Supplementary data

"The medaka fish *Tol2* transposable element is in an early stage of decay: identification of a nonautonomous copy" by Sakura Hayashi et al.

Figure S1. Procedures used for BLAST search and partitioning of sequences. A BLAST search was performed using the left terminal region (200 bp) of *Tol2* as a query sequence. For each of the returned hits, a long genomic sequence (20 kb that included the query-aligned region at the center) was drawn from the genome browser and partitioned into segments S1 to S5. The same BLAST search and partitioning were conducted using the right terminal region (200 bp) of *Tol2* as a query sequence.

No.	← S4 →	← S2 →	← S1 →	← S3 →	← S5 →	
01	GTAAGCCCTGGTTTGC	TGA GAAATTA	CAGAGGTGTAAA...	TTTACACCTCTG	GAAATTA	CTTCTCGGTTCACTTGCTAA
02	AGGACAGCAGCAGCCAGTCA	CTACCACA	CAGAGGTGTAAA...	TTTACACCTCTG	CTACCACA	TGAATATCGCTTCTTACTTT
03	AGCACTTCAGGGCTGCATGG	ACAATGTA	CAGAGGTGTAAA...	TTTACACCTCTG	ACAATGTA	AGTGTGTTGATAAAATAATAAT
04	CGTAAACGCAATAGATAAGG	GAGGCTTT	CAGAGGTGTAAA...	TTTACACCTCTG	GAGGCTTT	GAATGTCCTTTCCGAACGTC
05	ATTGTTGCTATAATATATA	TGAATAAC	CAGAGGTGTAAA...	TTTACACCTCTG	TGAATAAC	GTCTTTGATCAGCCCCAATTA
06	GGGACACAATCATGAAGTGG	AACAAAAT	CAGAGGTGTAAA...	TTTACACCTCTG	AACAAAAT	TTATTTGGATATTTCAAAC
07	ACATTGGTGCCCTTTGTTTA	GTCACGAG	GAGAGGTGTAAA...	TTTACACCTCTG	GTCACGAG	CGGTTAAACGTCGGCGAATA
08	CAAAATAGTGGATGAATTAC	TGCATGCC	CAGAGGTGTAAA...	TTTACACCTCTG	TGCATGCC	TCTGTAGAGATGCACAAAGC
09	TGCTGTTGATGAAGATGTTA	ATTAATGA	CAGAGGTGTAAA...	TTTACACCTCTG	ATTAATGA	TGCTTTTAAATGCAAGGAGGG
10	GAATATTAAAGTCATGTTA	CACAAAGA	CAGAGGTGTAAA...	TTTACACCTCTG	CACAAAGA	TAAATGGCGCTTAACTAGCA
11	GCTCACTTTTCTTTTCCCTT	CCAAGGTC	CAGAGGTGTAAA...	TTTACACCTCTG	CCAAGGTC	TGAAGATGTCCTGAAAGCAAT
12	CAGAGCCTTGATACAGTACA	GTTAAGCT	CAGAGGTGTAAA...	TTTACACCTCT	GTTAAGCT	AGGTGACCGGGACTTCTTTT
13	CACAAAGCCAACTGAGCTGGC	ATGATCTC	CAGAGGTGTAAA...	TTTACACCTCTG	ATGATCTC	CCATATTTTGCCCTGATGTTT
14	ACCTGGCACTGCTCTGATCT	GTTGTTGA	CAGAGGTGTAAA...	TTTACACCTCTG	GTTGTTGA	GCTCCTCAGACTCCCTCTG
15	AGAGCACAACTCTGCACCTGT	TGTGTATC	CAGAGGTGTAAA...	TTTACACCTCTG	TGTGTATC	CATTTAACTCACCCTGATAGC

No.	Location of query-aligned region	Length of S1 (bp)
01	1 7498021 to 7498220 (+)	4688
02	1 29953385 to 29953584 (+)	4687
03	1 30776735 to 30776934 (-)	4688
04	5 14707138 to 14707337 (+)	4676
05	7 18031704 to 18031903 (+)	4687
06	8 21653114 to 21653313 (+)	4688
07	10 1688406 to 1688604 (-)	4670
08	10 1751688 to 1751887 (+)	4683
09	10 19085880 to 19086079 (-)	4687
10	13 14657274 to 14657473 (+)	4568
11	13 20193109 to 20193308 (+)	4682
12	14 18002920 to 18003119 (-)	4685
13	20 23235089 to 23235288 (-)	4688
14	22 14903387 to 14903586 (+)	4686
15	23 7003245 to 7003444 (-)	4687

Figure S2. Genomic sequences containing putative *Tol2* units. A total of 15 sequences (indicated by No.) were identified, and they were partitioned into 5 segments (S1 to S5; see Figure S1). The location of the query-aligned region on the medaka HDR genome browser is also shown in the format used therein.

Figure S3. Verification of the focal variations by sequencing. **A.** PCR primers. The triangles indicate the location and direction of the PCR primers used. The left primer represented a 30-nt segment in *Tol2*, which was located just downstream of the IIRs. The right primer represented a 30-nt segment in the right-flanking region, which was 300 bp from the right end of *Tol2*-HDR07. Using a left primer corresponding to part of the left-flanking region would have been ideal. However, a PCR for a long region including the IIRs would not yield a clear single fragment because the IIRs form a hairpin structure that interrupts the movement of DNA polymerase. Therefore, a left primer was designed at an internal region. **B.** Result of PCR. The reaction mixture was electrophoresed on a 1.0% gel. The template DNAs used for the respective lanes are shown above the gel image. Orange is an orange-colored medaka fish variant purchased from a pet shop in Nagoya City. **C.** Sequences and sequence chromatograms obtained. The DNA fragment amplified from the HDR genomic DNA was sequenced. The sequence obtained was identical throughout its entire length to the *Tol2*-HDR07 sequence. Sequence chromatograms are shown for regions containing variations J, K, and L.

Abbreviations

Pro	CMV promoter
Neo	neomycin-resistance gene
PA	poly(A) addition signal
Ori	plasmid replication origin
Amp	ampicillin-resistant gene

GenBank accession numbers

<i>Tol2</i> element	D84375 (4682 bp)
<i>Tol2</i> cDNA	AB032244 (2156 bp)
pBluescript II SK+	X52328 (2961 bp)
pCS2+	Not registered (4095 bp)
pCMV-Tag1	AF025668 (4319 bp)

Figure S4. Structure of the donor plasmid and helper plasmids and their source materials. pDon and pHel-Int were constructed by combining the components shown here using the In-Fusion cloning system (product of Takara Bio Inc.). Note that the components are not to scale.

Figure S5. Structure of the coding region of helper plasmids. pHel-Int is a helper plasmid that contains the intact transposase cDNA. Other helper plasmids were prepared by PCR using pHel-Int as a template. The variation to be generated was included at the 5' end(s) of the primers, and the product fragment was circularized by self-ligation. pHel-Def contained a coding sequence for a defective transposase protein where an extra 4 nucleotides were inserted after the 1142nd nucleotide. This caused a frame shift mutation resulting in the alteration of amino acids at and after the 494th site and the emergence of a stop codon at the 502nd site. Elements of the *hAT* and many other superfamilies are known to carry 3 conserved amino acids (D, D, and D/E) in the transposase sequence (Yuan and Wessler 2011), which play an essential role in capturing metal cation cofactors (Bolland and Kleckner. 1996). These conserved amino acids are located at sites 192, 258, and 611 of the *Tol2* transposase (Jiang 2016). In pHel-Def, the last residue was lost due to the 4-nucleotide insertion. Plasmids pHel-J, pHel-K, and pHel-L, contained variations J, K, and L, respectively.