance	5.5e-07
nervous system development	1.4e-06
filopodium assembly	2.3e-06
insulin receptor signaling pathway	7.7e-06
positive regulation of transcription, DNA-dependent	1.1e-05
cell-cell adhesion	1.8e-05
organ morphogenesis	2.2e-05
neuron projection development	3.6e-05
transmission of nerve impulse	4.5e-05
patterning of blood vessels	0.0001
Molecular function	
protein binding	1.0e-18
sequence-specific DNA binding	1.6e-09
receptor activity	5.0e-09
actin binding	4.3e-08
sequence-specific DNA binding transcription factor activity	1.4e-07
calcium ion binding	1.0e-06
protein tyrosine kinase activity	1.3e-06
cytoskeletal adaptor activity	9.2e-06
signal transducer activity	2.0e-05
guanyl nucleotide binding	2.5e-05
protein serine/threonine kinase activity	3.0e-05
protein heterodimerization activity	3.6e-05
protein kinase activity	4.3e-05
double-stranded DNA binding	0.0001
insulin-like growth factor binding	0.0002
lipid phosphatase activity	0.0004
Cellular component	
plasma membrane	2.8e-24
cell junction	4.9e-10
integral to plasma membrane	4.4e-08
synapse	6.8e-07
cell-cell junction	4.7e-06
ruffle	4.7e-06
cell surface	1.2e-05
lateral plasma membrane	3.9e-05
dendrite	4.7e-05
basolateral plasma membrane	4.7e-05
presynaptic membrane	8.1e-05

integral to membrane	9.3e-05
cell-cell adherens junction	9.5e-05
actin cytoskeleton	0.0001
desmosome	0.0001
neuron projection	0.0002
protein complex	0.0002
membrane fraction	0.0002
myosin complex	0.0002
melanosome	0.0002
axon	0.0003
synaptic vesicle membrane	0.0003
synaptic vesicle	0.0003
cell cortex	0.0004
cell projection	0.0007
neuronal cell body	0.0008
apical part of cell	0.001
unconventional myosin complex	0.001
dendrite membrane	0.001
I band	0.001
muscle myosin complex	0.001
postsynaptic membrane	0.002

Supplementary Table 4. Functional enrichments for the four categories of ohnologs. We used orthology relationships to transfer human gene ontology annotations to their orthologs in the trout genome. Each set of ohnologs, based on their expression characteristics (highly correlated or non-correlated expression between the pair: HC/NC; same expression levels or different expression levels between the pair: SE/DE), were tested for functional enrichments compared to the remainder of the set of ohnolog genes using a random sampling procedure (10,000 iterations, resulting in an empirical p-value corrected for multiple tests using Benjamini-Hochberg false discovery rate correction with an FDR of 0.1). The exact enrichment p-value for each empirically significant ontology term was then calculated using Fisher's exact test.

HCSE		**HCDE**	
GENE ONTOLOGY TERM	**P-value**	**GENE ONTOLOGY TERM**	**P-value**
Biological process		**Biological process**	
regulation of transcription, DNA-dependent	1.7e-13	translational termination	3.6e-29
positive regulation of transcription from RNA polymerase II promoter	4.5e-10	viral transcription	3.6e-29
multicellular organismal development	4.6e-09	viral infectious cycle	2.1e-28
positive regulation of neuron differentiation	8.4e-09	translational elongation	7.8e-27
pattern specification process	3.8e-08	mRNA metabolic process	9.4e-27
anterior/posterior pattern specification	5.3e-08	RNA metabolic process	4.7e-23
regulation of sequence-specific DNA binding transcription factor activity	9.5e-08	cellular protein metabolic process	5.3e-23
neuron differentiation	1.3e-07	translation	7.0e-23
positive regulation of transcription, DNA-dependent	1.8e-07	viral reproduction	4.4e-18
spinal cord association neuron differentiation	4.0e-07	endocrine pancreas development	9.6e-16
skeletal muscle tissue development	6.8e-07	respiratory electron transport chain	3.6e-13
neurogenesis	1.2e-06	rRNA processing	2.2e-12
forebrain development	3.6e-06	gene expression	3.3e-12
cell differentiation	6.2e-06	M/G1 transition of mitotic cell cycle	6.6e-11
embryonic skeletal system morphogenesis	7.2e-06	G1/S transition of mitotic cell cycle	1.1e-09
response to amphetamine	7.7e-06	cell cycle checkpoint	3.9e-08
proximal/distal pattern formation	7.7e-06	S phase of mitotic cell cycle	6.2e-08
enteric nervous system development	7.7e-06	antigen processing and presentation of peptide antigen via MHC class I	3.8e-07
locomotory behavior	1.4e-05	interspecies interaction between organisms	1.7e-06
homophilic cell adhesion	1.5e-05	mitotic cell cycle	1.7e-06
tissue development	1.5e-05	generation of precursor metabolites and energy	1.8e-06
learning or memory	1.6e-05	protein folding	3.7e-06
pancreas development	1.6e-05	protein complex assembly	3.7e-06
positive regulation of branching involved in ureteric bud morphogenesis	1.6e-05	regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	6.1e-06
axon guidance	2.1e-05	DNA-dependent DNA replication initiation	7.9e-06
organ morphogenesis	4.1e-05	ribosomal small subunit biogenesis	7.9e-06
central nervous system development	4.8e-05	mitochondrial electron transport, NADH to ubiquinone	1.1e-05
inner ear morphogenesis	4.8e-05	anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process	2.0e-05
acetyl-CoA metabolic process	0.0001	positive regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	2.2e-05

dorsal/ventral axis specification	0.0001	regulation of establishment of cell polarity	8.3e-05
peripheral nervous system neuron development	0.0001	nuclear mRNA 3-splice site recognition	8.3e-05
type B pancreatic cell differentiation	0.0001	endoplasmic reticulum organization	8.3e-05
forebrain neuron differentiation	0.0001	ribosomal large subunit biogenesis	8.3e-05
motor axon guidance	0.0002	protein N-linked glycosylation via asparagine	0.0001
dorsal/ventral pattern formation	0.0002		
negative regulation of BMP signaling pathway	0.0002		
negative regulation of protein binding	0.0002		
ureteric bud development	0.0002		
negative regulation of transcription from RNA polymerase II promoter	0.0002		
muscle cell differentiation	0.0002		
cell-cell adhesion	0.0004		
neuron migration	0.0004		
integrin-mediated signaling pathway	0.0008		
embryo development	0.002		
nervous system development	0.003		
regulation of transcription from RNA polymerase II promoter	0.005		
NCSE		**NCDE**	
GENE ONTOLOGY TERM	**P-value**	**GENE ONTOLOGY TERM**	**P-value**
visual perception	1.8e-10	DNA repair	5.6e-05
positive regulation of myoblast differentiation	2.9e-06		
lens development in camera-type eye	5.3e-06		
response to organic substance	5.5e-06		
camera-type eye development	5.5e-06		
response to stimulus	2.7e-05		
activation of phospholipase C activity by G-protein coupled receptor protein signaling pathway coupled to IP3 second messenger	7.1e-05		
long-term memory	7.1e-05		
response to pH	7.1e-05		
cellular response to growth factor stimulus	0.0001		
synaptic transmission	0.0005		
germ cell development	0.001		
positive regulation of osteoblast differentiation	0.001		

Supplementary Table 5. Raw sequencing data overview. Coverage estimations are based on a genome size of 2.4 Gb.

Sequencer type	Library type	Number of reads (millions)	Number of bases (Mb)	Genome Coverage in Sequenced Bases	Genome Coverage in library insert	Insert size (Kb)
Sanger	Bac Ends	0.2	149.1	0.06X	4.7X	118<>137
Illumina	Single end	427	39660	16.5X	NA	NA
	Pair End	898	89133	37.1X	55X	0.3 - 0.5
Roche/454	Single End	86.4	26017	10.8X	NA	NA
	Long Single End	39.0	14883	6.2X	NA	NA
	Mate Paire	7.9	2343	1X	10.7X	6.5
		6.4	2179	0.9X	16.7X	12.5

Supplementary Table 6. Rainbow trout genome assembly overview. *Large contigs are contigs longer than 500bp.

	Number	Cumulative size (Mb)	Average size (Kb)	N50 (Kb and number)	N80 (Kb and number)	N90 (Kb and number)	Longest (Kb)
Large contigs *	445 600	1 684.6	3.8	7.7 57 637	2.6 169 086	1.5 254 103	118.4
Scaffolds	79 941	1 877.5	23.4	383.6 1 014	28.2 8 202	7.6 2 120	5 466.1

Supplementary Table 7. Rainbow trout protein-coding gene content and comparison with other vertebrates.

	Predicted genes	Genes in gene trees	Singletons	# of trees	Average number of genes per trees	Maximum number of genes in tree
Human	21860	20108	1752	9453	2,13	42
Mouse	23083	22226	857	8991	2,47	169
Cow	19994	19892	102	8680	2,29	54
Dog	19305	19156	149	8653	2,21	71
Horse	20436	20224	212	8780	2,30	114
Elephant	20033	19938	95	8564	2,33	100
Chicken	16736	15115	1621	7639	1,98	162
Coelacanth	19033	18902	131	7686	2,46	111
Opossum	19466	19169	297	8227	2,33	133
Zebrafish	26160	25618	542	8300	3,09	266
Stickleback	20787	20092	695	8006	2,51	70
Medaka	19686	18887	799	7562	2,50	70
Fugu	18523	18441	82	7131	2,59	38
Tetraodon	19602	19285	317	7306	2,64	24
Rainbow trout	46585	46585	-	8739	5,33	504

Supplementary Table 8. Anchorage statistics of the rainbow trout genome assembly on chromosomes.

Chromosome anchorage	Number	Cumulative size (Mb) and % of the assembly	Average size (Kb)	N50 (Kb and number)	N80 (Kb and number)	N90 (Kb and number)	Longest (Mb)
Scaffolds and remaining contigs anchored	4 413	1 023.3 48%	231.9	935.6 326	402.3 315	186.1 1178	5.5
Anchored scaffolds	4 094	1 022.9 54%	249.8	935.6 326	402.6 814	186.5 1176	5.5
Anchored and oriented scaffolds	120	86.8 4.6%	723.1	1491.0 20	647.2 46	391.7 64	4.2

Supplementary Table 9: Transposable element (TE) content of the rainbow trout genome. The table contains coverage and the number of copies of transposable elements, but not all types of repeats. The percentage given in the text, 38%, contains non-TE repeats such as simple repeats, low complexity regions and small RNA pseudogenes.

Class/Family	Genome coverage (%)	Number of copies
DNA Transposons	**6.67353**	**358747**
En-Spm	0.0154	1546
Harbinger	0.03303	1977
Kolobok	0.01591	1112
Maverick	0.01525	274
PiggyBac	0.21695	16542
Tc-Mariner	5.50661	262466
Tsn1-3	0.04536	4564
hAT	0.72344	64909
Unclassified DNA	0.10158	5357
LINE Retrotransposons	**10.90741**	**562632**
CR1	3.62551	157926
I	0.01977	1572
Jockey	0.10853	6837
L1	1.14679	74845
L2	1.25691	58550
Penelope	0.13573	9643
R2	0.00828	204
R4-Zebulon	0.06107	9445
RTE-Rex3	0.0942	4911
Rex1-Babar	5.09399	259572
Unclassified LINE	0.56761	33379
LTR Retrotransposons	**4.32384**	**252223**
BEL	1.00599	32967
DIRS	1.04601	76632
ERV	0.0887	3135
Gypsy	2.18314	139489
SINE Retrotransposons	**0.65932**	**83310**
5S	0.04087	3963
Deu	0.02416	2045
Hpa1	0.44347	61334
MIR	0.00283	852
Mermaid	0.00021	89
RSG1	0.08812	5817
Sma1	0.01812	4008
Unclassified SINE	0.04154	5202
Unclassified TEs	**5.17156**	**311601**
TOTAL TEs	**29.18343**	**1617408**

Supplementary Methods

Rainbow trout doubled-haploid genomic DNA preparation.

A single homozygous doubled haploid YY male from the Swanson River (Alaska) clonal line[5,6] was used for all the sequencing. The identity of this male was confirmed to the Swanson clonal line using AFLP comparisons and allelic variation at 10 microsatellite markers (OMM1081 (AF352752), OMM1083 (AF352751), OMM1087 (AF352756), OMM1107 (AF375022), OMM1128 (AF375030), OMM1279 (AF470043), OMM1662 (BV212290), OMM5005 (CO805111), Ogo4UW (AF009796), Ots1BML (AF107029). This clonal line was previously used for the production of BAC-end sequences[7] from the trout 10X HindIII bacterial artificial chromosome (BAC) library[8], for the construction of the rainbow trout physical maps[9,10] and for the characterization of rainbow trout transcriptome in various tissues[11]. Genomic DNA was extracted from fin clips preserved in absolute ethanol using the phenol/chloroform extraction procedure, after overnight incubation at 55°C in lysis buffer (Tris-HCI 10 mM, NaC1 0.3 M, SDS 2%, EDTA 10 mM, urea 4 M and proteinase K 0.4mg.mL-1, Roche Applied Science) and digestion with RNase A (0.032 mg. mL-1, Promega). DNA was dissolved in TE buffer (Tris-HCl 10mM, EDTA 1 mM). The DNA extraction quality was estimated by agarose gel electrophoresis and by fluoremeter method (Qubit ®).

Genome assembly error corrections with Solexa/Illumina reads.

Sequence quality of scaffolds from the Newbler assembly was improved by automatic error corrections with Solexa/Illumina reads[12], (70-fold genome coverage). These Illumina reads were mapped onto the 454 assembled scaffolds using the BWA pipeline[13] and Samtools[14]. Only uniquely mapped reads were retained in the analysis. By applying Samtools pileup we

obtained the alignment description of each site of the reference. A sequence nucleotide in a Newbler contigs was considered as an error, and the corresponding base was changed, if all the following conditions were encountered: (i) Read coverage greater than 3 (without the reads mapping on the extremities on that position), and (ii) Quality greater than 20 and (iii) 70% of reads mapping in agreement with at least one read mapping on each strand to validate the change. 65.9% of the Illumina reads were then mapped at unique positions on the assembly, and 623,465 bp (160,729 substitutions with 94,944 N, 429,014 deletions and 33,722 insertions) were corrected.

Sequence anchoring on the genetic and physical maps

Anchoring to the consensus genetic map: A total of 2,226 markers from the INRA linkage map[15] were used to perform blastn alignment and *in silico* PCR (e-PCR) amplifications using the assembled sequences as a template. Blastn searches against scaffold and contig sequences were carried out using an e-value cut off of 1e-5 with following parameters (-r=1 –q=-1 –G=4 –E=2 –W=9 –p97). When alignment length was equal to subject length ± 5 nucleotides, the blastn results were directly validated. Otherwise, manual stringent blastn analyses were further carried out to remove nonspecific blastn hits using the following parameters (-r=1 –q=-3 –G=5 –E=2 –W=11 –e=0.001). Validation was performed as follows: for non-rainbow trout salmonid markers, the best hit was validated even if blastn identity was < 97%. For trout markers, blastn identity should be > 97%. Disruption of blastn alignment for EST-derived microsatellite sequences was accepted due to possible intronic sequences. For the *in silico* PCR amplifications (e-PCR)[16], both forward and reverse primers generated from 5’ and 3’ microsatellite and SNP sequences were used. The e-PCR amplification products were subsequently filtered to keep only markers with hits showing 100% identity and alignment length equal to that of known corresponding marker sequence. Only alignments

corresponding to a unique location on the assembly sequence were kept for further steps. In fact, for duplicated markers having multiple alignments and for markers expected not to be duplicated sharing sequence location with others markers, assignment could not be disentangled. This final set of alignments corresponded to our anchoring starting point.

Anchoring to the physical map: The anchoring process was extended by using BAC-end sequences (BES) data[7] and physical information from the second-generation physical and genetic- integrated maps[17]. Firstly we added assembly sequences containing at least one BES pair and placed between two sequences already anchored that were located on the same physical contig. Secondly, we assigned every assembly sharing at least two BES pairs with any previously anchored sequence to the corresponding linkage group. The process was repeated with every newly assigned sequence as long as additional unassigned sequences were found. Blastn search results revealed a total of 1,536 markers showing alignment hits, out of which 358 were directly validated and 1,178 were manually validated. Electronic PCR amplification results provided 148 additional markers (plus 246 which were already identified by blastn searches). Among those 1,684 markers, 1,096 markers blasted to one unique sequence in the assembly and did not share this location with any other marker, 362 were known to be duplicated and had multiple hits on the assembly, 93 shared location with one or more markers and may be novel duplicated markers, and 133 overlapped several scaffold or contig sequences. Crossing information from the RAD-based linkage map and from the non-duplicated markers on the INRA consensus map allowed assigning 58 additional markers. Altogether, 1,287 markers allowed the assignment of 1,137 sequences to 29 linkage groups.

Through the initial BES information of the physical/genetic integrated map, 143 assembly sequences were newly assigned to linkage groups, increasing the cumulative number of assigned sequences up to 1,270. The following iterative anchoring process based on BES

information of this set of assigned sequences identified 1,361 novel sequences, resulting in the cumulative assignment of 2,631 sequences.

Anchoring to the RAD-based linkage map: A publicly available rainbow trout RAD-based linkage map[18] (4,563 markers assigned to 29 linkage groups) was used for the final step. Marker sequences were mapped on the scaffold sequences using megablast alignment with following parameters (-E=0, -G=0, -p=100, -W=68). These alignments were then used to extract non-redundant assignment with the previous anchoring data. A sequential use of data from linkage[15,18] and physical[10] maps was used to anchor the sequence assembly onto chromosomes. The first anchoring step was performed onto the INRA consensus linkage map[15]. The assigned sequences served as a starting point for an iterative anchoring process based on information from the physical map[10]. Final sequence anchoring was enriched using markers from a high density/medium resolution RAD-based linkage map[18]. Megablast searches on the RAD-based linkage map resulted in the alignment of 3,898 markers. The 3,881 markers with a unique hit (7 markers with multiple hits were discarded) were aligned on 2,368 assembly sequences. The RAD-based linkage information identified 58 scaffolds with multiple map assignments that were excluded from further analyses. Among the 2,310 remaining sequences that were assigned to unique linkage groups, 528 were assigned in earlier steps and 1,782 were newly assigned. Because of the low resolution of the map, this corresponded to 270 different locations, with 1 to 190 sequences per location (average equal to 9).

Rainbow trout transposable elements.

Classification of TEs was based on Wicker's classification[19]. The rainbow trout TEs database was built combining both manual and automatic annotation. Two TE sequences were included independently in the database if their sequence diverged by more than 20% at the nucleotide

level. Automatic repeat libraries, based on number of repeats, were built by RepeatModeler software (http://www.repeatmasker.org/). The first one was built using BES and contains 443 sequences. The second one was built using scaffolds from 454 sequencing and contains 2,142 sequences. Those libraries include TEs, low complexity regions (e.g. AT rich regions), simple repeats (e.g. microsatellites), non-coding RNA and highly repeated unidentified sequences. A detailed precise manual annotation allows the detection of more divergent and less reiterated TEs. Homology-based identifications using blast on the genome are used with already annotated nucleotide and protein sequences from different TE families. More precise analyses were performed on BAC sequences by sliding window screen, searching for specific TE features with adapted software: LTR_FINDER for Long Terminal Repeats, e-inverted (http://emboss.bioinformatics.nl/cgi-bin/emboss/einverted) for Terminal Inverted Repeats, and manual identification of Target Site Duplications. This library contains 88 sequences of TEs. To better classify sequences, alignment and phylogenetic tree construction were performed using the ClustalW and PhyML package[20], respectively, with default parameters. Phylogenetic reconstructions were based either on reverse transcriptase for retrotransposons or transposase for DNA transposons. TE homology was also compared using CENSOR software[21], which searches for homologies in Repbase[22,23]. Automatic libraries and manual library were combined, avoiding redundancy; the most exhaustive library contained 633 sequences. The genome was masked and percentage of TEs was determined using RepeatMasker software[24] (and http://www.repeatmasker.org/). Twenty two percent of the genome was masked with the manual library and an additional 16% of the genome was masked using the final combined library, totaling 38% of the genome maksed by TEs and repeat sequences. The output file from RepeatMasker was parsed using an in-house script in order to count the number of hits per family. To evaluate the age of TE copies, Kimura distances were calculated based on the alignment (consensus from the TE library versus copy

in the genome) generated by RepeatMasker. The Kimura calculation uses the rates of transitions and transversions. Those rates are then transformed in Kimura distances using [K=-1/2 ln(1 – 2p – q) – ¼ ln(1 – 2q)] where “p” is the proportion of site with transitions and “q” the proportion of site with transversions.

Repetitive elements represent about 38% of the genome, with a large proportion of transposable elements (TEs) (about 27.73% of the genome with a genome size of 2.4 Gb, Supplementary Table 6). In comparison to fish and other vertebrate genomes, the TE coverage of the genome is within the expected range, if we take into account the differences in genome size: takifugu (0.4 Gb / 2.7% of TEs)[25], Nile tilapia (1.2 Gb / 14%)[26], zebrafish (1.7 Gb / 20%)[27], mouse (2.9 Gb / 38.2%)[28], Atlantic salmon (3 Gb / 30%)[29], clawed frog (3.1 Gb / 37%)[30], 1989) and humans (3.4 Gb / 44.8%)[31].

Both retrotransposons and DNA transposons were identified in a wide variety of families. Only a few vertebrate TE families, such as Helitron transposon or Copia retrotransposons are absent from this genome. The LINE retrotransposon is the most abundant class (10%), mostly represented by CR1 and Rex1. DNA transposons are mainly represented by Tc-Mariner superfamilies which number more than 260,000 copies. Among LTR retrotransposons, Gypsy families are the most predominant. Finally, at least seven families of SINEs were identified, mainly represented by the Hpa1 family.

Using Kimura distances, we estimated the relative age of the different TE families in the genome of the rainbow trout (Supplementary Fig. 7). It appears that two or three main bursts of transposition occurred in the genome. The most ancient one is mainly due to a high activity of Tc-Mariner families (Kimura value 41). In the second (around Kimura value 12), an increase of all families and particularly CR1 is highlighted. Finally, the last one (Kimura value 8) shows a new burst of Tc-Mariner activity.

Interestingly, retrovirus sequences were identified in the MHC region. One of the retrovirus sequences presents high similarity (91% on 360bp) with a VHSV (Viral haemorrhagic septicaemia virus – Rhabdovirus)-induced mRNA of the rainbow trout (Accession number AF483545). This suggests that the retrovirus was expressed in response to infection by VHSV virus, suggesting that it might be involved in defense reactions against other viruses.

Alignment of predicted proteins on ohnologous genomic DNA.

To better understand the fate of inactivated gene copies, we aligned the protein sequence predicted from a given gene model on its paralogous region, yielding 569 high confidence paralogous regions. In the case of singleton genes, this allows us to model the structure of the gene or pseudogene that may be located in the region or, if no alignment can be found, to identify a case of deletion or a gap in the assembly. Alignments were performed using exonerate[32] with the "--model protein2genome" option. Each protein sequence was compared to the genomic sequence of the entire ohnologous scaffold. A custom-made python script was then developed to diagnose 5 possible situations: 1) Absent: no alignment could be generated at all, presumably because the corresponding genomic sequence is absent from the scaffold due to a gap in the assembly or a deletion in the genome. 2) Ambiguous: if 5 or more independent gene structures could be modeled by exonerate on the corresponding scaffold, we then considered that the correct paralogous copy of the query gene could not be reliably identified among the different copies. 3) Functional: if less than 5 gene structures could be modeled by exonerate, the largest model included more than 90% of the amino acids in the query sequence, and that model did not contain stop codons or frameshifts. This gene structure was then considered as potentially functional, i.e. may be a gene annotation missed by the annotation pipeline or a pseudogene with an essentially complete open reading frame. 4) Incomplete: Same as Functional, but the model includes less than 90% of the amino acids

of the query sequence. This may be due to a gap in the assembly or to a truncating deletion or to amino acids incorrectly incorporated in the query sequence during the annotation process. 5) Pseudogene: if less than 5 gene structures could be modeled by exonerate, the largest gene structure includes at least one premature stop codon (with reference to the gene structure of the query) or at least one frameshift, we then considered this gene structure as a pseudogene.

In 21.3% of cases we could not identify a match, probably because of gaps in the assembly. Pseudogenes and incomplete gene models account for 56.2% of the exonerate results, while 10.4% and 11.5% singletons generate functional and ambiguous gene models respectively. We therefore find that 66.4% of singletons generate a clear paralogous gene model (pseudogene, incomplete and functional models). Figure 3c and Supplementary Fig. 2 describe the %ID of the singletons against their pseudogene model.

RNA-Seq and small RNA Illumina libraries preparation

RNA-Seq Illumina Libraries were prepared starting with total RNA (2-4 µg), mRNA was polyA-selected, chemically fragmented and converted into single-stranded cDNA using random hexamer priming. Then, the second strand was generated to create double-stranded cDNA. Paired-end libraries were prepared using SPRI works apparatus without sizing. Briefly, fragments were end-repaired, then 3`-adenylated, and Illumina adapters were added. DNA fragments were PCR-amplified using Illumina adapter-specific primers and then purified. Finally, libraries were quantified by qPCR (MxPro, Agilent Technologies, USA) and library profiles were evaluated using an Agilent 2100 bioanalyzer (Agilent Technologies, USA).

Small RNA libraries were constructed according to the Small RNA v1.5 sample preparation guide (Illumina). 5µg of total RNA were used for the construction of each library. Briefly, a 3'RNA adapter and a 5'adapter were ligated to both small RNAs ends. RT-PCR amplification

was performed and the PCR product was run on a 15% polyacrylamide gel. The band corresponding to miRNAs plus adapters (around 90-100 bp) was excised from the gel and eluted. Quality and quantity of the product was checked using a DNA 1000 chip (Agilent).

Gene Ontology analyses

Enrichment for particular functional classes was performed for the different categories of ohnologs identified according to their expression patterns, and independently, for the genes that have been recurrently kept as ohnologs at the 1R-2R, Ts3R and Ss4R duplication events. In the first case, trout genes were functionally annotated by reporting the Gene Ontology (GO) annotations of their orthologs in the human genome. The sample sets were in turn the HCSE, HCDE, NCSE or NCDE ohnologs, and the control sets were the remaining ohnologs, in order to test whether different patterns of expression correspond to functionally distinct sets of genes within the entire ohnolog set. In the second case, the ancestral genes present in the ancestral vertebrate genome (Euteleostomi, as deduced from the gene trees generated by TreeBest) were functionally annotated by reporting the GO annotations of the modern human genes. The sample set was the ancestral genes that have been retained as 1R-2R, Ts3R and Ss4R ohnologs, while the control set was the remaining ancestral genes, in order to test whether particular functional classes have been preferentially amplified and retained through successive rounds of WGD. In both cases, the GO analyses were performed in two steps: we obtained statistically enriched functional annotations in the sample set using a random sampling procedure (10,000 iterations, custom Perl script) with corrected false discovery rate for multiple testing (Benjamini-Hochberg FDR correction, with a 10% FDR threshold). The exact enrichment p-values for GO terms detected as significant through the random sampling procedure were then calculated using Fisher's exact test in R.

Supplementary References

1 Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome biology* **11**, R106, doi:10.1186/gb-2010-11-10-r106 (2010).

2 Kasahara, M. *et al.* The medaka draft genome and insights into vertebrate genome evolution. *Nature* **447**, 714-719, doi:10.1038/nature05846 (2007).

3 Makino, T. & McLysaght, A. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 9270-9274, doi:10.1073/pnas.0914697107 (2010).

4 Howe, K. *et al.* The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496**, 498-503, doi:10.1038/nature12111 (2013).

5 Parsons, J. E. & Thorgaard, G. H. Production of androgenetic diploid rainbow trout. *The Journal of heredity* **76**, 177-181 (1985).

6 Young, W. P., Wheeler, P. A., Fields, R. D. & Thorgaard, G. H. DNA fingerprinting confirms isogenicity of androgenetically derived rainbow trout lines. *The Journal of heredity* **87**, 77-80 (1996).

7 Genet, C. *et al.* Analysis of BAC-end sequences in rainbow trout: content characterization and assessment of synteny between trout and other fish genomes. *BMC genomics* **12**, 314, doi:10.1186/1471-2164-12-314 (2011).

8 Palti, Y., Gahr, S. A., Hansen, J. D. & Rexroad, C. E., 3rd. Characterization of a new BAC library for rainbow trout: evidence for multi-locus duplication. *Anim Genet* **35**, 130-133, doi:10.1111/j.1365-2052.2004.01112.x (2004).

9 Palti, Y. *et al.* A first generation integrated map of the rainbow trout genome. *BMC genomics* **12**, 180, doi:10.1186/1471-2164-12-180 (2011).

10 Palti, Y. *et al.* A first generation BAC-based physical map of the rainbow trout genome. *BMC genomics* **10**, 462, doi:10.1186/1471-2164-10-462 (2009).

11 Salem, M., Rexroad, C. E., 3rd, Wang, J., Thorgaard, G. H. & Yao, J. Characterization of the rainbow trout transcriptome using Sanger and 454-pyrosequencing approaches. *BMC genomics* **11**, 564, doi:10.1186/1471-2164-11-564 (2010).

12 Aury, J. M. *et al.* High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. *BMC genomics* **9**, 603, doi:10.1186/1471-2164-9-603 (2008).

13 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

14 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

15 Guyomard, R., Boussaha, M., Krieg, F., Hervet, C. & Quillet, E. A synthetic rainbow trout linkage map provides new insights into the salmonid whole genome duplication and the conservation of synteny among teleosts. *BMC Genet* **13**, 15, doi:10.1186/1471-2156-13-15 (2012).

16 Schuler, G. D. Sequence mapping by electronic PCR. *Genome research* **7**, 541-550 (1997).

17 Palti, Y. *et al.* A second generation integrated map of the rainbow trout (Oncorhynchus mykiss) genome: analysis of conserved synteny with model fish genomes. *Marine biotechnology* **14**, 343-357, doi:10.1007/s10126-011-9418-z (2012).

18 Miller, M. R. *et al.* A conserved haplotype controls parallel adaptation in geographically distant salmonid populations. *Molecular ecology* **21**, 237-249, doi:10.1111/j.1365-294X.2011.05305.x (2012).

19 Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nature reviews. Genetics* **8**, 973-982, doi:10.1038/nrg2165 (2007).

20 Gouy, M., Guindon, S. & Gascuel, O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular biology and evolution* **27**, 221-224, doi:10.1093/molbev/msp259 (2010).
21 Kohany, O., Gentles, A. J., Hankus, L. & Jurka, J. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC bioinformatics* **7**, 474, doi:10.1186/1471-2105-7-474 (2006).
22 Jurka, J. Repbase update: a database and an electronic journal of repetitive elements. *Trends in genetics : TIG* **16**, 418-420 (2000).
23 Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research* **110**, 462-467, doi:10.1159/000084979 (2005).
24 Huda, A. & Jordan, I. K. Analysis of transposable element sequences using CENSOR and RepeatMasker. *Methods in molecular biology* **537**, 323-336, doi:10.1007/978-1-59745-251-9_16 (2009).
25 Aparicio, S. *et al.* Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* **297**, 1301-1310, doi:10.1126/science.1072104 (2002).
26 Shirak, A. *et al.* Identification of repetitive elements in the genome of Oreochromis niloticus: tilapia repeat masker. *Marine biotechnology* **12**, 121-125, doi:10.1007/s10126-009-9236-8 (2010).
27 Ivics, Z., Izsvák, Z. & Hackett, P. B. Repeated sequence elements in zebrafish and their use in molecular genetic studies. *The Zebrafish Science Monitor* **3** (1995).
28 Mouse Genome Sequencing, C. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-562, doi:10.1038/nature01262 (2002).
29 Davidson, W. S. *et al.* Sequencing the genome of the Atlantic salmon (Salmo salar). *Genome biology* **11**, 403, doi:10.1186/gb-2010-11-9-403 (2010).
30 Carroll, D., Knutzon, D. S. & Garrett, J. E. in *Mobile DNA* (eds M. Howe & D. Berg) 567-574 (1989).
31 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921, doi:10.1038/35057062 (2001).
32 Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC bioinformatics* **6**, 31, doi:10.1186/1471-2105-6-31 (2005).