学 位 論 文

Structures and Transposition Activity of A Novel Giant DNA Transposon, *Teratorn*

(巨大な新規 DNA トランスポゾン *Teratorn* の構造と転移活性)

平成 29 年 6 月博士　(理学)　申請

東京大学大学院理学系研究科

生物科学専攻

井 上　　雄 介

Contents

Abbreviations

- BAC — bacterial artificial chromosome
- BLAST — basic local alignment search tool
- dpf — days post fertilization
- ds — double-stranded
- ERV — endogenous retrovirus
- EVE — endogenous viral element
- HHV — human herpesvirus
- kb — kilo base pairs
- MGE — mobile genetic element
- ORF — open reading frame
- RT-PCR — reverse transcription polymerase chain reaction
- SNP — single nucleotide polymorphism
- ss — single-stranded
- TE — transposable element
- TIR — terminal inverted repeat
- TSD — target site duplication
- *zic1*/*zic4* — *zic1* and *zic4*

Abstract

Mobile genetic elements (e.g. transposable elements and viruses) display significant diversity with various life cycles, but how novel elements emerge remains obscure. Recent studies demonstrate that some elements with distinct life cycles share "hallmark genes" associated with replication, implying the existence of a network-like evolutionary relationships among them. In my doctoral thesis, I report a novel and giant (180-kb long) mobile element, *Teratorn*, originally identified in the genome of medaka, *Oryzias latipes*. *Teratorn* belongs to the *piggyBac* superfamily and retains the transposition activity. Remarkably, *Teratorn* is largely derived from a novel herpesvirus of the *Alloherpesviridae* family that infects fish and amphibians. In addition, through the genomic survey of *Teratorn*-like elements in other species, I identified a novel group of alloherpesvirus species widespread in the genomes of teleost fish species. Importantly, some of them exist as a fused form of *piggyBac* transposon and herpesvirus genome, implying the generality of transposon-herpesvirus fusion. I propose that *Teratorn* was created by a unique fusion of DNA transposon and herpesvirus, leading to life cycle shift. My study thus supports the idea that recombination or fusion of multiple elements belonging to distinct classes is a key event in generation of novel mobile genetic elements.

General Introduction

Mobile genetic elements : key players in genome evolution

All cellular organisms on the earth are parasitized by selfish-replicating genetic entities, named mobile genetic elements (MGEs). These elements include transposable elements (TEs) and viruses, as well as plasmids and self-splicing elements[1]. MGEs are very abundant biological entities; viruses outnumber 10 to 100-fold to cellular organisms[2], while TEs can occupy a large part of the host genome (e.g. ~45% in human[3] and ~85% in maize[4]). Even though each class of elements has its own different life cycle (e.g. TEs for persistent stay and propagation inside host genomes and viruses for transient stay and massive transfer between cells), what all elements have in common is the ability to parasitize cells and propagate by usurping host metabolism. Owing to their selfish nature, MGEs frequently cause harmful effects on host organisms, such as insertional mutation by TEs and pathogenesis by viruses. Thus, relationships between MGEs and hosts are generally competitive, exerting enormous impacts on each other[5,6]. For example, host organisms have developed several defense systems against MGEs, such as the restriction modification and CRISPR-Cas systems in bacteria[6,7], small RNA-mediated silencing[8] and innate immunity in eukaryotes[6], KRAB zinc finger proteins[9,10] and V(D)J recombination in vertebrates[7]. To overcome these defense systems, MGEs have developed several mechanisms to evade and/or antagonize them, such as target-specific integration of TEs[5,11] and evasion from host

immunity by fast evolution and acquisition of antagonistic molecules in viruses[12,13]. Such continuous arms race between MGEs and host organisms may have led to the massive diversification of each entity. On the other hand, host organisms have sometimes co-opted MGEs for their own functions. For example, prokaryotic MGEs are frequently utilized for horizontal transfer of genes encoding enzymes of antibiotics or toxins, enabling prokaryotic cell population to survive under harsh environments or kill off competitors[14,15]. In eukaryotes, TEs have contributed to construction and re-wiring of gene regulatory networks[11,16–18]. In addition, MGEs can provide host organisms with some specific traits. Examples include V(D)J recombination in vertebrates from a *Transib*-like transposon, the CRISPR/Cas system in bacteria from *Casposons*[7], introns in eukaryotes from group II introns[19] or DNA transposons[20], polydnavirus in parasitoid wasps[21] and placenta in mammals from retroviruses[16,22,23]. Together, MGEs have been key players in the evolution of cellular organisms. Despite their importance, however, we are far from full understanding of how MGEs themselves have evolved; that is, the origin of MGEs, relationships between different classes of elements, and the underlying mechanisms of the emergence of novel elements.

How did novel mobile elements emerge?

As described above, MGEs are characterized by the capacity to parasitize cells and propagate in a selfish manner. In addition, although there are no single gene conserved among all mobile genetic elements, all elements contain "hallmark genes"

associated with replication (e.g. polymerase, helicase and transposase) and structural constitution (e.g. capsid protein), which are not shared by cellular organisms[24,25]. Despite their common characteristics, MGEs are quite diverse in their genome constitution (either DNA or RNA of double-stranded (ds) or single-stranded (ss) nucleotides), replication manner (e.g. DNA polymerization, RNA-dependent RNA transcription, reverse transcription) and life cycle (with/without extracellular phase)[1,26,27] (Fig. 1). How did this diversity arise and what mechanism underlined this process? It has been proposed that they are evolutionarily linked with each other, forming network-like phylogenetic relationships[24,28]. Indeed, in prokaryotic MGEs, massive gene exchange between phages, transposons and plasmids occurred, making the borders of different classes of elements ambiguous[29]. In eukaryotic MGEs, the existence of network-like evolutionary relationships has also been proposed[24,25,28]. The most clear example is the transition from LTR retrotransposons to retroviruses and vice versa[23,30–32]. The former transition is thought to be mediated by the gain of an envelope gene from other viruses such as baculovirus and herpesvirus[30], while the latter could be a result of chromosomal integration into germline cells followed by propagation inside host genomes[23,32], constituting the large proportion of the host genome (e.g. ~8% of the human genome[3]). In addition, recent phylogenetic analyses implied that the transition events between transposon and virus also occurred for other elements in the distant past. For example, *Polintons*, replicative large DNA transposons widely distributed among eukaryotes, share genes with several other MGEs[33–35],

suggesting the evolutionary links between them. However, due to the low sequence similarity between diverged mobile elements as well as the incongruence of phylogeny for each gene[36,37], in most cases it remains ambiguous which element is actually a result of recombination events, leading to a transition in life cycle.

In my doctoral thesis, I report a novel DNA transposon "*Teratorn*" in a small teleost fish medaka (*Oryzias latipes*), and provide direct evidence that genetic interaction between different classes of mobile elements can produce novel elements. In chapter 1, I demonstrate that *Teratorn* is a ~180-kb long mobile element created by a unique fusion of *piggyBac* transposon and herpesvirus. In chapter 2, I report that *Teratorn*-like elements are widely distributed among teleost genomes, implying the generality of this type of fusion event between *piggyBac* transposon and herpesvirus.

Chapter 1

Identification of active and giant transposon, *Teratorn*, in the medaka genome - complete fusion of transposon and herpesvirus -

Introduction

Transposable elements (TEs) are DNA sequences that can move and propagate inside the genomes of host organisms. They constitute a large portion of eukaryotic genomes and contribute to genome evolution via providing genetic novelties[5,38]. TEs are categorized into two classes, according to whether they include RNA intermediates or not (Fig. 1a). Class I elements (retrotransposons) include long terminal repeat (LTR) and non-LTR elements, both of which transpose in a "copy-and-paste" manner mediated by reverse transcription. Class II elements (DNA transposons) are further classified into two subclasses depending on the type of transposition, i.e. in a "cut-and-paste" or "copy-and-paste" manner[26]. Even though each class holds dozens of superfamilies, they are generally small (in most cases less than 10-kb long) and possess only 1–3 genes[26,38]. Exceptions of this tendency are *Polintons* (self-synthesizing DNA transposons, 10–25 kb with more than 10 virus-like genes) and *Helitrons* (rolling-circle DNA transposon, 6–17 kb with a single large multidomain protein) (Fig. 1a)[33,38–40].

Teratorn was originally identified by Y. Moriyama as an inserted DNA element in the medaka spontaneous mutant *Double anal fin* (*Da*)[41] (Fig. 2a). In *Da*, *Teratorn* is inserted in the regulatory region of *zic1*/*zic4* gene and blocks their expression in somites while maintaining in neural tube, leading to the dorsal part of the trunk morphology ventralized (e.g. fin morphology, pigmentation pattern, lateral line distribution and body shape). It was initially named *Albatross*, but

recently renamed *Teratorn* (after the extinct group of very large birds of prey), to avoid confusions of the same name of the gene encoding a Fas binding factor[41]. In the previous study, partial sequencing suggested that *Teratorn* is a DNA transposon that moves in a "cut-and-paste" manner, since it contains 18-bp terminal inverted repeats (TIR) at its termini, and is flanked by target site duplications (TSD)[26]. Interestingly, *Teratorn* was unique in size (at least 41-kb long[41]) and in gene number (at least six genes). In addition, at least another three medaka spontaneous mutants were found to be caused by insertion of *Teratorn*; *rs-3* (*ectodysplasin-A receptor*)[42], *pc* (*glis*)[43] and *abc*^{def} (*pkd1l1*)[44] (Fig. 2b-d). Thus, I supposed that *Teratorn* is a novel active and giant DNA transposon with unique characteristics in evolutionary origin, life cycle and degree of impact on the host genome. However, due to multicopies as repetitive sequences in the medaka genome, the entire structure of *Teratorn* could not be deduced from the previously published medaka draft genome constructed by shotgun sequencing (ref. Kasahara et al., 2007[45]). Thus, I decided to perform BAC-based full-length sequencing of *Teratorn* copies in the medaka genome to unravel the characteristics of *Teratorn*.

Results

***Teratorn* is ~180-kb long.** I screened a BAC library of the Hd-rR strain to obtain clones containing the entire *Teratorn* (Fig. 3). Using PacBio long-read sequencing I determined the full-length sequences of six *Teratorn* individual copies and found that in five clones, *Teratorn* was ~180-kb and in one clone it was ~155-kb (Fig. 4,5). The 155-kb clone is a shorter version of the other five with a partial deletion. In addition, I identified another subtype of *Teratorn*, which exhibits ~88% sequence identity but little similarity in the repetitive region including terminal region, by blast search of the full-length *Teratorn* sequence against a new, higher quality, version of medaka genome (Fig. 6a,b) (http://utgenome.org/medaka_v2/#!Top.md). Structurally, the two subtypes are very similar to each other. Except for an 80-kb long inversion in the middle, the overall structure, such as the size, gene order (see below) and position of long repeats, is conserved (Fig. 6c–e). Thus, I concluded that they were derived from a common ancestor, and referred to the initially identified copies as *Teratorn* 'subtype 1' and the second one as *Teratorn* 'subtype 2'. Blast search and southern blot analysis of their terminal sequences suggested that there exist 30–40 copies of subtype 1 and ~5 of subtype 2 per haploid genome of the medaka Hd-rR inbred strain (Fig. 6f).

***Teratorn* belongs to *piggyBac* superfamily.** I first looked for a gene encoding transposase and found a gene homologous to *piggyBac* transposase inside *Teratorn*

(magenta arrow in Fig. 4b). *piggyBac* is one of the major "cut-and-paste" DNA transposon superfamilies and is widely distributed among eukaryotes, especially in metazoans[38]. The transposase gene of *Teratorn* exhibits high sequence similarity with other *piggyBac* superfamily transposase genes with essential amino acid residues strictly conserved, e.g. the catalytic domain (DDD motif), especially the four aspartic acid residues required for the transposition reaction (Fig. 7a)[46]. Furthermore, the sequence composition of TIRs and TSDs at its termini follows the rule of the *piggyBac* superfamily; *piggyBac* transposons contain TIRs with 12–19bp, beginning with a "CCYT" motif, and preferentially target TTAA (Fig. 7b)[26,38]. These characteristics indicate that *Teratorn* belongs to *piggyBac* superfamily. *Teratorn* is thus the biggest among known transposons, as the largest transposons reported so far are a *Polinton*-like DNA transposon (up to 25-kb long)[33,40] and a *Gypsy*-like LTR retrotransposon (up to 25-kb long)[47].

***Teratorn* contains the genome of a herpesvirus.** In addition to the transposase gene, GENSCAN[48] and GeneMarkS[49] predicted ~90 putative genes, covering more than 60% of *Teratorn* (Fig. 4b, Table 1). Remarkably, among the predicted genes, 17 genes show sequence similarity to those of herpesviruses (E-value < 0.01; blue arrows in Fig. 4b, Table 1). Those include genes required for virus propagation, such as DNA replication (DNA polymerase, primase and UL21 homolog DNA helicase), virion maturation (capsid maturation protease), packaging of viral DNA (large subunit terminase) and virus structural proteins (major capsid protein, subunit 2

capsid triplex protein and envelope glycoprotein), and are known to be essential for the life cycle of herpesviruses.

Besides the essential genes, ~20 genes show no sequence similarity to those of other herpesviruses but are found in other organisms, which may have been secondarily obtained from infected host genomes or other sources such as bacteria and viruses (yellow arrows in Fig. 4b, Table 1). Intriguingly, most of them appear to function in regulation of immune response or cell proliferation in hosts. The former includes cell surface proteins (CD276 antigen-like, CXCR-like, TNFR-like) and immune signal transduction factors (ZFP36-like, TARBP-like), while the latter includes cell cycle regulator (CDK-like), apoptosis inhibitor (Mcl1-like) and oncogene (pim-like). The existence of these gene repertoires is a common feature of herpesvirus species that infect the haematopoietic cell lineages (like betaherpesviruses and gammaherpesviruses of amniotes)[50]. The rest ~50 genes (gray arrows in Fig. 4b) show no significant blast hit against the current non-redundant protein database. Given that large DNA viruses tend to carry many unknown genes in addition to essential ones[51,52], it is likely that all predicted genes present in *Teratorn* (except for the tranposase gene) constitute a whole herpesvirus genome.

Herpesviruses are double-stranded DNA viruses that infect a wide variety of vertebrates and some invertebrates. Their genome is relatively large, ranging from 124-kb to 295-kb, and contains 70 to 200 genes[53,54]. They share common characteristics, such as the way of DNA replication and packaging into the capsid,

the three-dimensional structure of the capsid and the ability to establish a life-long persistent infection as a viral episome floating in the nucleus[50]. Herpesvirus species are classified into three families : *Herpesviridae* (amniotes), *Alloherpesviridae* (fish and amphibians) and *Malacoherpesviridae* (molluscs)[53]. Phylogenetic analysis based on amino acid sequences of the terminase gene, the only gene confidently conserved among all herpesvirus species, suggests that *Teratorn* belongs to the family *Alloherpesviridae* (Fig. 8a). Within *Alloherpesviridae*, the same topology was obtained from the phylogenetic analysis using the concatenated amino acid sequences of four genes (terminase, DNA polymerase, DNA helicase and major capsid protein, Fig. 8b). Importantly, *Teratorn* contains all 13 genes which are conserved among all alloherpesvirus species, and also shares other genomic features with alloherpesviruses such as the genome size, number of genes, GC content and existence of long repeats (Fig. 8c)[53,54]. All these data support the idea that *Teratorn* contains the complete genome of a novel alloherpesvirus.

***Teratorn* exists as a fusion of DNA transposon and herpesvirus.** The overall structure of the six *Teratorn* copies sequenced suggested that *Teratorn* is a fusion of the *piggyBac*-like DNA transposon and a novel alloherpesvirus. This raised the question of whether these two mobile genetic elements exist only in the fused form or also exist separately in the medaka genome. To address this question, I searched for genomic neighborhoods in all contigs from the Hd-rR genome assembly, which contain the transposase gene (Fig. 9a). Among the 70 obtained contigs, all but one

included the above herpesvirus genes at either 5' or 3' side to the transposase gene. For the remaining one contig, I was unable to examine the presence of viral genes because the assembly was interrupted by tandem repeats (Fig. 9b). Thus, essentially no *Teratorn* copy exists as a typical configuration of "cut-and-paste" DNA transposons (the transposase gene and TIRs). Furthermore, no herpesvirus-like sequences exist alone in the medaka genome (data not shown). Taken together, I conclude that all *Teratorn* copies exist as a fusion of the *piggyBac*-like transposon and herpesvirus genome (Fig. 4).

***Teratorn* retains the transposition activity.** The above structural characteristics suggest that *Teratorn* has at least the transposon-like life cycle. To test this idea, I first examined if the encoded transposase is active by the *in vitro* assay that directly detects transposition activity (Fig. 10–12)[55–57]. In this assay using HEK293T cells, I co-transfected helper plasmid that ubiquitously expresses transposase, and indicator plasmid that includes the puromycin-resistant gene flanked by the *Teratorn* terminal sequences, and tested excision and chromosomal integration of the transposon cassette (Fig. 10a,d). Excision of the transposon cassette from the indicator plasmid was examined by PCR using primers that flank the transposon cassette. I found that a PCR band was detected only when the helper plasmid was co-transfected with the indicator plasmid (Fig. 10b). I then tested the integration activity of *Teratorn* transposase. I cultured transfected cells in puromycin-containing medium for about two weeks, to screen for cells that became

puromycin-resistant after chromosomal integration of the cassette. However, few colonies were detected after puromycyn selection (Fig. 10e).

Since *Teratorn* possesses the additional TIRs at the boundary of a pair of long inverted repeats, i.e. 'internal TIRs' (Fig. 4, 6e), I hypothesized that internal TIRs are also required for integration reaction. Thus, I inserted it into the indicator plasmid so as to mimic the endogenous structure of *Teratorn* and performed the same assays again. I found that both excision and integration reaction took place this time (Fig. 10c,f). For excision assay, sequencing of PCR products revealed that transposon cassette was precisely excised at the boundary, which is the characteristics of *piggyBac* superfamily DNA transposons[57] (Fig. 11). For integration assay, southern blot and sequencing analyses[58] of genomic DNA of survived colonies confirmed the chromosomal integration of the plasmid sequence via transposition, although all of the copies sequenced so far were integrated via internal TIRs (Fig. 12). Since DNA cleavage reaction could occur both at external TIRs and internal TIRs in the artificial circular form of the indicator plasmid (Fig. 12b), this data implies that, for integration reaction, external TIRs are less frequently used than internal TIRs by unknown reason. In any case, these results indicate that the transposase of *Teratorn* is capable of mediating transposition *in vitro*.

To examine the transposition of *Teratorn in vivo*, I searched for integration site polymorphisms between the two groups of the Hd-rR medaka inbred strains, which had been kept separately for more than 20 years in our

laboratory (University of Tokyo) and at the National Institute for Basic Biology. Southern blot analysis of *Teratorn* terminal sequences showed distinct band patterns betIen the two groups (Fig. 13), suggesting that endogenous *Teratorn* transposition indeed occurred *in vivo*. Taken together, these results indicate that *Teratorn* still retains the activity of transposition *in vivo* and adopts the life cycle of the *piggyBac* superfamily DNA transposon.

***Teratorn* includes intact herpesvirus genes.** *Teratorn* is an active transposon, but could also retain some aspects of the herpesvirus. Indeed, *Teratorn* contains genes involved in virion morphogenesis (major capsid protein, capsid triplex protein, DNA packaging terminase and capsid maturation protease, Fig. 4b, 8c, Table 1). Importantly, sequence comparison revealed that catalytic sites in the ATPase and nuclease domain of terminase[59,60], as well as the catalytic triad of capsid maturation protease[61], are conserved for *Teratorn* (Fig. 14). For major capsid protein and capsid triplex protein, moderate sequence similarity to other alloherpesvirus species was detected (Fig. 15, blue). Although there is little sequence similarity to *Herpesviridae* family, similar pattern of secondary structure (α-helix and β-sheet) was detected, suggesting the conserved 3D structure of major capsid protein (Fig. 15, orange)[62,63]. These data suggest that *Teratorn* equips virion morphogenesis machinery.

To know whether *Teratorn* is still active as a virus, I examined the extent of ORF degradation in *Teratorn* copies. Since Illumina whole-genome shotgun sequencing data are publically available for the Hd-rR strain[64], which include

sequences of all *Teratron* copies, I called SNPs and indels inside *Teratorn* by aligning *Teratorn*-derived reads to the *Teratorn* reference sequences. The reference sequences of subtype 1 and subtype 2 were constructed based on sequences from a single BAC clone (73I9) and on the consensus sequences deduced from all copies in the Hd-rR genome, respectively. I found that the number of nonsense mutations is much smaller than that of missense and synonymous mutations in coding regions (Table 2). Consistently, for the six *Teratorn* copies sequenced so far, all ORFs were found to be intact, suggesting that genes of *Teratorn* are currently functional.

Furthermore, I examined transcription of selected *Teratorn* genes including transposase and virus-related genes. RT-PCR analysis detected their ubiquitous and low levels of expression in nearly all tissues (e.g. brain, liver, muscles, gonads) and in developing embryos (5 dpf (days post fertilization)) (Fig. 16). This pattern of expression could represent latency-associated conditions, and could change upon reactivation. To explore this possibility, I treated medaka embryonic-derived fibroblast cells with 5-azacytidine, N-butylate and 12-*O*-Tetradecanoylphorbol 13-acetate (TPA), all of which are chemicals known to induce reactivation of latently infected herpesviruses in human cells[50,65,66]. I found that, although N-butyrate and TPA alone didn't have an effect (data not shown), some of the *Teratorn* genes were moderately derepressed after administration of 5-azacytidine; yet, the expression level of most genes was not sufficiently high (10^{-5}–10^{-2} to *β*-actin, Fig. 17), as compared with herpesvirus genes in activated mammalian cells[67]. However, virus structural proteins derived from *Teratorn* were

not detected at these conditions by western blotting (major capsid protein, capsid triplex protein and envelope glycoprotein, Fig. 18).

Taken together, *Teratorn* encodes intact herpesvirus genes and thus likely utilize virus replication machinery for their propagation. However, conditions for full reactivation and virion formation remain elusive.

Discussion

***Teratorn* : a novel giant transposon formed by a unique fusion of *piggyBac* and herpesvirus.** Here I have characterized the mobile element, *Teratorn*. My study demonstrated that *Teratorn* has the gene encoding transposase of the *piggyBac* superfamily and retains transposition activity. One of the unique features of *Teratorn* is its size (~180-kb long), by far the biggest reported for a transposon. Thus, the transposition of *Teratron* would have greater impact on host genes and genomes. Indeed, among the previously published mutants, at least four medaka spontaneous mutants are caused by insertion of *Teratorn* (Fig. 2). Because of its serious impact, the transposition activity of *Teratorn* is likely suppressed *in vivo* possibly through an epigenetic mechanism. Indeed, a previous study demonstrated that *Teratorn* inserted in the *zic1/zic4* locus is highly methylated[68]. Nonetheless, I detected insertion site polymorphism between the two long-separated groups of the medaka inbred line Hd-rR, suggesting that transposition of *Teratorn* occasionally occurs under natural conditions. Because of its big size, *Teratorn* usually imposes deleterious effects on host gene function, but less frequently, can modulate the activity of widely distributed enhancers of a developmental gene, creating a novel trait in a host. One example is the medaka *Da* mutant (homozygous viable) in which *Teratorn* disrupts the somite enhancers of the *zic1/zic4* locus without affecting the neural enhancer, leading to drastic changes in body shape and fins from larva to adult[41].

Surprisingly, *Teratorn* contains all essential genes shared by alloherpesvirus species with a similar genome configuration (Fig. 8c). I also found genes involved in promotion of cell proliferation, inhibition of apoptosis, and immunomodulation, which is typical of lymphotropic herpesviruses. Importantly, in the majority of *Teratorn* copies, the above key genes are maintained intact and transcribed at low levels in medaka hosts. Based on those findings, I propose that *Teratorn* was descended from a novel herpesvirus that has propagated in the medaka genome by acquiring the *piggyBac* transposon system (Fig. 19).

Besides accidental chromosomal integration, there are only two endogenous herpesviruses, human herpesvirus 6 (HHV-6) and tarsier endogenous herpesvirus[69,70], reported so far, out of >100 herpesvirus species. They belong to the *Betaherpesviridae* and have an array of telomeric repeats (TMRs) at their termini, which are identical to the telomere sequences (TTAGGG)[69]. These herpesviruses are specifically integrated into the telomeric region via homologous recombination through TMRs, presumably catalyzed by U94, an endonuclease-like gene homologous to parvovirus Rep[69,71]. However, it remains unclear whether these endogenous viruses are genuine genomic parasites, because the tarsier endogenous herpesvirus has accumulated deleterious mutations in ORFs[70] and HHV-6 has not yet reached a fixation state among human populations (approximately 1% of the human population)[71]. In contrast, *Teratorn* has acquired the ability of transposition using its own transposase and colonized in the genomes of medaka species keeping the viral genes intact. Although the mechanism remains a mystery, an ancestral

Teratorn was accidently created by fusion of the two elements in other organisms, and the virus form of *Teratorn* then invaded into the medaka lineage. Since one of the hallmarks of herpesviruses is to establish a life-long persistent infection inside hosts as a viral episome without integration[50], I reason that chromosomal integration is an adaptive consequence to escape from host immunity and ensure a stable transmission of their progeny across host generations. The absence of genes involved in nucleotide metabolism in *Teratorn* (e.g. thymidine kinase (TK), deoxyuridine triphosphatase (dUTPase) and uracil DNA glycosylase), which are utilized under massive virus propagation by modulating the nucleotide pool[50], is suggestive, because those genes might no longer be needed after endogenization. Taken together, *Teratorn* represents the first example of herpesvirus adaptation to intragenomic life cycle (Fig. 19).

Current life cycle of *Teratorn*. The entire life cycle of *Teratorn* is still largely unknown, although it behaves at leasts as an active transposon *in vivo*. However, given that most herpesvirus-related genes also appear functional, those genes might be coordinately utilized for its propagation together with the transposase gene. Indeed, there is circumstantial evidence that chromosomally integrated HHV-6 (ciHHV-6) can be reactivated, e.g. in ciHHV-6-positive cultured cells[72], healthy ciHHV-6-positive individuals and X-linked severe combined immunodeficiency (X-SCID) patients[71]. Reactivation of these HHV-6 genes appeared to be accompanied by the formation of circular viral DNA molecules via

excision of the telomeric t-loop[71,73]. *Teratorn* could also undergo similar processes during reactivation; excision mediated by the *piggyBac*-like transposase followed by genome replication and virion formation (Fig. 19b). However, I have not succeeded in full activation of *Teratorn*. 5-azacytidine treatment of medaka fibroblasts only caused moderate reactivation of viral genes (Fig. 17), and under this condition, I failed to detect virus proteins derived from *Teratorn* by western blot (major capsid protein, capsid triplex protein and envelope glycoprotein, Fig. 18). Given that all essential viral genes, especially genes involved in virion morphogenesis such as capsid proteins, DNA packaging terminase and capsid maturation protease, appear to be functional, it is still likely that, under as-yet-unknown conditions, *Teratorn* actually produces virus particles that infect new individuals. Similarly, virions have not yet been detected experimentally from *Polinton*/*Maverick* superfamily transposons, replicative large DNA transposons widespread among eukaryotic genomes[33,40], although they contain a set of intact virus-like genes required for virion formation, including A32-like DNA packaging ATPase, Ulp1-like cysteine protease and two capsid proteins[34,74].

Despite the potentially hazardous consequences of viral genes, there might be some benefits to the host having this large mobile element. One possible scenario is that *Teratorn* serves as a shield against an otherwise lethal virus, by inhibition of virus entry by receptor block, inhibition of functional virion assembly or establishment of immunotolerance. On the other hand, *Teratorn* shows characteristics of selfish genetic elements, since all herpesvirus essential genes and

transposase gene remain intact. Thus, I propose that *Teratorn* might serve two purposes; one is a guardian against exogenous virus infection, and the other is a selfish intragenomic parasite. Recently, it was shown that endogenous virophages in some protozoan genomes were reactivated by superinfection of giant viruses (preys of the virophages), which facilitates host survival and their own propagation[75,76]. *Teratorn* could undergo the similar response under the infection of exogenous viruses. In any case, further experimental efforts to detect virions will be necessary to understand the life cycle of *Teratorn* and the biological significance of the existence of *Teratorn* in the medaka genome.

Because of the transposition activity and intact herpesvirus genes, I propose that *Teratorn* is a mobile element that was created by a unique fusion of *piggyBac*-like transposon and herpesvirus and as such, acquired a "bivalent" life form, behaving both as a transposon and a virus. To know how general this phenomenon occurs, next I conducted the genomic survey of *Teratorn*-like elements in vertebrates.

Chapter 2

Widespread distribution of *Teratorn*-like elements in teleosts

Introduction

Viruses are the most abundant organismal entities on earth and have intimate or adverse relationships with host organisms, from symbiosis to infectious disease. Occasionally, viruses are integrated into chromosomes of germline cells and become a part of the host genome, being referred to as endogenous viral elements (EVEs)[23]. The impacts of EVEs on host organisms again vary from beneficial (e.g. immunity to exogenous viruses, novel gene regulatory networks, emergence of new traits such as placenta)[13,23,32,77–79] to detrimental effects (e.g. genomic parasites like transposable elements)[23,80,81]. Endogenous retroviruses (ERVs), constituting up to ~8% of the human genome, are the most common EVEs, since they have the chromosomal integration step during their life cycles[31]. Recent accumulation of genome sequencing data have revealed that non-retroviral viruses, too, to a lesser extent, can become EVEs[23,82–86]. However, since the life cycle of non-retroviral viruses in eukaryotes normally does not include the chromosomal integration step, integration mechanisms of most endogenous non-retroviral elements are thought to be rather passive (homologous recombination, accidental integration at double-strand breaks, or retrotransposon-mediated integration[23,82]). Indeed, there have been only a few reports of endogenous non-retroviral elements with nearly complete viral genomes to propagate[23].

In chapter 1, I described the mobile element, *Teratorn*, that exists as a fusion of a *piggyBac* transposon and a herpesvirus. Importantly, *Teratorn* still

retains the transposition activity and encodes a set of intact herpesvirus genes. Since herpesviruses usually don't have the chromosomal integration step in their life cycles, essentially nothing is known about endogenous herpesviruses that retain the propagation capacity in host genomes. Thus, I suppose that *Teratorn* represents the first example of herpesviruses that had transformed into an intragenomic propagator, via gaining the *piggyBac* transposition system. From these findings, I became interested in the generality of transposon-herpesvirus fusions.

Here I performed a comprehensive search for *Teratorn*-like elements in other species and analyzed their phylogeny. Analysis in the genus *Oryzias* suggests that *Teratorn* has widely colonized in this lineage, as a fused form of *piggyBac* and herpesvirus. Furthermore, genomic survey against all vertebrates showed that *Teratorn*-like elements are widely distributed among teleost fish species in the endogenized form, some of which were certainly mediated by *piggyBac* transposon. I propose that *Teratorn*-like elements form a new herpesvirus genus, *Teratornivirus*, that tends to propagate in the host genome with the help of a transposon.

Results

Widespread colonization of *Teratorn* in the genus *Oryzias*. As a first fstep to know the distribution of *Teratorn*, I performed genomic survey against Japanese medaka strains. The Japanese medaka, *Oryzias latipes*, is known to exist as four major local populations, i.e. northern Japanese, southern Japanese, eastern Korean and Chinese/western Korea[87]. Whole-genome shotgun sequencing data are currently available for five medaka inbred strains established from each local population (Hd-rR from southern Japanese, HNI and Kaga from northern Japanese, HSOK from eastern Korean and Nilan from Chinese / western Korean populations; Fig. 20a)[64]. Taking advantage of these genome resources, I first searched for *Teratorn*-related sequences and performed a comparative analysis in these strains. Preliminary mapping of Illumina reads to the *Teratorn* sequence of Hd-rR suggested the presence of *Teratorn* in the genome of all strains. To estimate copy number, I mapped whole-genome shotgun reads of each medaka inbred strain to the reference medaka genome (Hd-rR) (see Methods). I then calculated the putative copy number, by dividing the average coverage of all *Teratorn* genes by that of all host genes, assuming that the copy number is proportional to the depth of coverage. This calculation revealed that the copy number varies greatly depending on the strain (a few to 60) (Fig. 20b). The phylogenetic tree of *Teratorn* was then constructed based on the deduced consensus sequence of each species, and the tree topology was found to be identical to that of the host medaka species (Fig. 20c).

Furthermore, I calculated F_{ST} value for every pair of strains using popoolation2[88], based on the variants among *Teratorn* copies in each medaka inbred strain, assuming multiple *Teratorn* copies within a single genome as a population. I found that, for both subtype 1 and subtype 2, F_{ST} tends to be higher as the evolutionary distance between the two strains gets larger (Figure 20d). Indeed, topology was found to be identical between the trees inferred from the pairwise F_{ST} values and that of host phylogeny (Figure 20e). These data imply the vertical inheritance of *Teratorn* from the common ancestor of Japanese medaka, along with occasional transposition and propagation.

Next I expanded my focus to the genus *Oryzias*. There are more than 20 medaka related species inhabiting the South East Asia, from India to Japan. They are subdivided into three groups, *latipes* species group, *javanicus* species group and *celebensis* species group (Fig. 21a,b)[89]. PCR analysis of five herpesvirus core genes (DNA polymerase, DNA helicase, ATPase subunit of terminase, major capsid protein and envelope glycoprotein) and transposase gene of *Teratorn* in 13 medaka-related species revealed that *Teratorn* exists in the *latipes* and *javanicus* species groups, but not in the *celebensis* species group (Fig. 21b). Phylogenetic analysis suggests that the phylogeny of *Teratorn* and transposase genes is almost the same as that of the host species (Fig. 21c). In addition, BAC screening and sequencing of the large part of *Teratorn* in the species of *javanicus* species group revealed that the configuration of *Teratorn*, particularly the location of *piggyBac* transposase, is conserved among *Oryzias* genus (Fig. 21d,e). Taken together, these

data demonstrate that *Teratorn* have widely colonized in the genomes of *Oryzias* genus, presumably using the *piggyBac* transposition machinery. Similar topology of phylogenetic trees between host genes and *Teratorn* genes implies that *Teratorn* was vertically descended from the common ancestor of *Oryzias* genus. However, considering the absence of orthologous copy among the three medaka inbred strains (Hd-rR, HNI and HSOK), high sequence similarity among copies inside one species and the absence of degraded copies for all species except for *O. luzonensis*, the possibility that recent invasion of *Teratorn* into each species took place cannot be ruled out.

***Teratorn*–like elements are widely distributed in teleost genomes.** To further address the distribution of *Teratorn* in other organisms, I performed blast search against publically available vertebrate genome dataset. Tblastn search of 13 herpesvirus core genes showed that *Teratorn*-like elements were detected in at least 19 of the 67 teleost fish species (E-value < 10^{-3}, more than 9 of the 13 core genes, Fig. 22a). In contrast, I didn't get any significant hits against amphibian, chondrichthyes or sarcopterygi genomes, indicating that *Teratorn*-like elements populate only in teleosts. Interestingly, tblastn search using the sequences of another distantly-related alloherpesvirus species, Cyprinid herpesviris 3 (CyHV-3), didn't provide any positive hits other than *Teratorn*-like elements, suggesting that *Teratorn*-like elements are the only herpesvirus integrated in teleost genomes. Phylogenetic analysis based on concatenated amino acid sequences of four obtained

herpesvirus core genes demonstrated that those *Teratorn*-like elements were closely related with each other and form a cluster inside *Alloherpesviridae* (Fig. 22b). Indeed, almost all *Teratorn*-like elements showed high sequence similarity to the medaka *Teratorn* (~70% identity in nucleotide sequence), except for one from *K. marmoratus* (~30% identity in amino acid sequence).

The patchy distribution of *Teratorn*-like elements among teleost fishes might result from multiple independent endogenization events in teleosts. The phylogeney of *Teratorn*-like elements exhibits little correlation with that of the hosts (Fig. 22). However, the local consistency of phylogeny is observed at the tip of some branches in cichlids and medaka, suggesting vertical inheritance in these fish groups.

For some species, I was able to obtain contigs or scaffolds covering over 100-kb regions with *Teratorn*-like elements (Fig. 23a,b). Gene annotation demonstrates that some of them certainly contain a set of intact virus essential genes, suggesting that *Teratorn*-like elements are maintained intact in those species. To exmained the possibility of intragenomic propagation, I estimated the copy number of *Teratorn*-like elements by mapping whole-genome shotgun reads of several teleost species to their reference genome (see Methods). I then calculated the putative copy number, by dividing the average coverage of herpesvirus core genes by that of all host genes, assuming that the copy number is proportional to the depth of coverage. I found that *Teratorn*-like elements are present in multiple copies in some teleost species (~18 copies in yellow croaker (*L. crocea*), ~14 copies in

nile tilapia (*O. niloticus*), ~8 copies in atlantic salmon (*S. salar*) and northern pike (*E. lucius*), ~6 copies in annual killifhsh (*A. limnaeus*); Fig. 23c), implying that they increased their copy number inside host genomes, like in medaka. Together, *Teratorn*-like elements are teleost-specific, widely distributed in teleost genomes, and retain the ability of propagation.

***piggyBac*-herpesvirus fusion broadly occurred in the teleost lineage.** In several species (yellow croaker (*L. crocea*), nile tilapia (*O. niloticus*), ocean sunfish (*M. mola*), turquoise killifish (*N. furzeri*), annual killifish (*A. limnaeus*) and Atlantic salmon (*S. salar*), I identified *piggyBac* transposase genes next to the *Teratorn*-like elements (Fig. 22, 23a,b, magenta), suggesting that the fusion of the two mobile elements occurred similar to medaka. To further explore this possibility, I first focused on yellow croaker (accession: GCA_000972845.1)[90] and nile tilapia (accession: GCA_001858045.2), since these fish seem to have high copy number of *Teratorn*-like elements (Fig. 23c). I searched for genomic neighborhoods in all scaffolds that harbor *piggyBac* transposase genes and compared to the reference sequence of *Teratorn*-like element.

For yellow croaker, I obtained 40 scaffolds that include the *piggyBac* (Fig. 24a). Among the 17 obtained scaffolds that showed high sequence similarity to the reference *piggyBac* transposase (i.e. their p distance to the reference transposase sequence was <0.041), five scaffolds contained the herpesvirus-like sequences next to the *piggyBac* transposase gene, in the same configuration to the reference as

shown in Fig. 24a. For 11 other scaffolds, I was unable to determine their neighborhoods because their length was too short. For one remaining scaffold (KQ041719.1), no herpesvirus-like sequence was found around the transposase gene, presumably caused by misassembly because the coverage of shotgun reads is peculiarly high at its flanking region (>100-fold to the transposase region, data not shown). By contrast, for the remaining 23 contigs (their p distance was >0.069), their transposase genes are not connected to any herpesvirus-like sequences (Fig. 24b). Indeed, phylogenetic analysis revealed that *piggyBac* copies adjacent to the herpesvirus-like sequence were clustered together (Fig. 24c). Thus, it is highly likely that the fusion of a particular *piggyBac* transposon and herpesvirus occurred in yellow croaker.

I next focused on nile tilapia (*O. niloticus*). There are two types of *Teratorn*-like element in the nile tilapia genome, referred to as '*Teratorn*-like 1' and '*Teratorn*-like 2'. Those two elements are distantly related to each other (less than 80% nucleotide identity), and only *Teratorn*-like 1 is linked to the *piggyBac* transposon, existing at a high copy number (*Teratorn*-like 1, ~12 copies; *Teratorn*-like 2, ~2 copies) (Fig. 23c, 25a). I thus focused on *Teratorn*-like 1 for further analyses. Like in the analysis of yellow croaker, I searched for genomic neighborhoods for all copies of the same *piggyBac* and compared with the reference sequence of *Teratorn*-like 1 element (MKQE0100015.1 : 40163260–40398786), in which the herpesvirus-like sequence is flanked by the *piggyBac* sequences at both ends (Fig. 25a). This configuration was indeed observed in many of the identified

elements (Fig. 25b). Importantly, in those elements, the regions flanked by the transposons are highly homologous, i.e. the region from the 5' TIR of the 5'-side *piggyBac* to the 3' TIR of the 3'-side *piggyBac*, while no sequence similarity was found outside (Fig. 25c), indicating the transposition of *Teratorn*-like 1 mediated by *piggyBac*. Therefore, I conclude that *Teratorn*-like 1 in nile tilapia is indeed the fusion of *piggyBac* and herpesvirus, which allows for transposon-mediated propagation in the host genome.

I conducted the same analysis for the rest of the fishes having *Teratorn*-like elements in their genomes. In ocean sunfish and turquoise killifish, only two and one regions were identified that contain a long range of *Teratorn*-like sequences together with a specific type of *piggyBac*-like transposase (Fig. 23b). On the other hand, for annual killifish and Atlantic salmon, *piggyBac* copies that are in close proximity to *Teratorn*-like elements are not monophyletic (Fig. 26), and the spatial relation between *piggyBac* and herpesvirus-like sequence varies among copies. Thus, I was unable to conclude that *Teratorn*-like element in those two species experienced the fusion event. Intriguingly, *piggyBac* transposase genes connected with *Teratorn*-like elements are polyphyletic among teleosts, suggesting the independent and multiple fusion events in these teleost fishes (Fig. 27).

Discussion

In this chapter, I have extended the analysis outside medaka to address the distribution and form of existence of *Teratorn*-like elements in other organisms. Database search against all vertebrate genomes demonstrated that *Teratorn*-like elements are not restricted to the genus *Oryzias*, but rather widespread in specific lineages of teleosts.

Evolutionary history of *Teratorn* in the genus *Oryzias*. Genomic survey in the genus *Oryzias* showed that *Teratorn* exists in the fused form in *latipes* and *javanicus* species groups, and its phylogeny is almost consistent with that of their hosts. In addition, the structure of *Teratorn*, especially the spatial relation with the *piggyBac* transposase gene, is conserved among them. Thus, it might be possible that *Teratorn* invaded the genome of the common ancestor of *Oryzias* genus, as a fused form with *piggyBac*, and has been vertically transmitted in this lineage. In this case, the absence of *Teratorn* in *celebensis* species group can be explained by its accidental loss sometime after divergence from the two species groups. However, given the absence of orthologous copy between any pair of medaka strains (Hd-rR, HNI and HSOK), as well as the absence of degraded copies for almost all species, the possibility of recent horizontal transfer of *Teratorn* at each lineage cannot be ruled out. Further research is needed to uncover the origin and the history of *Teratorn* in the genus *Oryzias*.

***Teratornivirus*, a new herpesvirus genus, that tend to integrate into host genomes.** Genomic survey demonstrated that *Teratorn*-like elements are distributed in a variety of teleost fish species. Although I can not exclude the possibility that positive hits in genomic database search can be a result of contamination of exogenous virus DNA into draft genomes. However, the following facts support that at least some are genuine genomic integrants. First, I frequently observed the link between herpesvirus-like sequences and other endogenous genomic regions in contigs or scaffolds, suggesting insertion. Second, some *Teratorn*-like elements are interrupted by various types of transposons. Finally, I sometimes observed ORF degradation in *Teratorn*-like sequences, which is unlikely in exogenous viruses (Table 3).

Teratorn-like elements identified in this study are phylogenetically close to each other, forming a single cluster inside *Alloherpesviridae*. To date, four genera have been established in *Alloherpesviridae*; *Batrachovirus*, *Cyprinivirus*, *Ictalurivirus* and *Salmonivirus*[91]. However, given its evolutionary distance from all those genera and broad distribution in teleosts, the group of herpesviruses that include the *Teratorn*-like elements should be given a new genus name and I tentatively propose "*Teratornivirus*".

This novel herpesvirus genus is characterized by high tendency for endogenization, and indeed my database search failed to find endogenous herpesviruses in alloherpesvirus species other than *Teratornivirus*. Furthermore,

some of *Teratornivirus* sequences are present in multiple copies in host genomes, suggesting that they can also propagate inside a host, like in medaka. Then, what character made *Teratorniviruses* broadly endogenized? My analyses imply that the acquisition of the *piggyBac* transposon system has contributed to their propagation in host genomes. In particular, *Teratornivirus* in yellow croaker (Fig. 24), nile tilapia (Fig. 25), turquoise killifish and ocean sunfish (Fig. 23b) are likely to be a genuine fusion form with the *piggyBac* transposon, and thus the hijack of the transposon system by *Teratornivirus* is not restricted to the medaka species. However, for other species, I failed to find out clear evidence for the fusion of *Teratorniviruses* with transposons. This might be a result of the failure of identification of cryptic transposons fused with *Teratorniviruses*, due to incomplete assembly of genome data. However, it is also possible that some other unique characteristics of *Teratornivirus* (e.g. cell tropism, possession of telomeric repeats, life cycle) might facilitate initial endogenization and that *piggyBac* mainly contributes to their intragenomic propagation. In any case, unraveling the terminal region of each *Teratorn*-like element, as well as characterization of their life cycle (e.g. detection of virions), will be required to further understand how *Teratornivirus* became endogenized in various fishes.

Conclusion and General discussion

In my doctoral thesis, I have characterized a novel mobile element, *Teratorn*. In chapter 1, I demonstrated that *Teratorn* was created by a unique fusion of *piggyBac* transposon and herpesvirus genome, based on the result that it contains *piggyBac* transposon machinery (transposase and TIRs) and all virus essential genes. Intriguingly, *Teratorn* is capable of transposition and maintains all virus genes intact. Thus, *Teratorn* has been descended from a unique herpesvirus that adopted the intragenomic life style by gaining *piggyBac* transposon machinery. In chapter 2, I reported the widespread distribution of *Teratorn*-like elements in the genomes of teleost fish species, forming a novel genus among *Alloherpesviridae*. Thus I propose a new genus, *Teratornivirus*, for this family. Importantly, some of the identified *Teratornivirus* species may be a genuine fusion of *piggyBac* transposon and herpesvirus like in medaka, suggesting the generality of herpesvirus-*piggyBac* fusion. Taken together, I propose that fusion with DNA transposon is a driving force for the intragenomic propagation of the herpesvirus species of this genus.

Because of the absence of chromosomal integration machinery, almost all viruses, except for retroviruses, were thought to have lost propagation capacity after endogenization. Indeed, except for ERVs, there have been only a few reports of EVEs that behave as genomic parasites. Such examples are endogenous pararetroviruses and single-stranded DNA viruses. The former constitutes up to ~1% of the genome of some plants, although chromosoal integration is not a

prerequisite[83,86], and the mechanism underlying chromosomal integration and propagation is unknown. The latter is present in tens to hundreds of copies in some species (e.g. geminiviruses in plants and fungi, circoviruses and parvoviruses in animals) and is thought to integrate via RC-Rep endonuclease, an enzyme involved in replication of virus DNA, although this is not fully confirmed[84,85]. Thus, I propose that *Teratorn* (and, some *Teratornivirus* species) is the first endogenous non-retroviral element that had shifted into the intragenomic lifestyle to promote propagation in host genomes.

So far, *Teratorn* is the first virus that experienced the fusion with a cut-and-paste DNA transposon in eukaryotes. However, *Teratorn* cannot be the only example of fusion of two distinct mobile elements in the network of eukaryotic MGEs. For example, several studies reported the insertion of host-derived transposons into virus genomes (e.g. baculovirus[92,93], poxvirus[94] and pandoravirus[95]); thus, emergence of functional fusion could occur on rare occasions. Indeed, all viruses have the potential to shift into the intragenomic life style, if they acquire an integration system from other sources. In this context, *Polintons* (also known as *Mavericks*) is of particular interest, belonging to a group of replicative large DNA transposons (10–25 kb) widely distributed in eukaryotes[33,40,96]. Recent phylogenetic studies suggested that *Polintons* have evolutionary links with other mobile genetic elements (e.g. adenoviruses, nucleocytoplasmic large DNA viruses, virophages, *Polinton*-like viruses, linear plasmids and bacteriophage) and serve as a hotbed for recombination leading to a change in their life cyle[33–35,37]. In this scenario,

Polinton was initially emerged via the acquisition of a *Ginger 1*-like DNA transposon by a bacteriophage (*Tectivirus*), at the onset of eukaryogenesis. Then, after that, some of them have returned back to virus life forms, concurrent with the loss of integrase gene and/or expansion of their genomes to form a large part of current eukaryotic DNA viruses[24,34,36]. Furthermore, recombination events between distantly related viruses have been reported, such as chimeric viruses (chimera of ssDNA and ssRNA(+) virus) and bidnavirus (chimera of parvovirus, *Polinton*, reovirus and baculovirus)[24,34,97,98]. Thus, the recombination and fusion between mobile elements of distinct classes are more frequently than previously thought, leading to the diversification of mobile genetic elements (Fig. 28). My doctoral research has provided concrete evidence for this concept. Further identification of novel and peculiar mobile genetic elements will provide further insights into mechanisms underlying their diversification and evolution of MGEs.

Materials and Methods

Fish strains

Hd-rR, d-rR and HNI inbred strain of *Oryzias latipes* were maintained in our laboratory. Fish strains of medaka-related species were obtained from the laboratory stocks maintained at Niigata University and National Institute for Basic Biology. Lists of each species and their original collection sites are described in Table 4. All experimental procedures and animal care were performed according to the animal ethics committee of the University of Tokyo.

Screening and sequencing of BAC clones

Teratorn insertion sites were identified by screening of the 5' and 3' flanking regions by blastn of *Teratorn* terminal sequences against the public genome assembly of medaka Hd-rR inbred strain[45], followed by mapping of those obtained sequences to the genome of another medaka inbred strain HNI to know which of the 5' and 3' flanking sequences are derived from the same loci. For the 12 *Teratorn* copies with identified integration sites, BAC clones that were PCR positive for both *Teratorn* ends were screened from the medaka Hd-rR BAC library[99]. For the screened six individual *Teratorn* copies, BAC DNA was purified by QIAGEN Large-Construct Kit (QIAGEN). Sequencing of BAC clones was implemented using PacBio RS-Ⅱ. Assembly was carried out using HGAP pipeline.

For medaka related species (library name : IMBX for *O. dancena*; OHB1

for *O. hubbsi*; OJV1 for *O. javanicus*; LMB1 for *O. luzonensis*), BAC clones were screened by PCR of the four *Teratorn* genes (*piggyBac*-like transposase, membrane glycoprotein, DNA polymerase and DNA packaging terminase) BAC DNA was then purified using NucleoBond® Xtra BAC (MACHEREY-NAGEL). Sequencing library was prepared by sonicating to a size ranging from 800 to 1200bp, followed by adaptor ligation and amplification using KAPA Hyper Prep Kit (KAPA BIOSYSTEMS). Paired-end sequencing of the prepared libraries were executed on the illumina Miseq platform. After filtering out low-quality reads by trimmomatic v0.33[100], *de novo* assembly was carried out by CLC Genomic Workbench (https://www.qiagenbioinformatics.com/).

Gene annotation

Gene annotation was initially carried out by the GeneMarkS web server[49]. For genes that start with codons other than "ATG", prediction by ORF finder (http://www.ncbi.nlm.nih.gov/gorf/gorf.html) was used instead. If adjacent multiple ORFs seemed to be derived from a single gene (i.e. different portions of the same gene were obtained as blastp output), gene annotation by GENSCAN web server[48] was used to generate a more plausible gene model including introns.

Identification of subtype 2 *Teratorn*

Subtype 2 *Teratorn* was identified by blastn search of subtype 1 *Teratorn* sequence against the updated version of Hd-rR genome assembly

(http://utgenome.org/medaka_v2/#!Top.md). The putative full-length subtype 2 *Teratorn* sequence was constructed by conjugating partial sequences of two contigs (ctg7180000008209 : 1–30344, ctg7180000008207 : 2815545–2982619), followed by calling consensus sequence by Bcftools[101], based on the Illumina whole-genome shotgun short reads (see below).

Multiple alignment of active *piggyBac* transposase genes

Amino acid sequences of *piggyBac* superfamily transposase genes (*piggyBac* in *T. ni*, *Uribo2* in *X. tropicalis*, *yabusame-W* in *B. mori*, *piggyBat* in *M. lucifugus*, *AgoPLE* in *A. gossypii* and *PLE-wu* in *S. frugiperda*) were downloaded from GenBank. Those transposases were aligned using Clustal W with default parameters[102]. The multiple alignment was visualized using TrimAl[103].

Search for neighboring region to the *piggyBac* transposase

Contigs or scaffolds including the transposase gene were screened by blastn search of the transposase gene, followed by extraction of flanking region by BEDtools (upstream 25 kb and downstream 25 kb region for medaka, 60 kb flanking region for yellow croaker, and 20 kb flanking region for nile tilapia). Gene order around the transposase genes were then assessed by displaying dot plot matrix between the reference *Teratorn*-like element and the obtained contigs, using mafft on line server[104]. Neighbor-joining analysis were carried out based on the partial region of the transposase gene using Tamura-Nei method as substitution model with 1000

bootstrap replicates, without considering evolutionary rate difference among sites.

Plasmid construction for the transposition assay

Indicator plasmid was constructed as follows. First, the puromycin-resistant gene cassette was extracted from the pMXs-puro retroviral vector by digesting with *BamH*I and *Xho*I. In parallel, the AcGFP expression cassette was amplified by PCR. These two DNA fragments were then conjugated using in-fusion HD cloning kit (TaKaRa). This puro^{R}-AcGFP fragment was further conjugated with *Teratorn* TIRs at both sides by joint-PCR, and was subcloned into pCR-Blunt Ⅱ-TOPO (Invitrogen). Finally, *Teratorn* internal TIR was inserted into the boundary of the 5' TIR and the puro^{R}-AcGFP cassette so as to mimic the endogenous *Teratorn* structure. Helper plasmid was constructed by RT-PCR of the *Teratorn* transposase followed by subcloning into pCSf107mT vector[105].

Transfection, excision assay and integration assay

HEK293T cell line was a kind gift from Prof. Yoshinori Watanabe (University of Tokyo). Cells were maintained in DMEM with 10% FBS and 100 U/ml penicillin and 100 μg/ml streptomycin (Gibco) at 37℃ with 5% CO_2. The day before transfection, cells were seeded in 6-well plates to achieve 70–90% confluent on the next day. Lipofectamine 3000 reagent (invitrogen) was used to co-transfect cells with 1500 ng of helper plasmid and 2500 ng of indicator plasmid per well. For the excision assay, 40% of cells were passaged to a new dish, cultured for about two

days to become fully-confluent and the plasmid was extracted by the Hirt method[106]. PCR was carried out using primers that flank the transposon cassette of the indicator plasmid. In the integration assay, 40% of transfected cells were passaged to a new dish and cultured in DMEM with 5.0, 7.5 or 10.0 μg/ml of puromycin for about two weeks. Medium was changed every two or three days. Surviving colonies were isolated by peeling off and sucking up with a pipette and transferred to a new dish. When cell clones grew in sufficient number, genomic DNA was isolated using Wizard Genomic DNA extraction kit (Promega) for Southern blotting and PCR-based integration site determination.

Southern blotting

5–10 μg of genomic DNA extracted from medaka adult whole body or HEK293T cells were digested with appropriate restriction enzymes for at least one hour and precipitated with ethanol. Digested DNA was resolved by gel electrophoresis with 25–30 V overnight, and transferred to a Hybond-N+ nylon membrane (GE Healthcare). AlkPhos Direct Labeling and Detection System with CDP-*Star* (GE Healthcare) was used for hybridization and signal detection. Purified PCR products of *Teratorn* terminal sequences were used as hybridization probes to visualize individual *Teratorn* insertions.

(EPTS)LM-PCR

Extension primer tag selection linker-mediated PCR ((EPTS)LM-PCR) was carried

out as previously described[58]. First, 1.5 μg of genomic DNA from HEK293T cells was digested with *Hae*III or *Apo*I overnight. After precipitation with ethanol, a biotinylated primer specific to *Teratorn* terminal sequence was used for tagging the transposon insertion sites. In this step, Phusion High-Fidelity DNA Polymerase (Thermo) was used for DNA extension (98℃ for 3 min, 68℃ for 30 min, and 72℃ for 30 min). After purifying the PCR products using the Wizard® SV Gel and PCR Clean-Up System (Promega), biotin-tagged DNA was isolated by incubating with Dynabeads® M280 Streptavidin (Thermo) for one hour, followed by twice washing out of non-target DNA. Finally, double-stranded oligonucleotide phosphorylated linkers were ligated at 16℃ overnight. PCR with primer pairs specific to *Teratorn* terminal sequence and linker oligonucleotide were carried out to amplify *Teratorn* insertion sites.

Multiple alignment of herpesvirus genes

Multiple alignment of amino acid sequences of DNA packaging terminase, capsid maturation protease, major capsid protein and subunit 2 capsid triplex protein were constructed by PROMALS3D[107]. Sequences of each species were downloaded from Genbank. Catalytic centers of terminase and protease gene was characterized based on Rao VB. and Feiss M.[59], Selvarajan Sigamani S. *et al*[60], and Cheng H. *et al*[61] Domains of major capsid protein was characterized based on Huet A. *et al*[62].

Variant calling and annotation. Reference medaka genome data was

reconstructed as follows, in an attempt to map all *Teratorn*-derived reads to the reference *Teratorn* sequences. *Teratorn* was masked from the public medaka draft genome[45] by blastn of both *Teratorn* subtypes, followed by maskFastaFromBed command of BEDtools. Then, the genome was conjugated with the sequences of *Teratorn*. Illumina whole-genome shotgun read data was downloaded from DDBJ Sequence Read Archives (accession : DRR002213). After filtering out low-quality bases and adapter sequences from short reads by trimmomatic v0.33[100], reads were aligned to the reconstructed reference genome using BWA (Burrows-Wheeler Aligner)-MEM[108], with the default parameter settings. Removal of PCR duplicates was carried out by Picard (http://picard.sourceforge.net). Local realignment was executed using RealignTargetCreator and IndelRealigner tools in GATK, with the default parameters[109]. Variant calling was performed by UnifiedGenotyper tool in GATK, with the following parameters (ploidy, 30 for *Teratorn* subtype 1 and 5 for subtype 2; stand-call-conf, 30; stand-emit-conf, 20; glm, BOTH). Variant annotation was implemented using SnpEff, with the default parameters[110].

Expression analysis of *Teratorn* genes in medaka

Total RNA was isolated from 5 dpf (days post fertilization) embryos and adult fish tissues (brain, liver, fin, muscle, gut, ovary, testis) of medaka d-rR strain, using ISOGEN (Nippon Gene), according to the manufacture's protocol. After removal of genomic DNA by DNaseI digestion (Invitrogen), cDNA was synthesized using SuperScript III (Invitrogen). RT-PCR was carried out using Phusion DNA

Polymerase (Thermo Fisher) for 40 cycles.

Test of reactivation of *Teratorn* genes by chemical administration

Medaka embryonic fibroblast cell line was previously established in our laboratory. Cells were maintained in L-15 medium (Gibco), supplemented with 15% FBS (biosera), 100 U/ml penicillin and 100 μg/ml streptomycin (Gibco) at 30℃. For reactivation, cells were administered with 2 μM of 5-azacytidine (Sigma) for five days. For a subset of samples, 500 ng/ml of 12-*O*-Tetradecanoylphorbol 13-acetate (Sigma) and 3 mM of sodium butyrate (Sigma) were also administered for the last 2 days. This experiment was repeated for three (no chemical administration, 5-azacytidine only) or four (TPA + N-butyrate, TPA + N-butyrate + 5-azacytidine) times. Total RNA extraction and cDNA synthesis were carried out as described above. qRT-PCR was carried out with the Stratagene mx3000p system (Agilent), using the THUNDERBIRD SYBR qPCR Mix (ToYoBo). *β*-actin was used as an internal control. Ratio of molar concentration of *Teratorn* genes relative to *β* -actin was quantified based on standard curves generated from purified PCR products, using Qubit® 2.0 Fluorometer (Thermo Fisher) to measure mass concentration.

Statistics

In the qRT-PCR experiment described above, no statistical method was used to predetermine sample size. Sample sizes were based on previously reported experiments which are similar to the present study. No data were excluded from the

analysis. For genes which met the assumption of normal distribution (all genes except for major capsid protein, pim, and terminase, Shapiro-Wilk test, $p<0.05$), statistical significance was tested by one-sided Welch Two Sample t-test. All statistical analyses were performed using R software (version 3.2.4).

Antibodies

Partial sequences of major capsid protein (amino acids 934–1074), capsid triplex protein (amino acids 51–170) and membrane glycoprotein (amino acids 711–830) were subcloned into pGEX4T-1 vector (GE Healthcare). After culture of transformed bacteria, GST-tagged protein fragment was purified using Glutathione Sepharose 4B (GE Healthcare). Polyclonal antibodies were raised by immunization of mouse with the GST-tagged truncated proteins (100 μg) for about two months, followed by extraction of serum.

Western blotting

Fore reactivation of *Teratorn* genes, medaka fibroblast cells were administered with 5-azacytidine (0.0 μM, 1.0 μM, 1.5 μM, 2.0 μM) for five days. For a subset of samples, 500 ng/ml of 12-*O*-Tetradecanoylphorbol 13-acetate (Sigma) and 2 mM of sodium butyrate (Sigma) were also administered for the last 2 days. For positive control, HEK293T cells transfected with the plasmids that express the GFP-tagged antigen protein fragment were used. Cells were lysed with 1x SDS buffer, separated by polyacrylamide gels (5.5% for major capsid proteins and membrane glycoprotein,

and 10.0% for capsid triplex protein, respectively) and transferred onto polyvinylidene difluoride membranes (Millipore). After blocking with 5% skim milk for 1 h at room temperature, membranes were incubated with primary antibodies (anti-major capsid protein, anti-capsid triplex protein and anti-membrane glycoprotein, 1:2000 dilution) for overnight at 4℃. After several washing, membranes were incubated with secondary antibody (Anti-Mouse IgG HRP-Linked Whole Ab Sheep (GE Healthcare, NA931V), 1:2500 dilution) for 1 h at room temperature. Protein bands were visualized using the ECL Select Western Blotting Detection Reagent (GE Healthcare) and detected by ImageQuant (GE Healthcare).

Screening of partial sequences of *Teratorn* genes in *Oryzias* genus

Partial sequences of five herpesvirus genes (DNA polymerase, DNA helicase, major capsid protein, membrane glycoprotein and ATPase subunit of terminase) were screened from genomic DNA of each medaka related species by PCR, using degenerate primers constructed from the multiple alignment of *Teratorn*-like sequences in several teleost fishes (*S. salar*, *O. mykiss*, *E. lucius*, *O. niloticus*, *P. nyererei*, *C. semilaevis*). For genes in which degenerate primers didn't work, primers designed from sequences of subtype 1 and subtype 2 *Teratorn* of *O. latipes* were used instead. *piggyBac*-like transposase gene was screened by using primers that were designed from *O. latipes* genome.

Comparative analysis of *Teratorn* in five medaka inbred strains

Reference genome data reconstruction, mapping of Illumina data and copy number estimation are performed as described below (see "Copy number estimation"). Consensus genomic sequences were constructed for each inbred strain using Bcftools[101], based upon the mapping data. For the phylogenetic analyses, sequenes of 18 host genes[111] and 16 *Teratorn* genes were extracted from the reconstructed genome data of each inbred strain. Pairwise F_{ST} was calculated for sliding windows and step size of 400, using popoolation2[88]. Pool size was set as follows, based on the estimated *Teratorn* copy number; Subtype 1 : 30 for Hd-rR, 25 for HNI, 40 for Kaga, 60 for HSOK, 15 for Nilan; Subtype 2 : 5 for Hd-rR, 15 for HNI, 7 for Kaga, 5 for Nilan.

Search for *Teratorn*–like elements in other species

Tblastn search of 13 herpesvirus core genes of *Teratorn* was carried out against all available teleost genomes with default parameters. In addition, tblastn of the four genes (DNA polymerase, DNA helicase, DNA packaging terminase and major capsid protein) was performed against amphibians, chondrichthyes or sarcopterygi in the web browser. Contigs or scaffolds that include a series of herpesvirus-like sequences were screened as follows. First, location of the 13 herpesvirus core genes was identified by tblastn. After merging the genomic loci which are within 40 kb of one another, sequences of the defined region and the flanking 40 kb region were extracted from the draft genome by BEDtools.

Consensus sequence calling of *Teratorn*–like elements

In yellow croaker, preliminary consensus sequence was constructed using Unipro UGENE[112], based on multiple alignment of seven contigs that include the relatively long range region of *Teratorn*-like sequence (>40 kb). Then the consensus sequence was created by mapping of the illumina whole-genome shotgun reads, followed by consensus sequence calling by vcftools[101]. In nile tilapia, all gained *Teratorn*-like sequences were aligned to the longest one (MKQE01000015 : 40163260-40398786) using BWA (Burrows-Wheeler Aligner)-MEM[108], followed by extraction of consensus sequence using Unipro UGENE.

Copy number estimation of *Teratorn* and *Teratorn*-like elements

Reference genome data was reconstructed as follows. First, *Teratorn* or *Teratorn*-like sequences were masked from the genome by blastn, followed by maskFastaFromBed command of BEDtools[113]. Then, the masked genome data was conjugated with the *Teratorn* or *Teratorn*-like sequences. Illumina whole-genome shortgun read data were downloaded from DDBJ Sequence Read Archives (accession number are listed in Table 6). After filtering out low-quality reads by trimmomatic v0.33, reads were aligned to the reconstructed reference genome using BWA (Burrows-Wheeler Aligner)-MEM[108], with the default parameter settings. After converting the output sam files into bam files, coverage at each position in all coding region was counted by coverageBed command of BEDtools with –d option. Copy number of *Teratorn*-like sequences were then calculated by dividing the

average coverage of 15 herpesvirus genes (DNA polymerase, DNA helicase, primase, ATPase subunit of terminase, major capsid protein, membrane glycoprotein, capsid triplex protein, capsid maturation protease, ORF34, ORF37, ORF44, ORF54, ORF56, ORF60, ORF64) by the average coverage of the rest of all nuclear genes (all species except for *N. furzeri*) or the partial region of 19 nuclear genes (*N. furzeri*), assuming that copy number is proportional to the depth of read coverage.

Phylogenetic analysis

Nucleotide or protein sequences were retrieved either by blast search, PCR screening or downloading from GenBank. Multiple alignments were built up using MUSCLE in MEGA6 package. Poorly aligned region were removed either by trimAl or manual procedure. Neighbor-joining or maximum-likelihood analysis was carried out using MEGA6, while bayesian inference was performed using MrBayes3.2. Gene(s), species, and parameters utilized in these analyses are listed in Table 5.

Primers

Primer sequences used in this study are listed in Table 7.

Data availability

Nucleotide sequence data and gene annotation of *Teratorn* and *Teratorn*-like elements are available in the online version of the published paper at https://www.nature.com/ncomms. (DOI: 10.1038/s41467-017-00527-2) Sequence of

subtype 1 *Teratorn* (73I9) has been submitted to the DDBJ/EMBL/GenBank databases (Accession No: LC199500).

Figures

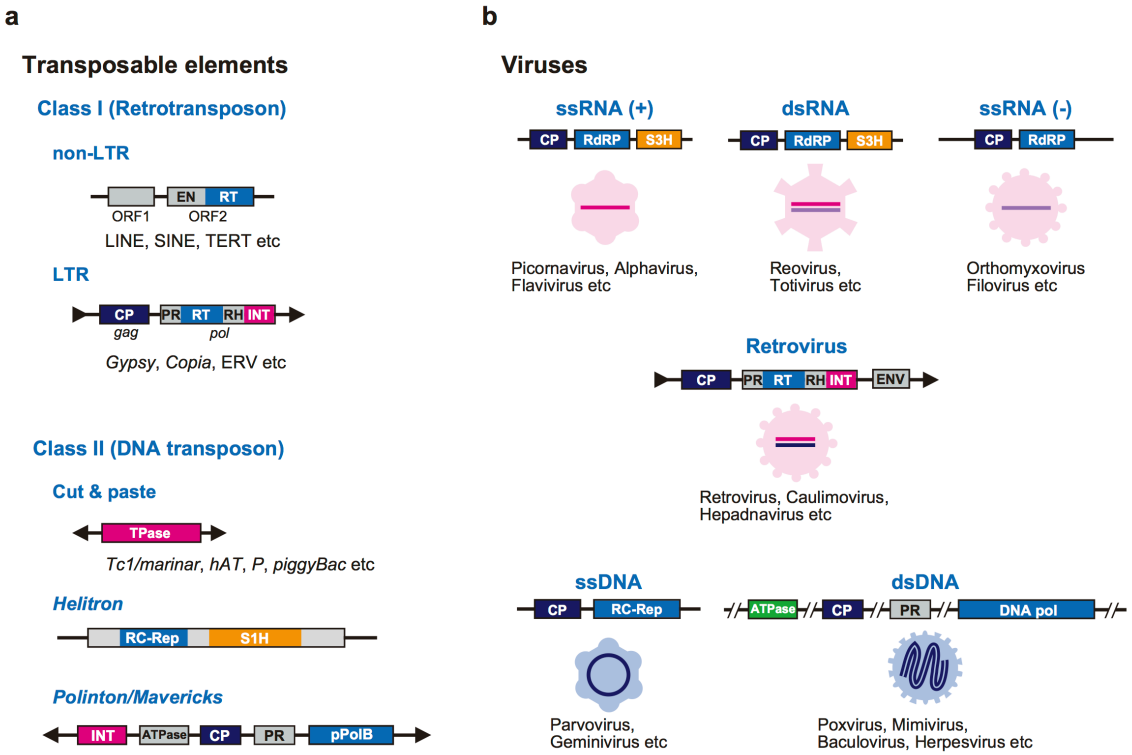


Figure 1 | Classification of mobile genetic elements in eukaryotes. (a) Transposable elements are classified into two major groups, according to the existence of RNA intermediate during transposition. Class I elements (retrotransposons) include long terminal repeat (LTR) and non-LTR elements, both of which transpose in a "copy-and-paste" manner. Class II elements (DNA transposons) include elements that transpose in a "cut-and-paste" manner and that in a "copy-and-paste" manner (*Helitron* and *Polinton/Mavericks*). The classification is based on Wicker T. et al., 2007[26]. **(b)** Viruses are classified into six groups, according to the constitution of their genome; 1) positive-sense single-stranded RNA viruses; 2) double-stranded RNA viruses; 3) negative-sense single-stranded RNA viruses; 4) Retroviruses; 5) single-stranded DNA viruses; 6) double-stranded DNA viruses. The classification is based on (eds) King A. M. Q. et al., 2011[27]. Abbreviations; ATPase, DNA packaging ATPase; CP, capsid protein; EN, endonuclease; ENV, envelope protein; INT, integrase; PR, protease; pPolB, protein-primed family B DNA polymerase; RC-Rep, rolling-circle replication initiation endonuclease; RdRP, RNA-dependent RNA polymerase; RH, RNA helicase; RT, reverse-transcriptase; S1H, superfamily 1 helicase; S3H, superfamily 3 helicase; TPase, DDE transposase.

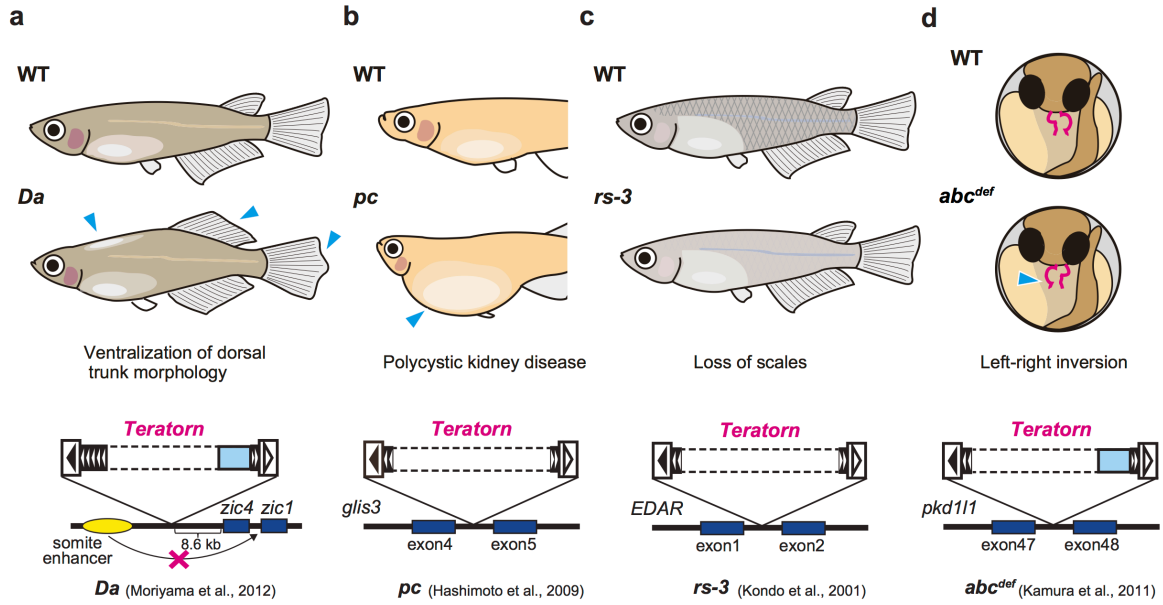


Figure 2 | Medaka spontaneous mutants caused by *Teratorn* insertion. (a) *Da* mutant (ventralization of dorsal trunk morphology, Moriyama et al., 2012[41]).**(b),** *pc* mutant (polycystic kidney disease, Hashimoto et al., 2009[43]). **(c)** *rs-3* mutant (loss of scales, Kondo et al., 2001[42]). **(d)** *abc*^{def} mutant (defect in left-right axis formation, Kamura et al., 2011[44]). Solid and dotted boxes inside *Teratorn* indicate the sequence-determined and undetermined region, respectively.

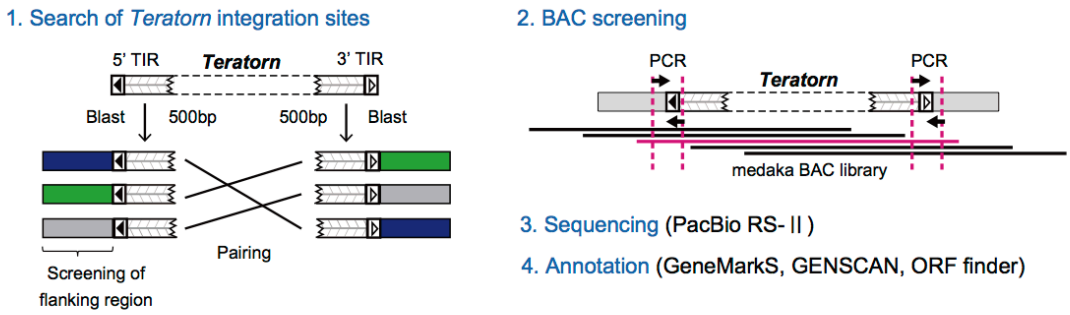


Figure 3 | Screening and sequencing of *Teratorn* copies. The procedure of screening and sequencing of the full-length *Teratorn* copies: 1. *Teratorn* insertion site determination by screening of 5' and 3' flanking regions from the public medaka genome database, followed by pairing of the 5' and 3' flanking regions, which are derived from the same loci; 2. Screening of BAC clones that include the whole sequence of *Teratorn*; 3. Sequencing of BAC clones; 4. gene annotation.

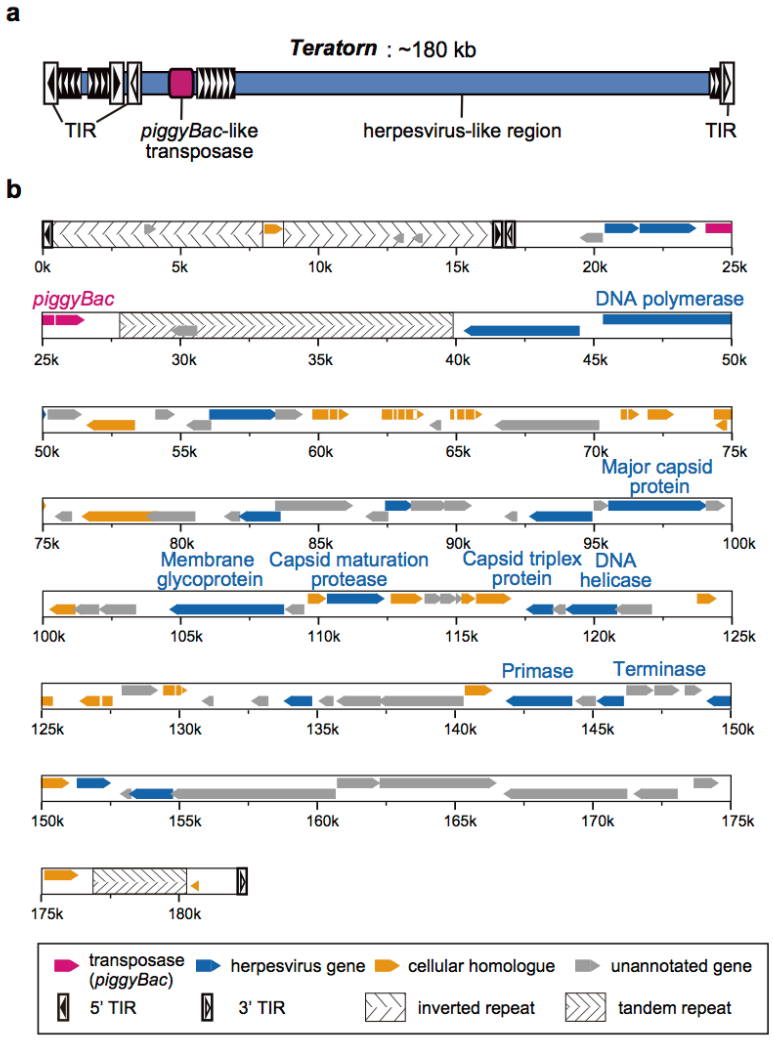


Figure 4 | Sequence characteristics of *Teratorn*. (a) The overall structure of *Teratorn*. **(b)** Gene map of the subtype 1 *Teratorn* copy (73I9; named from the BAC clone ID). Predicted genes are classified into four categories depicted by colored arrowheads; magenta, *PiggyBac*-like transposase gene; blue, herpesvirus-like genes; yellow, cellular homologues; gray, unannotated genes. Terminal inverted repeats (TIRs) of *PiggyBac*-like transposon are depicted by boxed triangles.

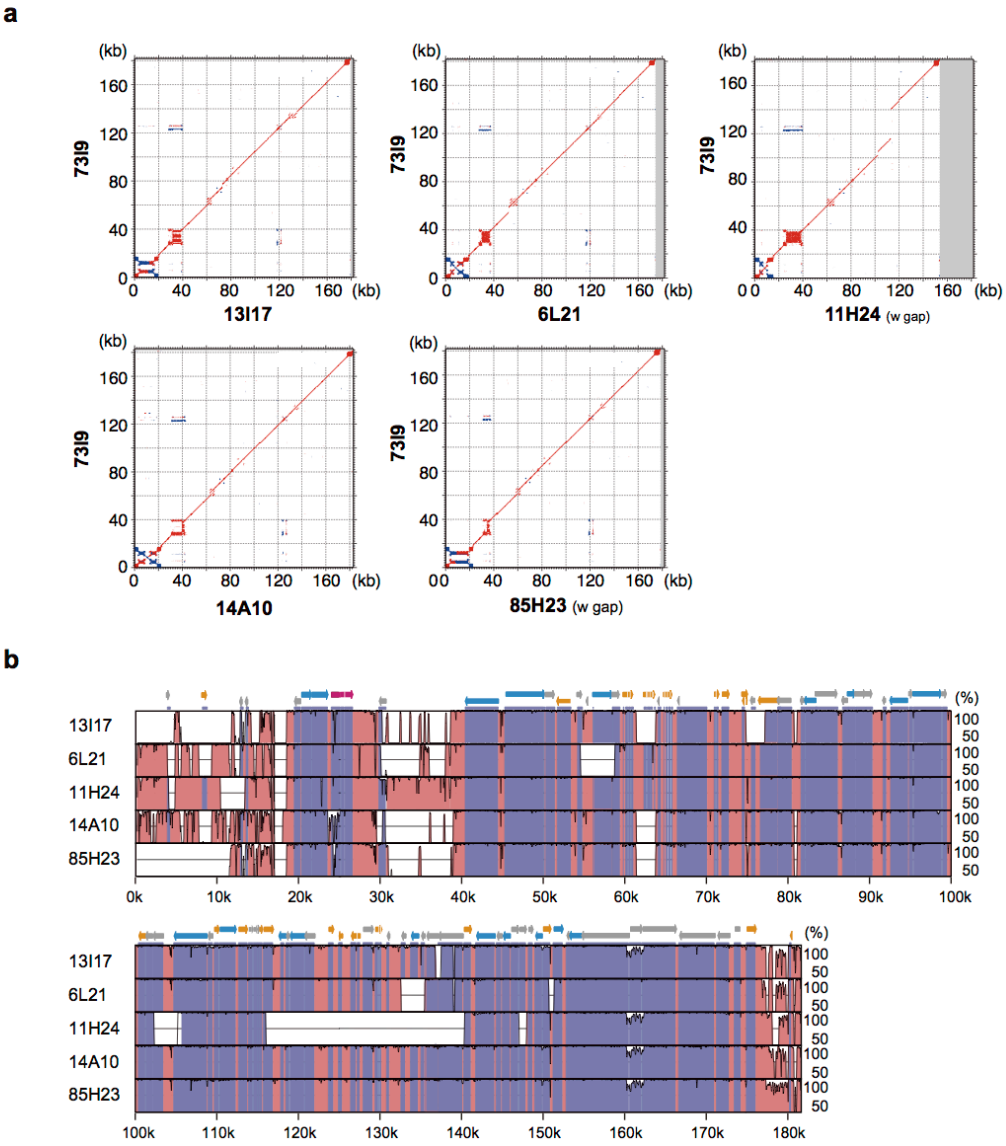


Figure 5 | High sequence similarity between *Teratorn* copies. (a) Dot plots showing the alignment of one *Teratorn* copy (73I9; named from the ID of BAC clone) with the other five copies (13I17, 6L21, 11H24, 14A10, 85H23). The red and blue dots indicate that the corresponding residues of the two sequences match in the forward and reverse direction, respectively. Note that the synteny, including the position of repetitive region, is conserved among all copies. **(b)** Sequence comparison between one copy and the other five copies, visualized by VISTA. Purple and red indicate putative coding and non-coding regions, respectively. Arrows above the histogram indicate the position of genes in 73I9 copy.

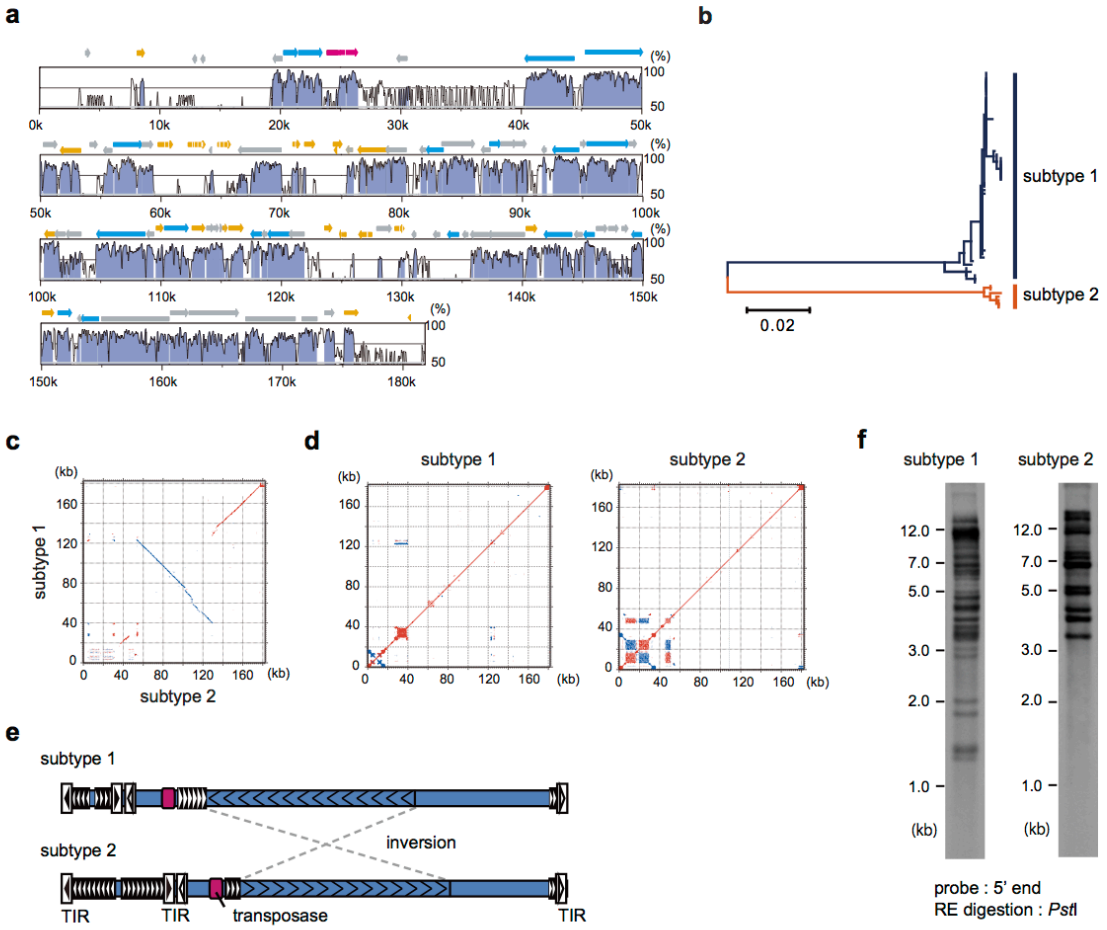


Figure 6 | There are two subtypes of *Teratorn*. **(a)** Sequence comparison between *Teratorn* subtype 1 and subtype 2, visualized by VISTA. Blue and white regions indicate coding and non-coding regions, respectively. Arrows above the histogram indicate the position of genes inside *Teratorn* subtype 1. **(b)** Neighbor-joining tree of all *Teratorn* transposase copies in the genome of medaka Hd-rR inbred strain. Note that they are separated into two clusters. **(c)** A dot plot matrix showing the alignment of *Teratorn* subtype 1 and subtype 2. Note that synteny is almost conserved except for a ~80-kb long inversion in the middle. **(d)** Dot plots showing the alignment of each *Teratorn* subtype with itself. Note that the position of inverted repeat and tandem repeats are common between the two subtypes. **(e)** Comparison of the whole structure of subtype 1 and subtype 2 *Teratorn*. **(f)** Southern blot displaying individual *Teratorn* insertions in the genome of medaka Hd-rR inbred strain. The ~300-bp 5' terminal region of *Teratorn* of each subtype was used as hybridization probes.

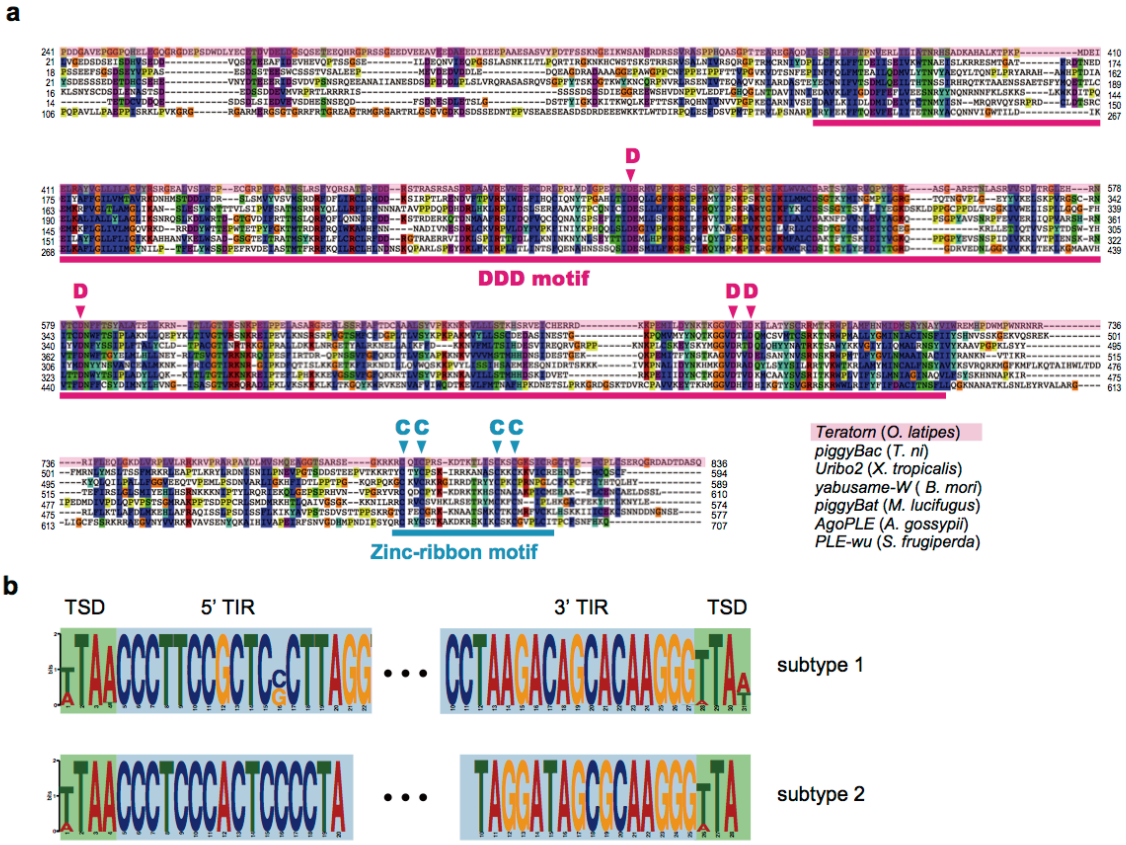


Figure 7 | *Teratorn* belongs to *piggyBac* superfamily. (a) Alignment of partial amino acid sequences of several active *piggyBac* superfamily transposase genes. Note the conservation of the four aspartic acid residues, which form the core of the 'DDD motif' and are required for transposition reaction of *piggyBac* (magenta arrowheads). In addition, four cysteine residues at the C termini, which form the core of Zinc-ribbon motif with unknown function, are also conserved (blue arrowheads). **(b)** Consensus sequences of TIR and TSD of all *Teratorn* copies in the genome of Hd-rR inbred strain, displayed by MEME. Note that sequence composition of TIR and TSD follow the rule of *piggyBac* family; TIR length ranges from 12 to 19bp beginning with the "CCYT" motif, and targets TTAA.

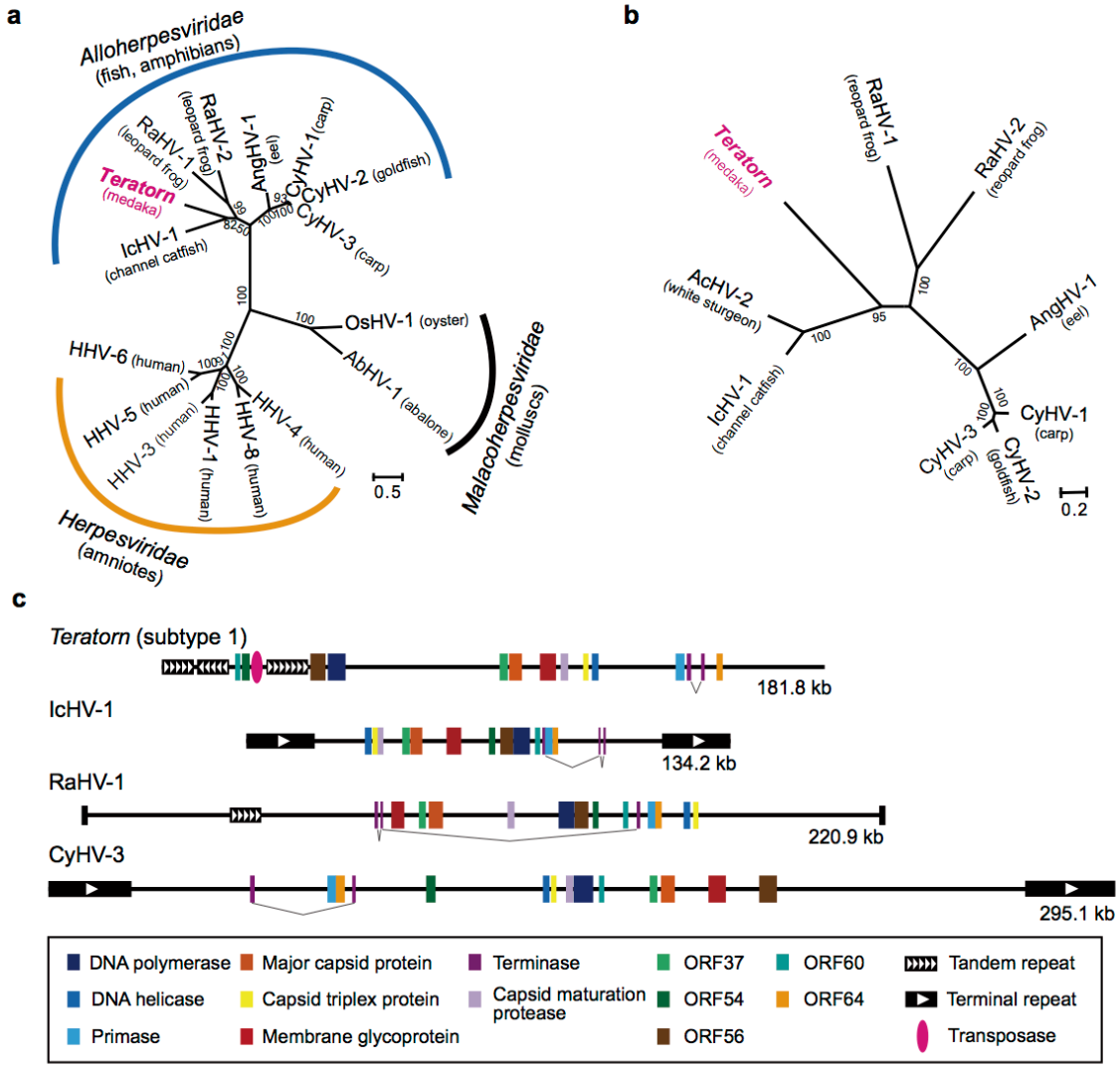


Figure 8 | *Teratorn* contains a full herpesvirus genome. (a) Maximum-likelihood tree based on the amino acid sequence of the DNA packaging terminase gene, the only gene confidently conserved among all herpesvirus species. Bootstrap values of branching are indicated at the nodes. Note that, among the three families (*Herpesviridae*, *Alloherpesviridae* and *Malacoherpesviridae*) in the *Herpesvirales* order, *Teratorn* belongs to the family *Alloherpesviridae* (infecting fish and amphibians). **(b)** Maximum-likelihood tree based on the concatenated amino acid sequences of major capsid protein, DNA helicase, DNA polymerase and DNA packaging terminase from all alloherpesvirus species with sequenced genome. Within *Alloherpesviridae*, *Teratorn* is only distantly related to any other species. **(c)** Comparison of genomic structure of subtype 1 *Teratorn* and several alloherpesvirus species, as representatives of genera in the family *Alloherpesviridae*. The colored squares indicate each herpesvirus core gene, and blacked boxes depict repeats. Note that *Teratorn* contains all 13 core genes conserved among all alloherpesvirus species, as well as long repeats.

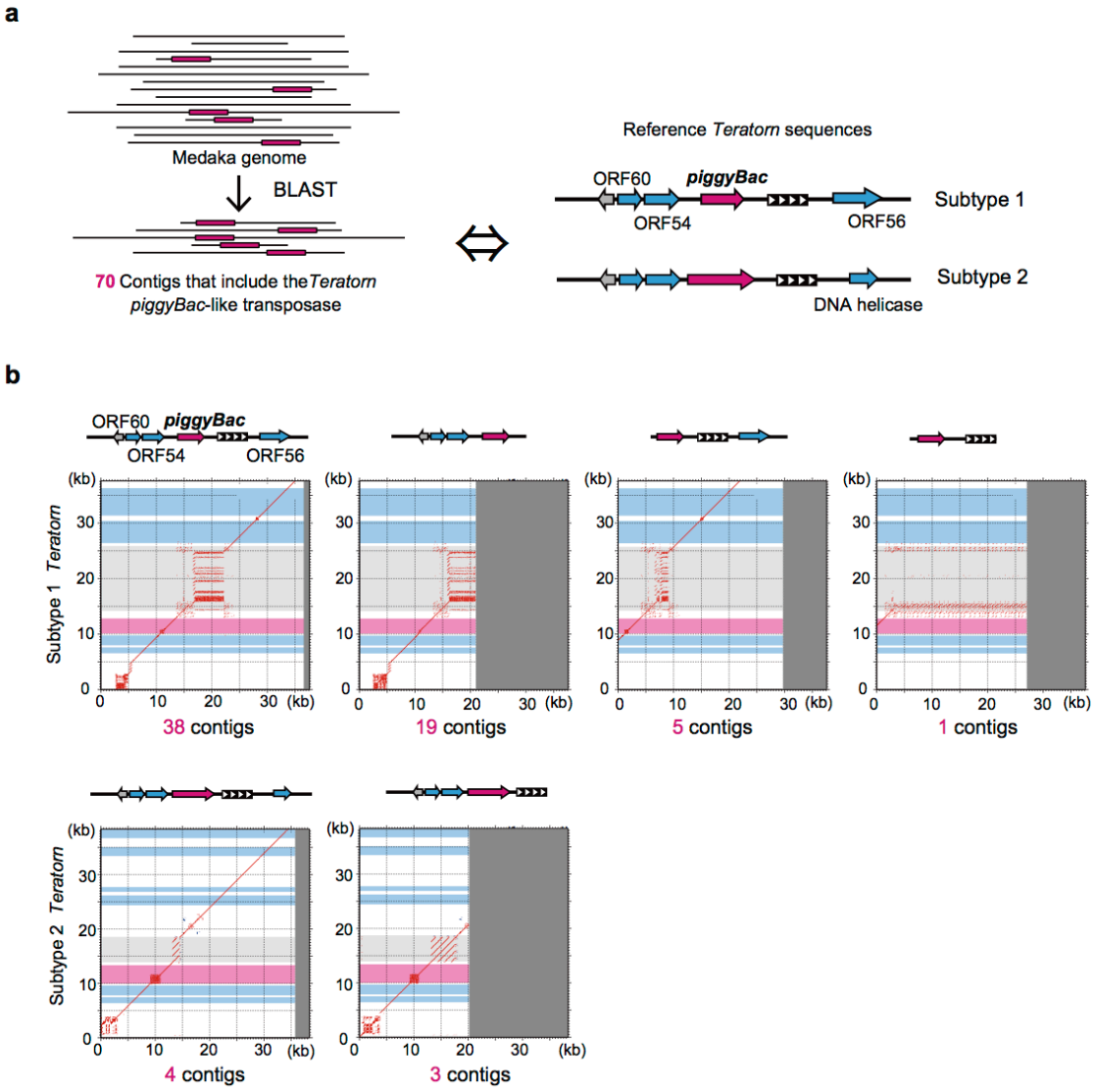


Figure 9 | All of the *piggyBac*-like transposase copies are adjacent to the herpesvirus genes in the medaka genome. (a) The procedure of screening all contigs that include the *piggyBac*-like transposase gene of *Teratorn*. First, all contigs that contain the transposase gene were screened from the medaka draft genome by blastn. For all contigs obtained, genomic neighborhoods around the transposase genes were tested by displaying alignment with the reference *Teratorn* sequence. **(b)** Dot plots showing the alignment of the reference sequence of *Teratorn* with the contigs screened from the medaka draft genome. Magenta, cyan and light gray indicate the coding region of the *piggyBac*-like transposase gene, herpesvirus genes and tandem repeats, respectively.

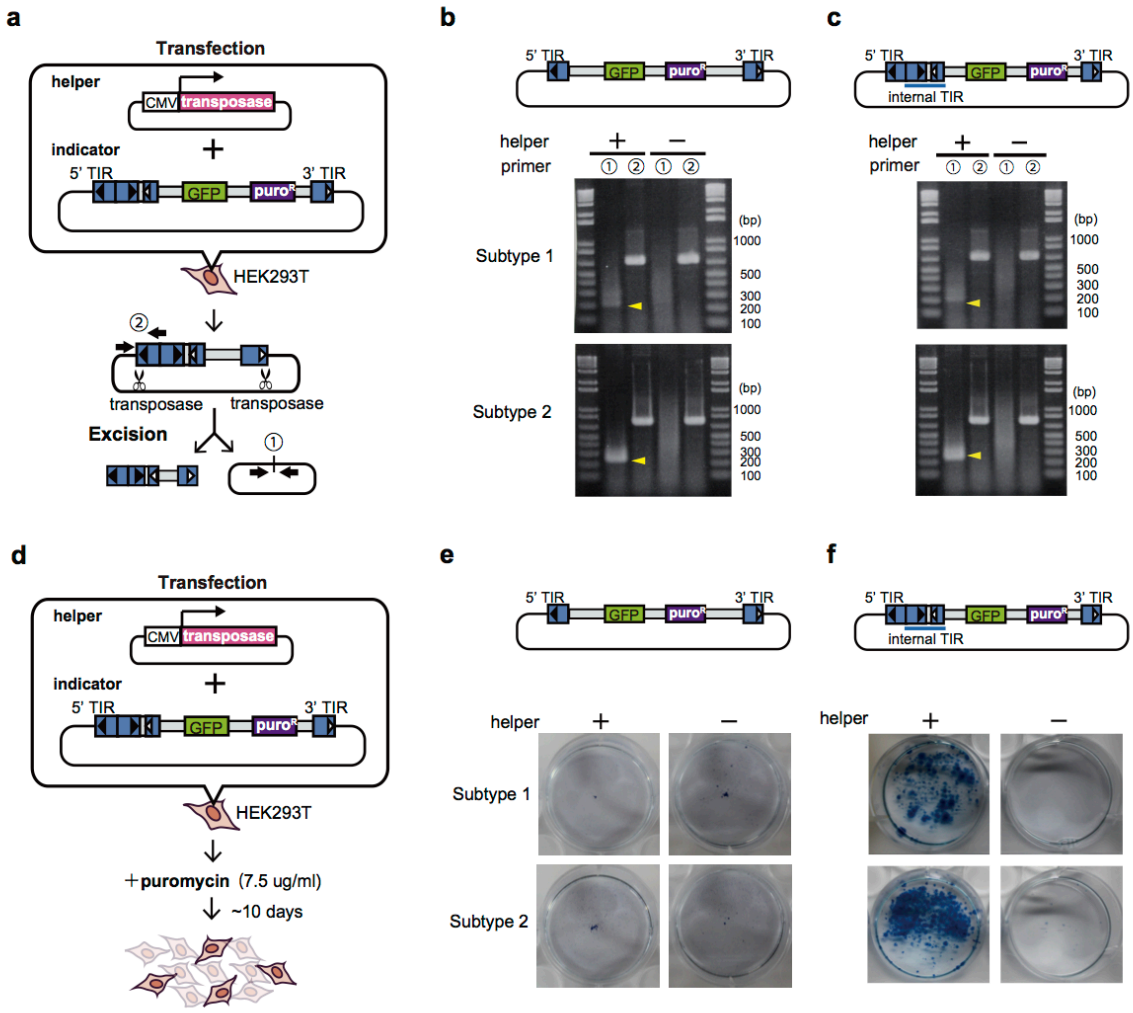


Figure 10 | *Teratorn* retains transposition activity. **(a)** A schematic of the excision assay. In the helper plasmid, *Teratorn* transposase gene was expressed under the CMV promoter. In the indicator plasmid, a GFP reporter and a puromycin-resistant gene were flanked by the 5' and 3' TIR. For a subset of samples, additional TIR was inserted at the boundary of 5' TIR and GFP cassette so as to mimic the endogenous *Teratorn* structure (internal TIR). Transposition activity was examined by co-transfection of those two plasmids into HEK293T cells, followed by PCR-based detection of transposon cassette excision from the indicator plasmid. Thick arrows indicate primer pairs used for the excision assay. **(b)** Results of the excision assay in the absence of internal TIR. The number indicates the target region of PCR depicted in a (1, flanking the transposon cassette, amplifying only when excision reaction takes place; 2, targeting a terminus of transposon cassette, positive control). Note that PCR product flanking the transposon cassette was detected only when the helper plasmid was co-transfected (subtype 1, 202 bp; subtype 2, 250 bp; arrowheads), suggesting that 5' and 3' TIRs are sufficient for excision reaction. **(c)** Excision assay in the presence of internal TIRs. **(d)** Schematic of the integration assay. In this assay, long-term chemical selection

(~10 days) was carried out following plasmid transfection to screen transgenic cell lines. **(e)** Results of the integration assay in the absence of internal TIR. 13 days after 7.5 μg/ml of puromycin selection, colonies were fixed and stained with methylene blue. Note that no colony was formed even in the presence of transposase. **(f)** Integration assay in the presence of internal TIR. Note that multiple colonies were observed when the helper plasmid was co-transfected, indicating that internal TIR is required for chromosomal integration.

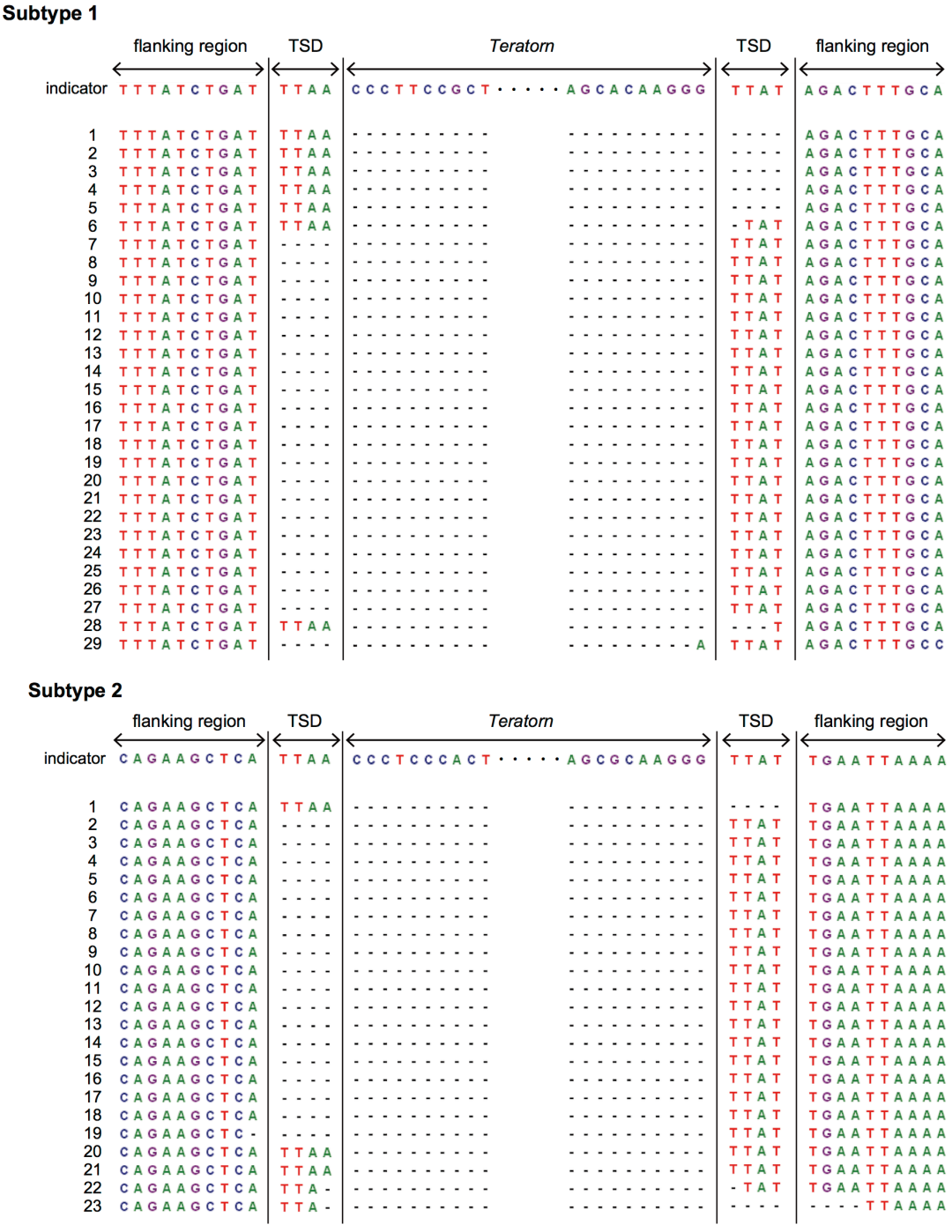


Figure 11 | Precise excision of *Teratorn* transposon cassette from the indicator plasmid. Nucleotide sequences of PCR products obtained from the excision assay, as described in Fig.10c. The top line indicates the sequence of the indicator plasmid, and the following 29 (subtype 1) and 23 (subtype 2) lines indicate the sequences of individual subcloned PCR products. Note that the excision reaction occurred precisely, at high frequency (subtype 1, 26/29 clones; subtype 2, 18/23 clones), similar to other *piggyBac* superfamily DNA transposons[57]. TSD; target site duplication.

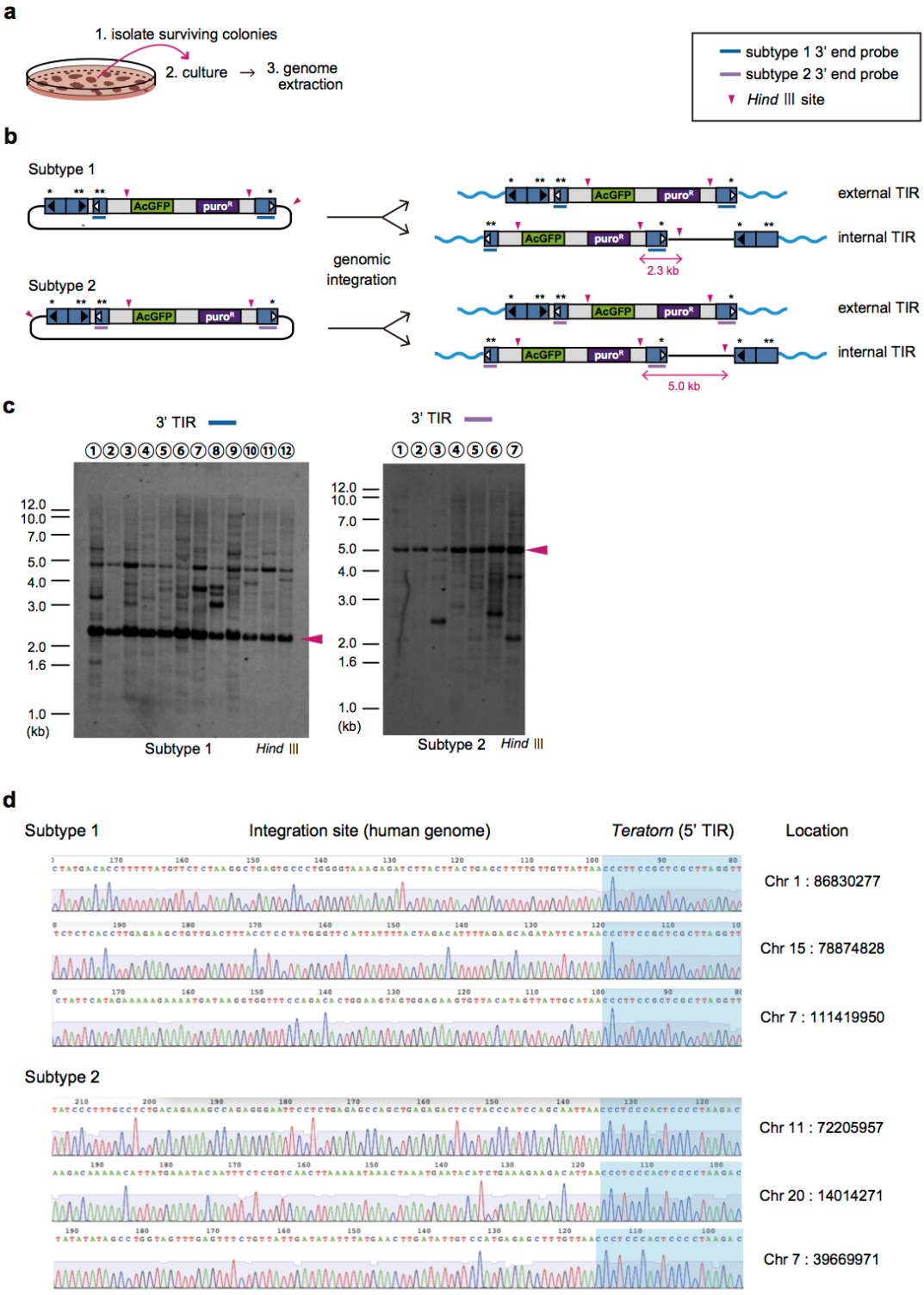

a
1. isolate surviving colonies
2. culture → 3. genome extraction
subtype 1 3' end probe
subtype 2 3' end probe
Hind III site
b
Subtype 1
Subtype 2
AcGFP
puroR
genomic integration
external TIR
internal TIR
2.3 kb
5.0 kb
c
3' TIR
1 2 3 4 5 6 7 8 9 10 11 12
1 2 3 4 5 6 7
12.0
10.0
7.0
5.0
4.0
3.0
2.0
1.6
1.0
(kb)
Subtype 1 Hind III
Subtype 2 Hind III
d
Subtype 1
Integration site (human genome)
Teratorn (5' TIR)
Location
Chr 1 : 86830277
Chr 15 : 78874828
Chr 7 : 111419950
Subtype 2
Chr 11 : 72205957
Chr 20 : 14014271
Chr 7 : 39669971

Figure 12 | Confirmation of in *vitro* integration of *Teratorn* transposon cassette into HEK293T cells. (a) Isolation method of genomic DNA from cell colonies that survived puromycin selection. **(b)** Possible way of chromosomal integration of the indicator plasmids. Integration can be occurred via 1) external TIRs (upper) and/or internal TIRs (lower). Blue and purple bars indicate the position of the target sequences for southern hybridization (blue, subtype 1; purple, subtype 2). Magenta arrowheads indicate *Hind*III sites. **(c)** Southern blotting of genomic DNA of surviving colonies of HEK293T cells, using 3' terminal sequences as hybridization probes. Numbers above the lanes indicate the genomic DNA of individual colonies. Although the band pattern was different among individual clones, there is a common band (arrowheads) which size corresponds to the DNA fragment formed by *Hind*III digestion (depicted as double-headed arrows in b), suggesting that transposition via internal TIRs also took place (as in b). **(d)** Examples of insertion sites of *Teratorn* transposon cassette in the genome of surviving HEK293T cell colonies. The blue region indicates the terminal sequence of *Teratorn*.

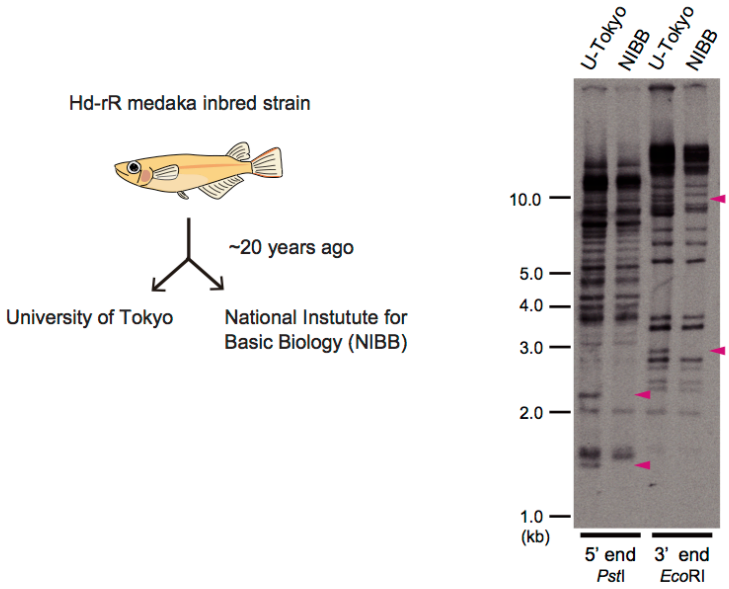
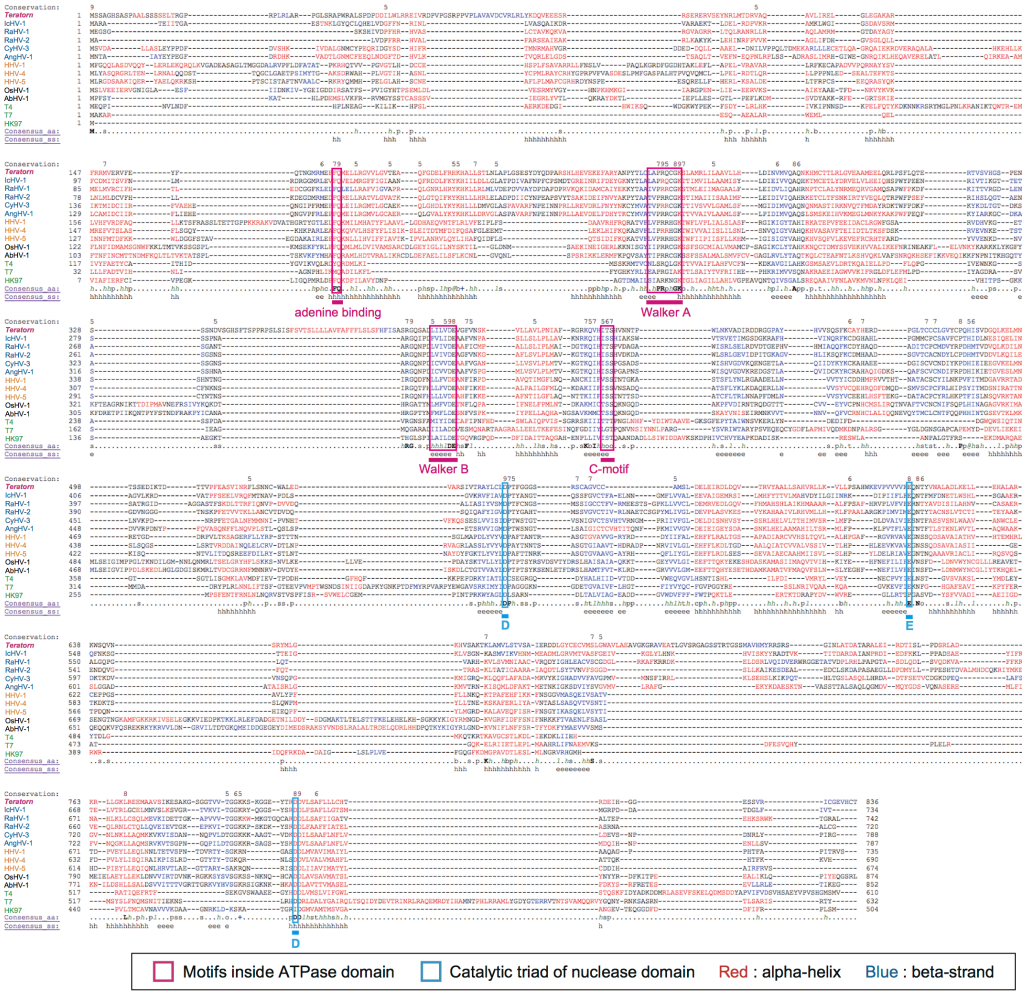


Figure 13 | Implication of endogenous *Teratorn* transposition in vivo. Southern blotting detecting individual *Teratorn* insertions in the Hd-rR individuals kept in the University of Tokyo and at the National Institute for Basic Biology, using sequences of *Teratorn* 5' and 3' ends as hybridization probes. Note that band patterns are different between the two individuals (arrowheads).

a


adenine binding
Walker A
Walker B
C-motif
D
E
Motifs inside ATPase domain
Catalytic triad of nuclease domain
Red : alpha-helix
Blue : beta-strand

b

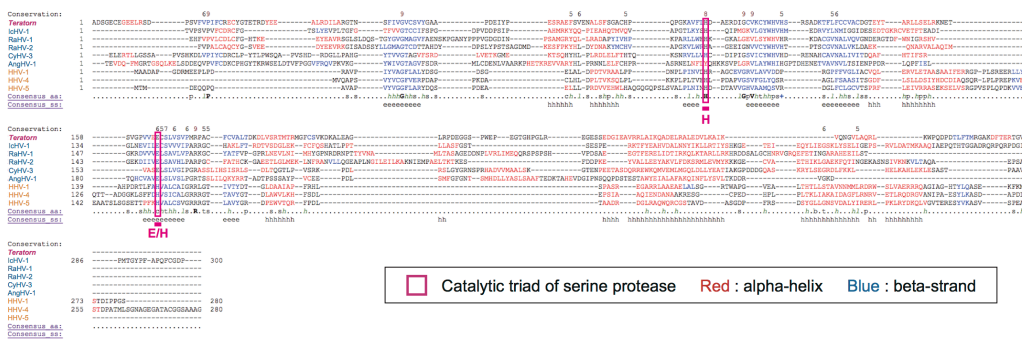
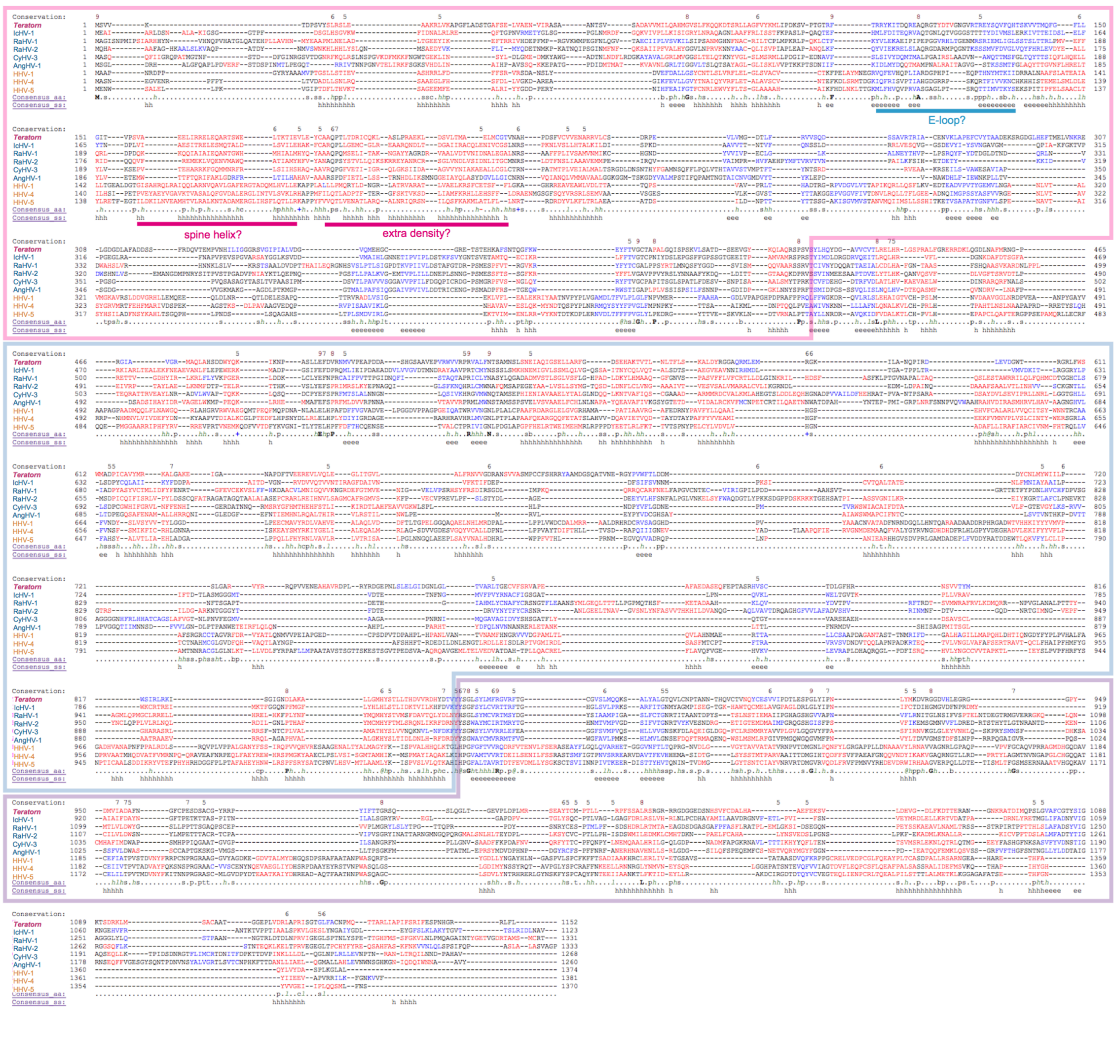

H
S
E/H
D
Catalytic triad of serine protease
Red : alpha-helix
Blue : beta-strand

Figure 14 | Catalytic residues of DNA packaging terminase and capsid maturation protease are conserved in *Teratorn*. (a) Multiple alignment of the full-length amino acid sequences of DNA packaging terminase gene, constructed by PROMALS3D (Pei J. et al., 2008[107]) (*Teratorn*, magenta; *Alloherpesviridae*, blue ; *Herpesviridae*, yellow; *Malacoherpesviridae*, black; bacteriophages, green). Catalytic center motifs are depicted by magenta (ATPase domain) and cyan (nuclease domain), respectively. Note the sequence conservation at catalytic centers in *Teratorn*. **(b)** Multiple alignment of partial sequences of capsid maturation protease gene (*Teratorn*, magenta; *Alloherpesviridae*, blue; *Herpesviridae*, yellow). Note the sequence conservation of catalytic triads (His-Ser-His/Glu).

a


E-loop?
spine helix?
extra density?

b

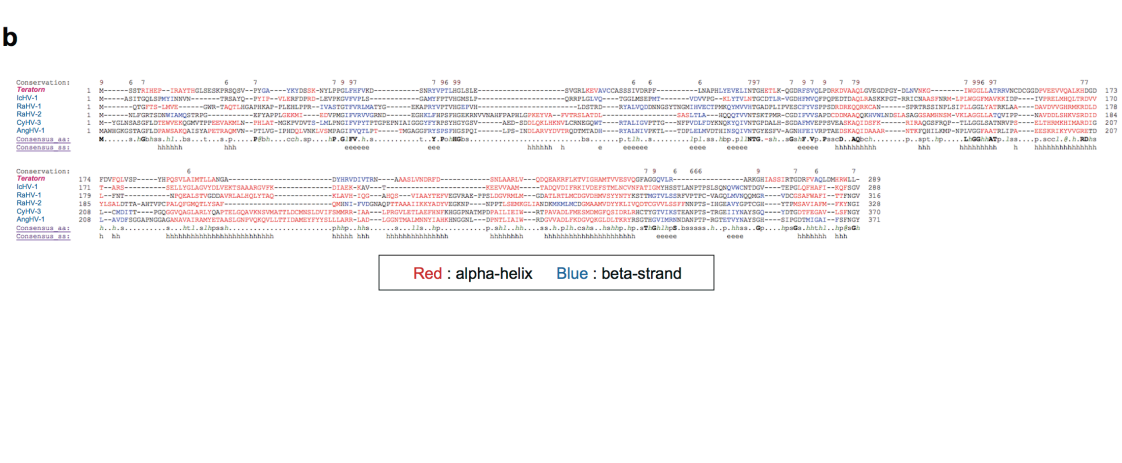

Red : alpha-helix
Blue : beta-strand

Figure 15 | Similar secondary structure pattern of capsid proteins between herpesviruses and *Teratorn*. (a) Multiple alignment of the full-length amino acid sequences of major capsid protein, constructed by PROMALS3D (Pei J. et al., 2008[107]) (*Teratorn* ,magenta; *Alloherpesviridae*, blue; *Herpesviridae*, yellow). Predicted secondary structures are depicted as red (**α**-helix) and blue (ß-strand). Herpesvirus major capsid proteins are known to be subdivided into three domains; Floor domain, which faces at the lumen of the capsid, Upper domain, which faces at the outer surface of the capsid, and Middle domain, which locates in the middle. Although sequence similarity was low, similar pattern of secondary structures was observed. **(b)** Multiple alignment of the full-length sequence of subunit 2 capsid triplex protein of *Teratorn* (magenta) and alloherpesvirus species (blue). Note the sequence similarity among them.

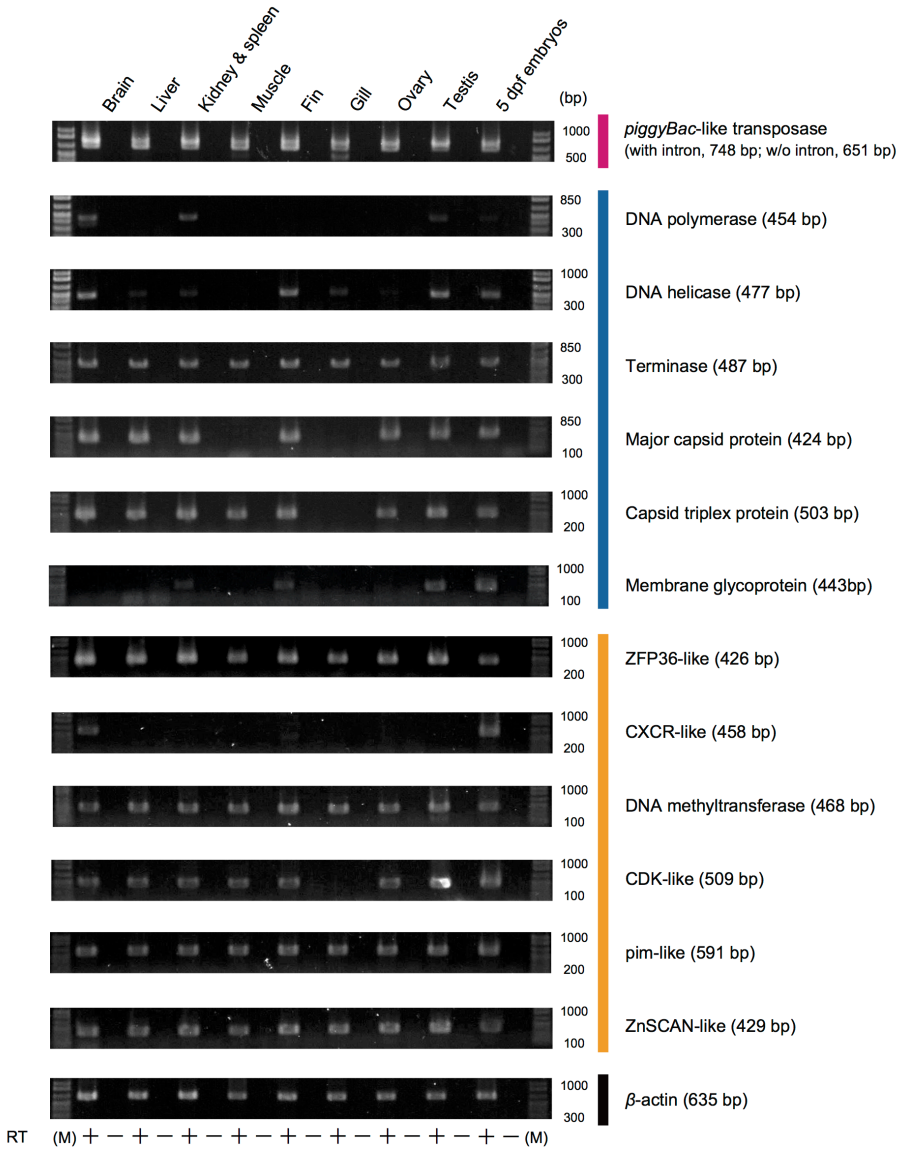


Figure 16 | Expression of some *Teratorn* genes in several medaka tissues. RT-PCR of *Teratorn* genes in several adult medaka tissue and st. 35 medaka embryos. "+" and "–" indicate that reverse-transcription reaction was carried out or not, respectively. PCR was performed for 40 cycles. The colors indicate the categories of genes shown as in Fig. 4b (magenta, *piggyBac* transposase; blue, herpesvirus genes with known funtion; yellow, cellular homologues that seem to be involve in evation from host immunity (ZFP36-like, CXCR-like and DNA methyltransferase-like) and cell proliferation (CDK-like, pim-like and ZnSCAN-like).

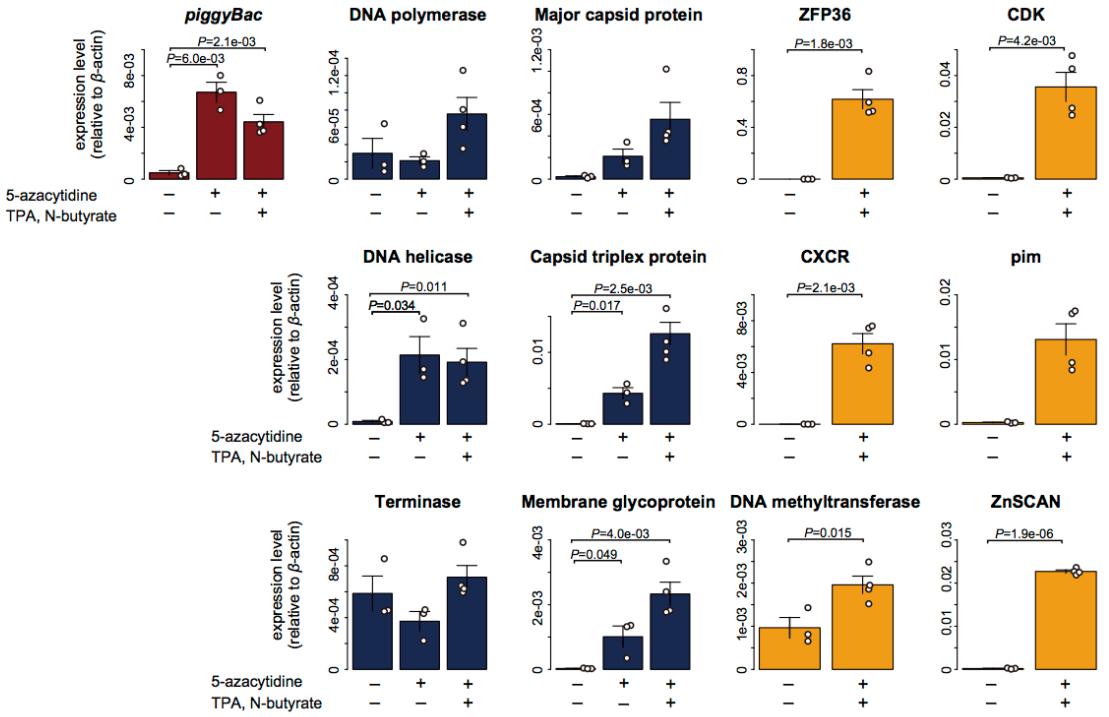


Figure 17 | Moderate upregulation of *Teratorn* genes by 5-azacytidine administration. qPCR analysis of *Teratorn* genes in medaka fibroblast cells administered with or without 2 μM of 5-azacytidine, 3 mM of N-butyrate and 500 ng/ml of 12-*O*-Tetradecanoylphorbol 13-acetate (TPA). "+" and "–" indicate that each chemical was administrated or not. The value indicates the ratio of molar concentration relative to *β*-actin. Note that expression levels of most genes were moderately increased by chemical administration, although the expression level was still low. Statistical significance was tested by one-sided Welch Two Sample t-test. Each data point indicates the raw value of each experiment, and bars represent the mean ± SEM of replicates. Number of biological replicates are as follows; n = 3 for no chemical treatment, n = 3 for 5-azacytidine treatment, n = 4 for 5-azacytidine, TPA and N-butyrate treatment.

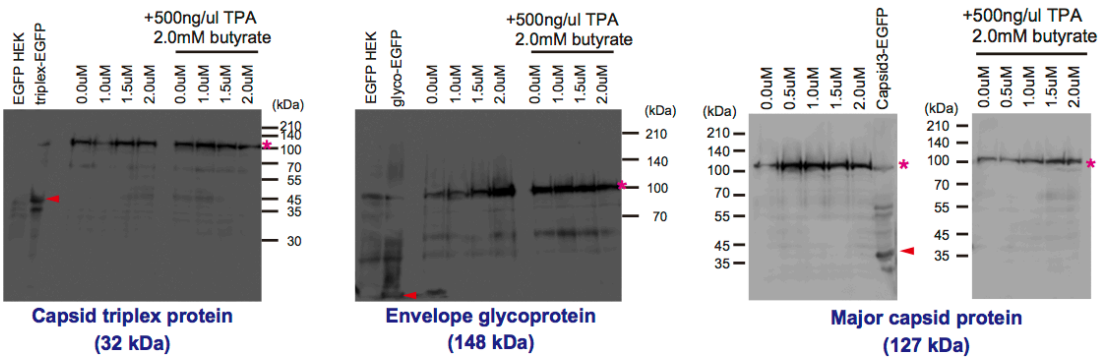


Figure 18 | No virus structural protein production after administration of 5-azacytidine, TPA and N-butyrate in medaka fibroblasts. Western blot of the lysate of medaka fibroblasts administered with 5-azacytidine, TPA and N-butyrate for the presence of three herpesvirus structural proteins (capsid triplex protein, envelope glycoprotein and major capsid protein) of *Teratorn*. Arrowheads indicate the positive control; lysate of cells transfected with the plasmids that express the antigen protein fragment fused with GFP. Note that no clear signal was observed at the corresponding molecular weight for each gene. Asterisks indicate a non-specific signal.

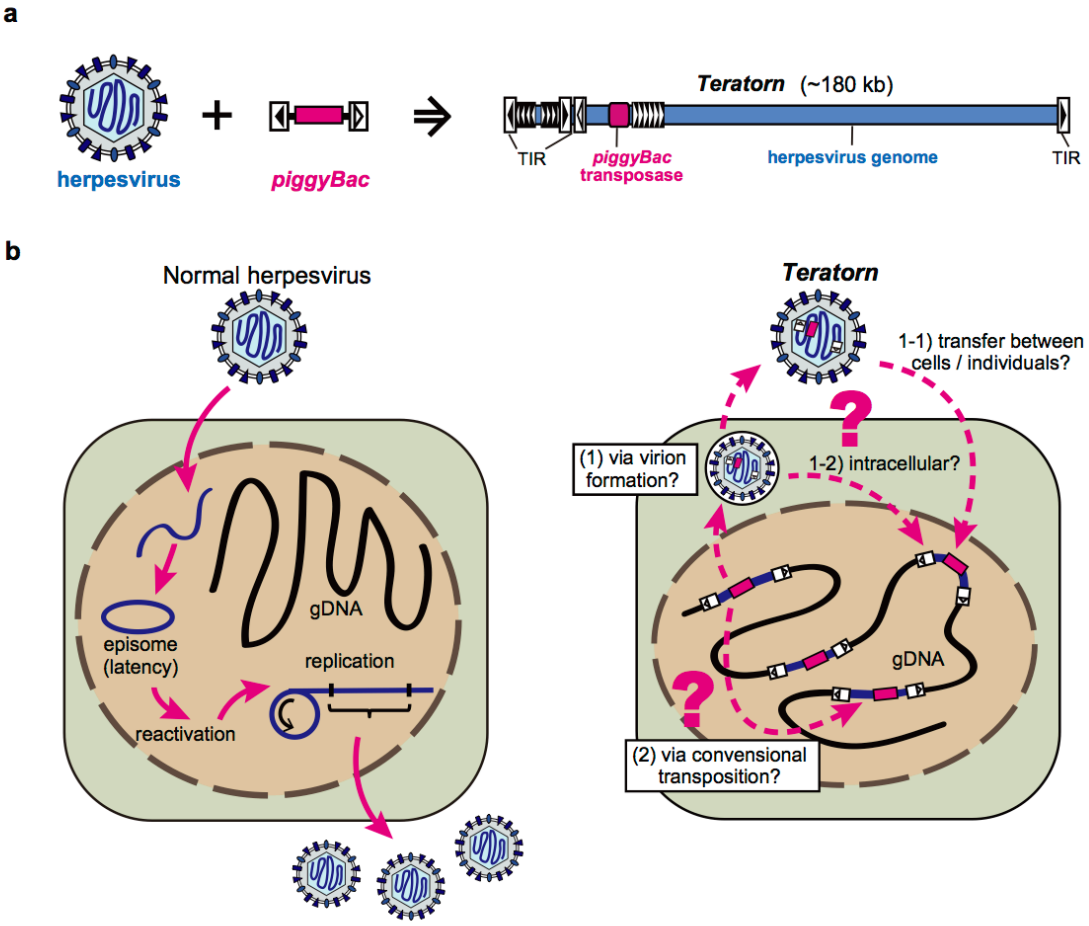


Figure 19 | Model of *Teratorn* derivation from a herpesvirus that shifted to an intragenomic life cycle by gaining the *piggyBac* transposon system. (a) The entire structure of *Teratorn*. *Teratorn* is a fusion of a *piggyBac*-like transposon and a herpesvirus. **(b)** Comparison of life cycles of normal herpesviruses and *Teratorn*. Normal herpesviruses don't integrate their genome into chromosomes of host cells during their life cycles; instead, they form episomal DNA molecules and persist inside cells, accompanied by recursive reactivation (left). *Teratorn* might be a novel herpesvirus that had shifted its life cycle to an intragenomic transposon-like parasite, by gaining the *piggyBac* transposon system. Transposition mechanism might be either via (1) DNA replication and virus particle formation or (2) conventional cut-and-paste transposition. Virus particles might be either (1-1) infectious and transmissible to other cells / individuals or (1-2) non-infectious and remain inside cells (right).

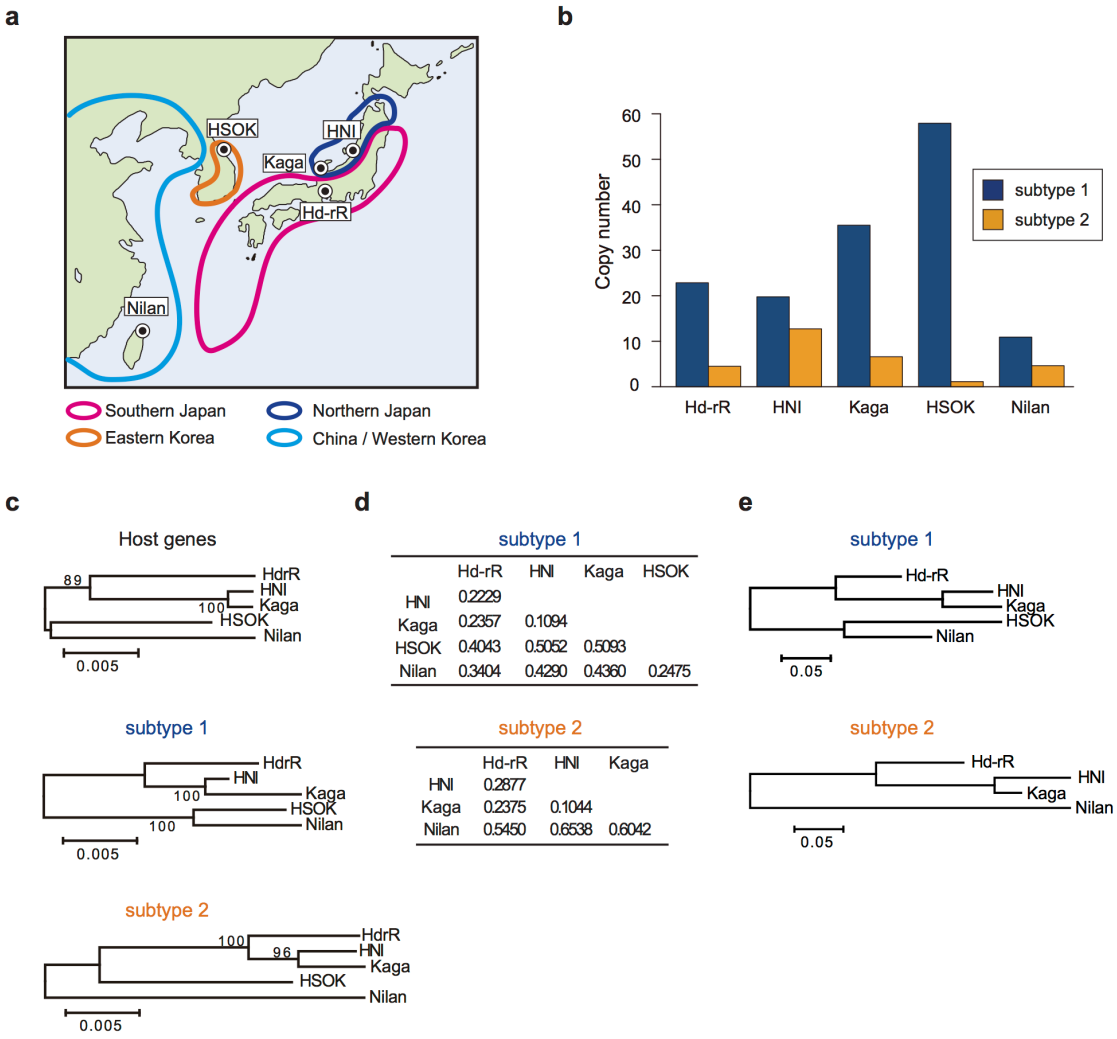


Figure 20 | Distibution of *Teratorn* in the Japanese medaka populations. (a) Geographic distribution of wild medaka populations. The geographic origins of each medaka inbred strain used in this analysis are depicted by circles. **(b)** Copy number of *Teratorn* in each medaka inbred strain (blue, subtype 1; orange, subtype 2). Copy number was estimated by mapping of Illumina whole-genome shotgun read data against the reference genome, followed by division of the average coverage value of all *Teratorn* genes by the average coverage of the rest of all nuclear genes. **(c)** Maximum-likelihood trees based on the third codons of the sequences of 18 host genes (Betancur-R R. et al., 2013[111]) and the consensus sequence of 16 *Teratorn* genes of each medaka inbred strain. Genome sequence of each strain was inferred from mapping of Illumina short read data against the Hd-rR reference genome, followed by consensus sequence calling. Bootstrap values of branching are indicated at the nodes. **(d)** Pairwise F_{ST} values among *Teratorn* copies in each medaka inbred strain. F_{ST} values were inferred from whole-genome shotgun sequencing data by popoolation2 (Kofler R. et al., 2011[88]), using a sliding window and step size of 400, followed by a calculation of average values throughout the *Teratorn* sequences. **(e)** Neighbor-joining trees based on the pairwise F_{ST} between each medaka inbred strain.

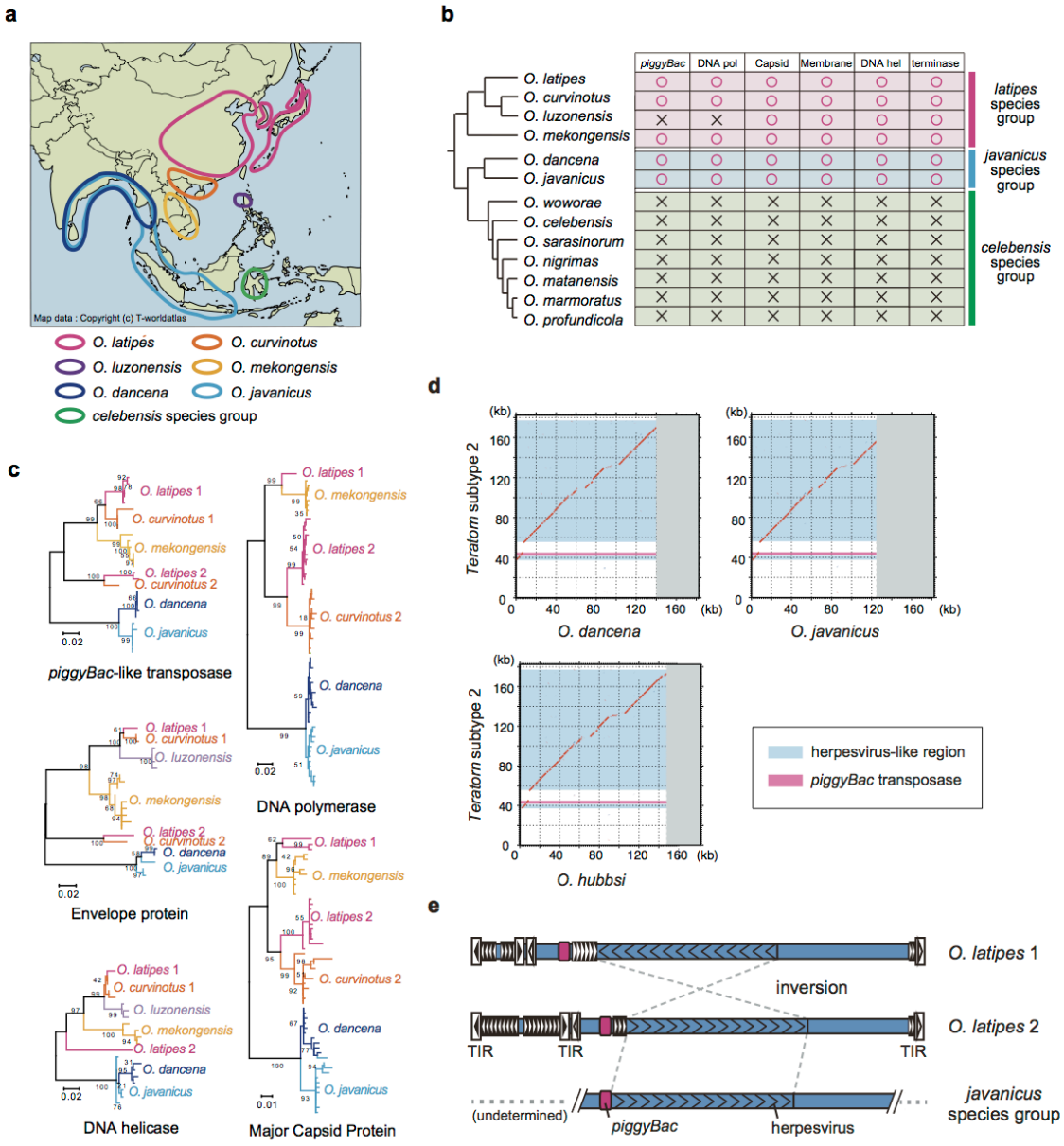


Figure 21 | Distribution and phylogeny of *Teratorn* in *Oryzias* genus. (a) Geographic distribution of medaka related species. **(b)** Results of PCR screening of six *Teratorn* genes in 13 medaka related species. Phylogenetic relationships of medaka related species used in this analysis are depicted on the left (Takehana et al., 2005[89]). Note that *Teratorn* genes are detected in *latipes* and *javanicus* species group but not in *celebensis* species group. **(c)** Maximum-likelihood trees based on the sequences of subcloned PCR products of *Teratorn* genes from each medaka related species. Note that the topology of the phylogenetic trees of each gene are almost the same as that of host species, except for the existence of two subtypes of *Teratorn* in *latipes* species group. The scale bar

represents the number of substitutions per site. **(d)** Dot plots showing the alignment of *Teratorn* in medaka species of *javanicus* species group with subtype 2 *Teratorn* in *O. latipes*. *piggyBac*-like transposase gene and herpesvirus-like region are shown in magenta and cyan, respectively. **(e)** Comparison of the structure of *Teratorn* in *Oryzias* genus. Note that gene synteny, including the position of the *piggyBac* transposase gene, is conserved in the *Oryzias* genus.

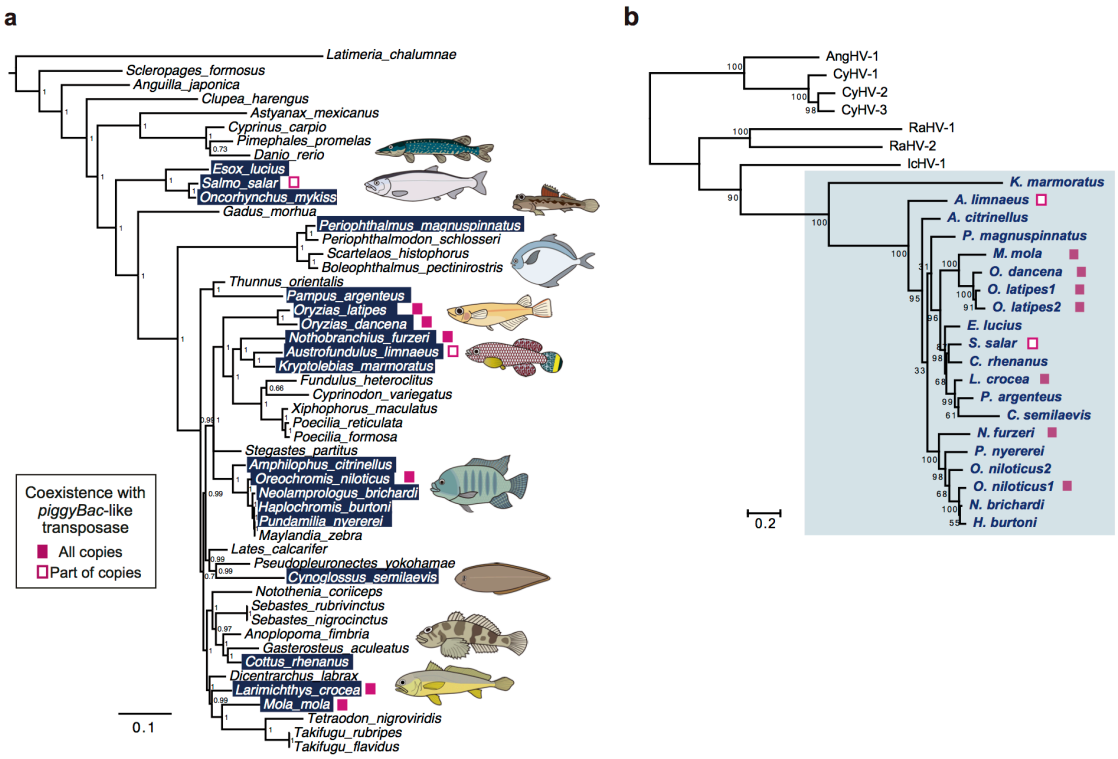
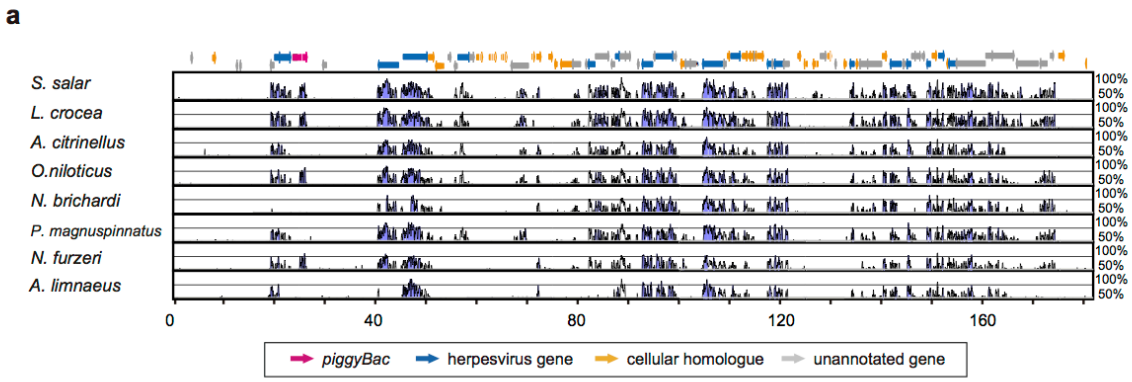
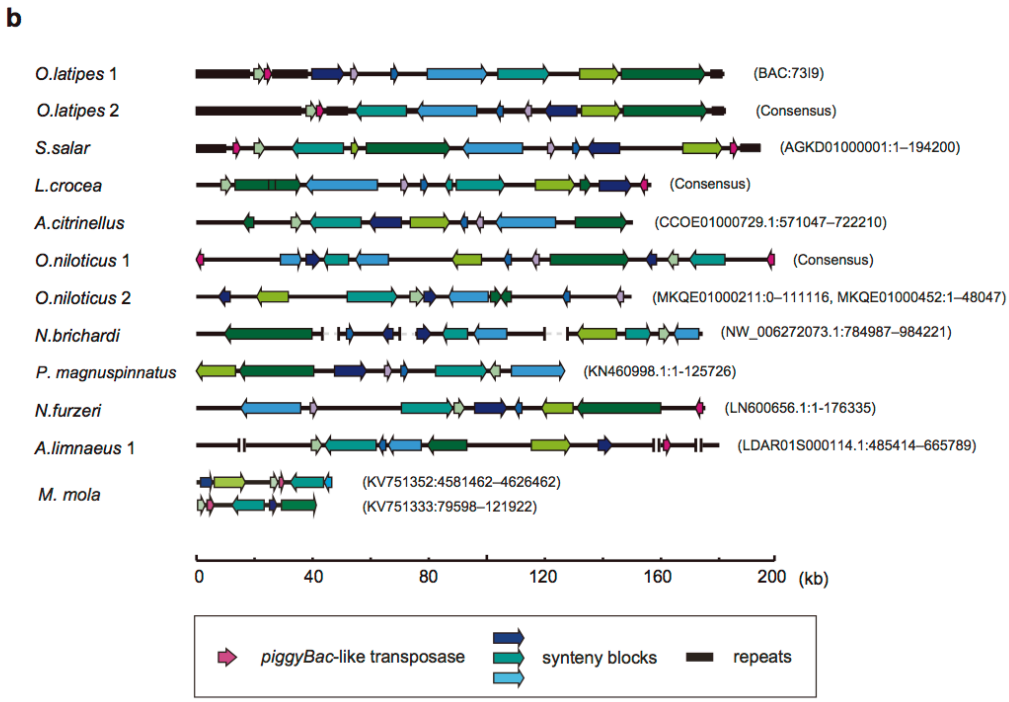


Figure 22 | *Teratorn*-like elements are widely distributed in teleost fish genomes. (a) Result of the tblastn search of 13 herpesvirus core genes in *Teratorn* against publically available genome data of teleost fish species. Species that seem to contain *Teratorn*-like element (more than 9 of the 13 herpesvirus core genes with their E-value < 10^{-3}) are highlighted in blue. Phylogenetic tree was constructed by bayesian inference, based on the concatenated nucleotide sequence of nine nuclear genes (Near T. J. et al., 2012[114]). Species in which *Teratorn*-like elements are adjacent to *piggyBac*-like transposase gene are marked by magenta squares. **(b)** Maximum-likelihood tree based on the concatenated amino acid sequences of major capsid protein, DNA helicase, DNA polymerase and DNA packaging terminase from *Teratorn*-like elements in teleosts and exogenous alloherpesvirus species. *Teratorn*-like elements are highlighted in blue. The scale bars represent the number of substitutions per site.


a
S. salar
L. crocea
A. citrinellus
O.niloticus
N. brichardi
P. magnuspinnatus
N. furzeri
A. limnaeus
100%
50%
0
40
80
120
160
piggyBac
herpesvirus gene
cellular homologue
unannotated gene


b
O.latipes 1
O.latipes 2
S.salar
L.crocea
A.citrinellus
O.niloticus 1
O.niloticus 2
N.brichardi
P. magnuspinnatus
N.furzeri
A.limnaeus 1
M. mola
(BAC:73I9)
(Consensus)
(AGKD01000001:1–194200)
(Consensus)
(CCOE01000729.1:571047–722210)
(Consensus)
(MKQE01000211:0–111116, MKQE01000452:1–48047)
(NW_006272073.1:784987–984221)
(KN460998.1:1-125726)
(LN600656.1:1-176335)
(LDAR01S000114.1:485414–665789)
(KV751352:4581462–4626462)
(KV751333:79598–121922)
0
40
80
120
160
200
(kb)
piggyBac-like transposase
synteny blocks
repeats

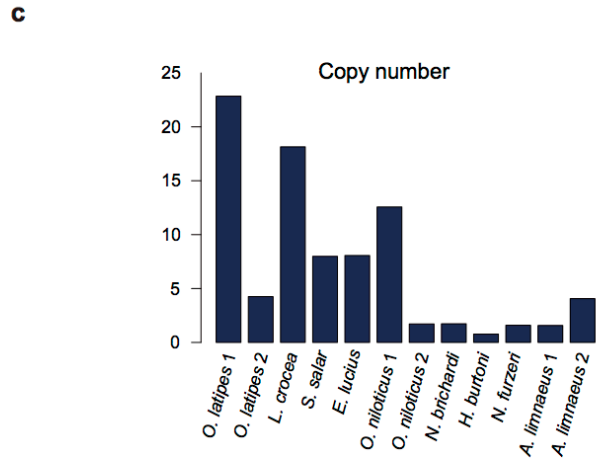

c
Copy number
25
20
15
10
5
0
O. latipes 1
O. latipes 2
L. crocea
S. salar
E. lucius
O. niloticus 1
O. niloticus 2
N. brichardi
H. burtoni
N. furzeri
A. limnaeus 1
A. limnaeus 2

Figure 23 | *Teratorn*-like elements encode a series of herpesvirus-like genes and contain multiple copies. (a) Homology plot of *Teratorn*-like element against subtype 1 *Teratorn* in medaka visualized by VISTA. The blue and white regions indicate coding and non-coding regions, respectively. Arrows above the histograms indicate the position of genes of medaka *Teratorn* (subtype 1). **(b)** Structures of *Teratorn*-like elements in several teleost fish species. Those sequences were (1) extracted from contigs or scaffolds (*S. salar*, *A. citrinellus*, *N. brichardi*, *N. furzeri, A. limnaeus* and *M. mola*) or (2) reconstructed by conjugating several contigs (*L. crocea* and *O. niloticus*). Conserved synteny blocks are depicted by the same colors. Magenta arrows indicate *piggyBac*-like transposase. Sources of each sequence are described at the right. **(c)** Estimated copy number of *Teratorn*-like elements in the genomes of some teleost fishes. Copy number was estimated by mapping of Illumina short read data against the reference genome data, caluculation of coverage at each nucleotide position, then dividing the average coverage value in herpesvirus core genes by the average coverage of coding region of nuclear genes.

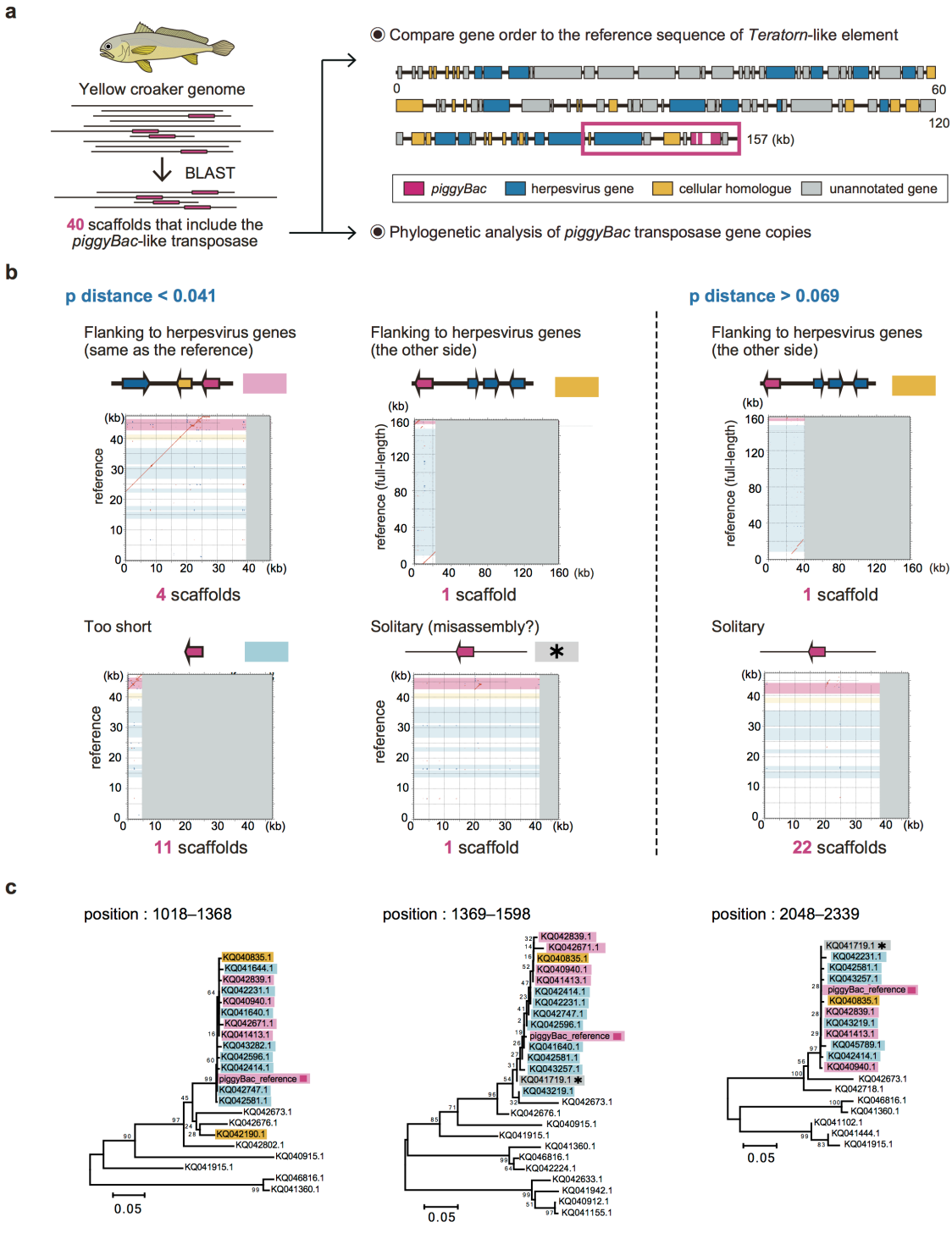

a
Yellow croaker genome
BLAST
40 scaffolds that include the piggyBac-like transposase
Compare gene order to the reference sequence of Teratorn-like element
0
60
120
157 (kb)
piggyBac
herpesvirus gene
cellular homologue
unannotated gene
Phylogenetic analysis of piggyBac transposase gene copies
b
p distance < 0.041
p distance > 0.069
Flanking to herpesvirus genes (same as the reference)
4 scaffolds
Flanking to herpesvirus genes (the other side)
1 scaffold
Flanking to herpesvirus genes (the other side)
1 scaffold
Too short
11 scaffolds
Solitary (misassembly?)
1 scaffold
Solitary
22 scaffolds
reference
reference (full-length)
(kb)
c
position : 1018–1368
position : 1369–1598
position : 2048–2339
piggyBac_reference
0.05

Figure 24 *piggyBac*-herpesvirus fusion in yellow croaker. **(a)** The procedure of screening all scaffolds that include the *piggyBac* transposase gene inside *Teratorn*-like element in yellow croaker (*Larimichthys crocea*). First, scaffolds that contain the transposase gene were screened from the yellow croaker genome by blastn. For all scaffolds obtained, genomic neighborhoods around the transposase genes was tested by displaying alignment with the reference sequence of *Teratorn*-like element. In parallel, phylogenetic relationship of transposon copies was analyzed. Predicted ORFs (exons) are depicted by colored squares according to the categories; magenta, *piggyBac* transposase; blue, herpesvirus genes; yellow, cellular homologues; gray, unannotated genes. **(b)** Dot plots showing the alignment of the partial region of the reference sequence of *Teratorn*-like element with scaffolds obtained by BLAST. Magenta, cyan and yellow indicate coding regions of the transposase, herpesvirus genes and OTU-like cysteine protease gene, respectively. **(c)** Neighbor-joining trees based on the sequences of *piggyBac* transposase copies obtained by blast search are displayed. Three regions (1018–1368, 1369–1598, 2048–2339) of the reference transposase sequence were utilized for the phylogenetic tree construction, since some copies contain only a partial sequence of transposase. Each sequence is named by the scaffold name (e.g. KQ042839.1). *piggyBac* copies are categorized into five groups described in b, according to the linkage to the herpesvirus genes. Note that *piggyBac* copies adjacent to the herpesvirus-like sequence (magenta) are clustered together, suggesting that a particluar type of *piggyBac* element is fused with *Teratorn*-like element. The scale bars represent the number of substitutions per site.

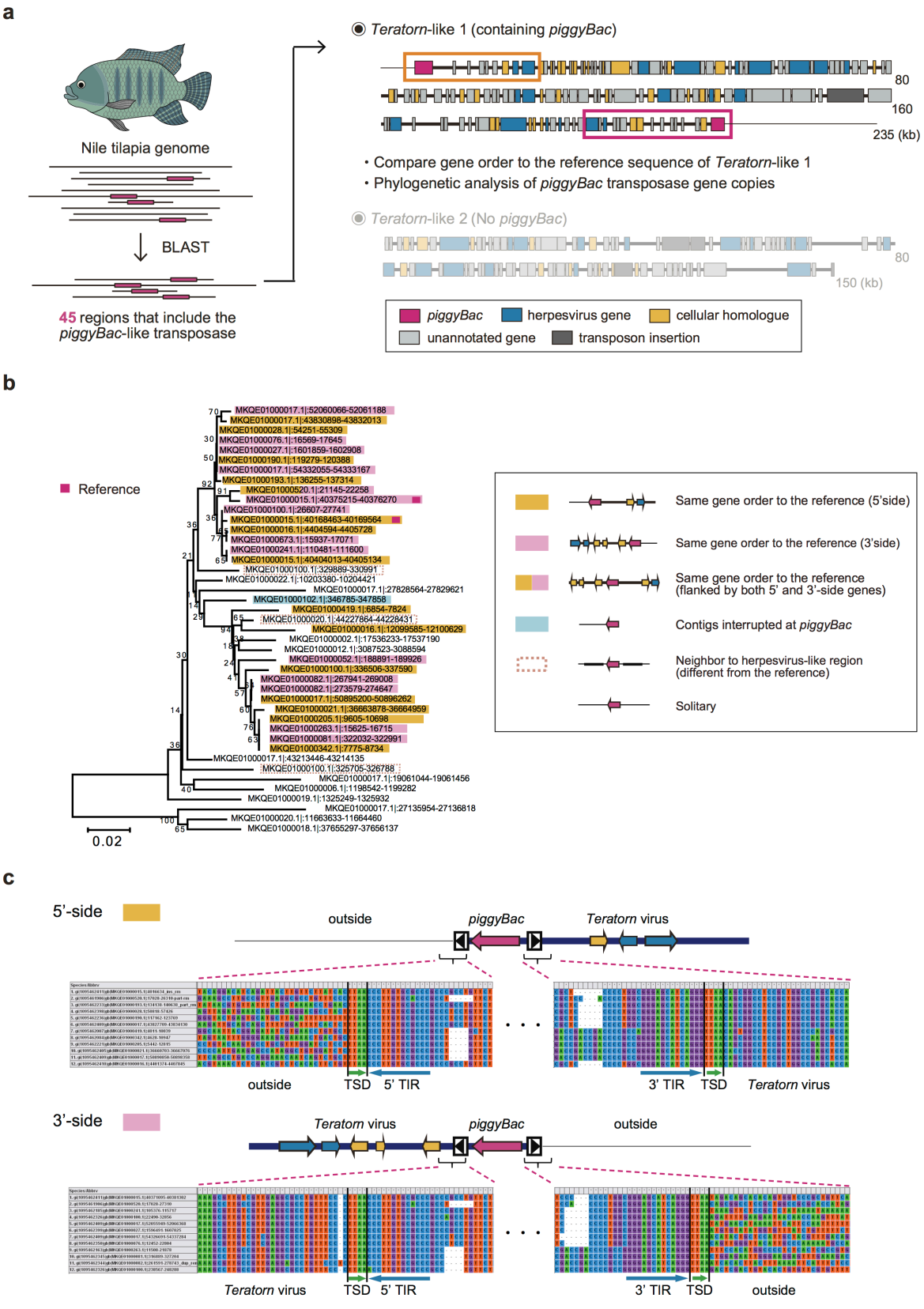

a
Nile tilapia genome
BLAST
45 regions that include the piggyBac-like transposase
Teratorn-like 1 (containing piggyBac)
80
160
235 (kb)
• Compare gene order to the reference sequence of Teratorn-like 1
• Phylogenetic analysis of piggyBac transposase gene copies
Teratorn-like 2 (No piggyBac)
80
150 (kb)
piggyBac
herpesvirus gene
cellular homologue
unannotated gene
transposon insertion
b
Reference
MKQE01000017.1|:52060066-52061188
MKQE01000017.1|:43830898-43832013
MKQE01000028.1|:54251-55309
MKQE01000076.1|:16569-17645
MKQE01000027.1|:1601859-1602908
MKQE01000190.1|:119279-120388
MKQE01000017.1|:54332055-54333167
MKQE01000193.1|:136255-137314
MKQE01000520.1|:21145-22258
MKQE01000015.1|:40375215-40376270
MKQE01000100.1|:26607-27741
MKQE01000015.1|:40168463-40169564
MKQE01000016.1|:4404594-4405728
MKQE01000673.1|:15937-17071
MKQE01000241.1|:110481-111600
MKQE01000015.1|:40404013-40405134
MKQE01000100.1|:329889-330991
MKQE01000022.1|:10203380-10204421
MKQE01000017.1|:27828564-27829621
MKQE01000102.1|:346785-347858
MKQE01000419.1|:6854-7824
MKQE01000020.1|:44227864-44228431
MKQE01000016.1|:12099585-12100629
MKQE01000002.1|:17536233-17537190
MKQE01000012.1|:3087523-3088594
MKQE01000052.1|:188891-189926
MKQE01000100.1|:336506-337590
MKQE01000082.1|:267941-269008
MKQE01000082.1|:273579-274647
MKQE01000017.1|:50895200-50896262
MKQE01000021.1|:36663878-36664959
MKQE01000205.1|:9605-10698
MKQE01000263.1|:15625-16715
MKQE01000081.1|:322032-322991
MKQE01000342.1|:7775-8734
MKQE01000017.1|:43213446-43214135
MKQE01000100.1|:325705-326788
MKQE01000017.1|:19061044-19061456
MKQE01000006.1|:1198542-1199282
MKQE01000019.1|:1325249-1325932
MKQE01000017.1|:27135954-27136818
MKQE01000020.1|:11663633-11664460
MKQE01000018.1|:37655297-37656137
0.02
Same gene order to the reference (5'side)
Same gene order to the reference (3'side)
Same gene order to the reference (flanked by both 5' and 3'-side genes)
Contigs interrupted at piggyBac
Neighbor to herpesvirus-like region (different from the reference)
Solitary
c
5'-side
outside
piggyBac
Teratorn virus
outside
TSD
5' TIR
3' TIR
TSD
Teratorn virus
3'-side
Teratorn virus
piggyBac
outside
Teratorn virus
TSD
5' TIR
3' TIR
TSD
outside

Figure 25 | *piggyBac*-herpesvirus fusion in nile tilapia. (a) The procedure of screening all genomic regions that include the *piggyBac* transposase gene inside *Teratorn*-like 1 in nile tilapia. First, genomic regions of the transposase copies were identified by blastn. For all regions obtained, genomic neighborhoods to the transposase were tested by displaying alignment with the reference sequence of *Teratorn*-like 1. In parallel, phylogenetic relationship of the transposase copies was analyzed. Predicted ORFs (exons) are depicted by colored boxes according to the categories; magenta, *piggyBac* transposase; blue, herpesvirus genes; yellow, cellular homologues; light gray, unannotated genes; dark gray, transposon insertion. **(b)** Neighbor-joining trees based on the *piggyBac* transposase copies obtained by blast search are displayed. *piggyBac* copies are categorized into six groups as described in the inset, according to the linkage to the herpesvirus-like sequence. Each sequence is named by the genomic location of each transposon copy (e.g. MKQE01000017.1|:52060066-52061188). Note that there are multiple *piggyBac* copies linked to the herpesvirus-like sequence in the same configuration as the reference sequence (either 5' and 3' end of *Teratorn*-like 1, yellow and magenta), although phylogenetically polyphyletic. The scale bar represents the number of substitutions per site. **(c)** Multiple alignments of *piggyBac* transposon copies adjacent to *Teratorn*-like element 1 (5' side, 12 loci; 3' side, 12 loci). Terminal region of *piggyBac* transposon was displayed. Note that sequence homology was seen from the 5' TIR (terminal inverted repeat) of the 5'–side *piggyBac* to the 3' TIR of the 3'–side *piggyBac*, while no homology was found outside, suggesting the transposition of *Teratorn*-like element via *piggyBac* of both ends. TSD : target site duplication.

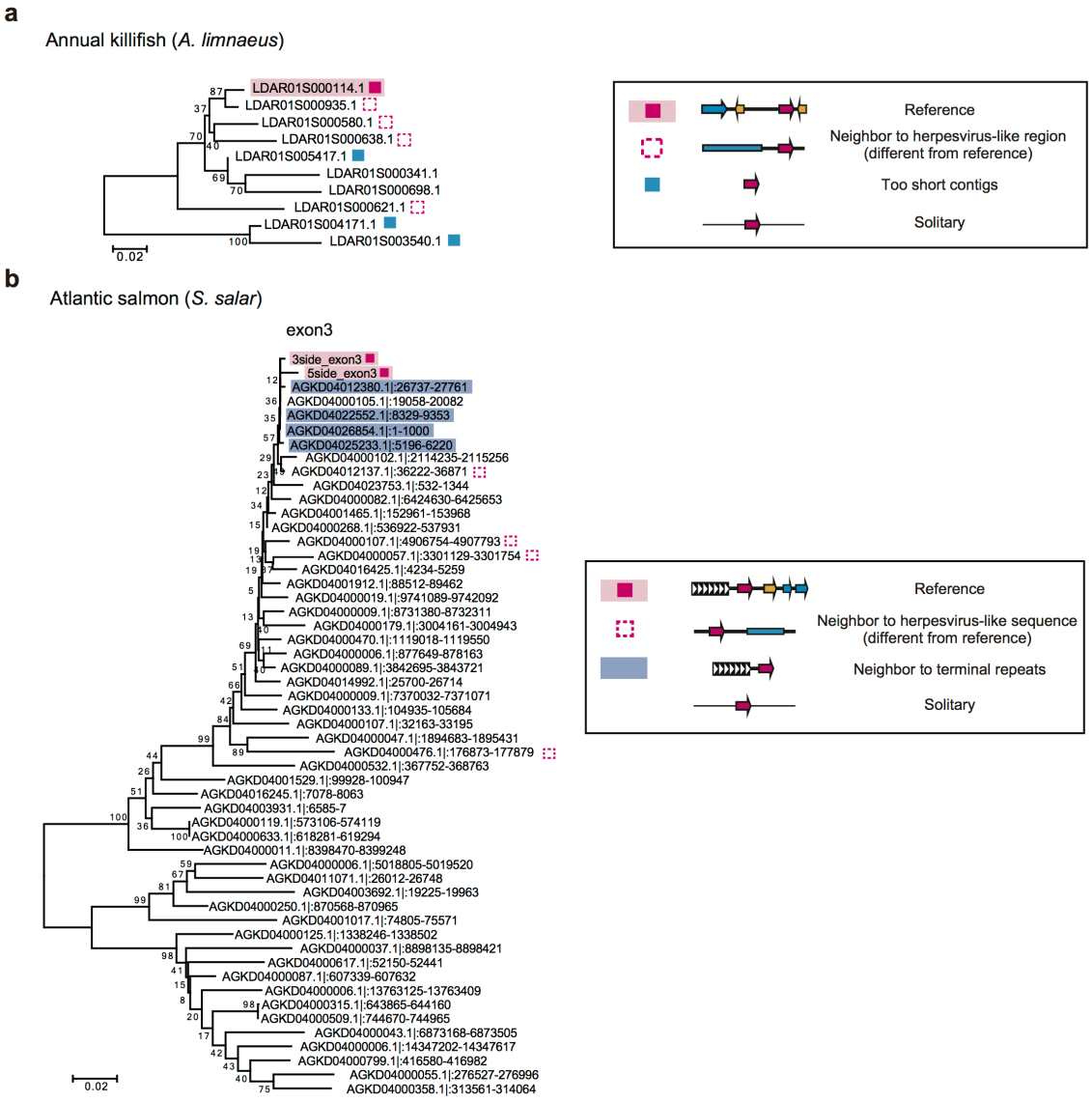


Figure 26 | Validation of *piggyBac*-herpesvirus fusion in annual killifish and Atlantic salmon. Neighbor-joining trees based on the *piggyBac* transposase copies of **(a)** annual killifish (*A. limnaeus*) and **(b)** Atlantic salmon *(S. salar*) are displayed. *piggyBac* copies are categorized into five groups described in the inset, according to the linkage to the herpesvirus-like sequence. For both species, *piggyBac* copies adjacent to the herpesvirus-like sequence are not clustered together, making it unlikely that *Teratorn*-like elements are fused with *piggyBac* transposon in those two species. The scale bars represent the number of substitutions per site.

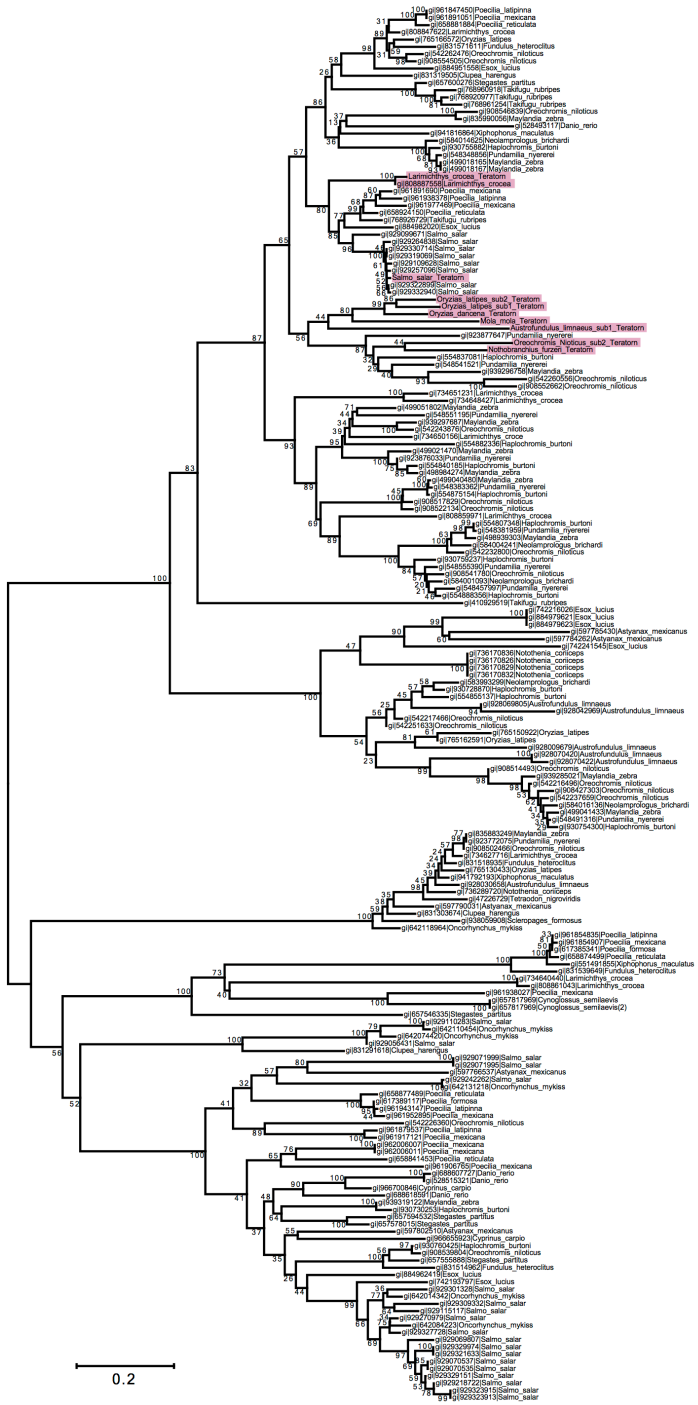


Figure 27 | Phylogenetic tree of *piggyBac*-like elements in teleost fishes. Neighbor-joining tree based on the amino acid sequence of *piggyBac*-like transposase genes in teleosts are shown. Poisson correction method was used as amino acid substitution model. All ambiguous positions were removed for each sequence pair. A total of 264 positions were used in the final dataset. The bar represents the number of substitutions pre site.

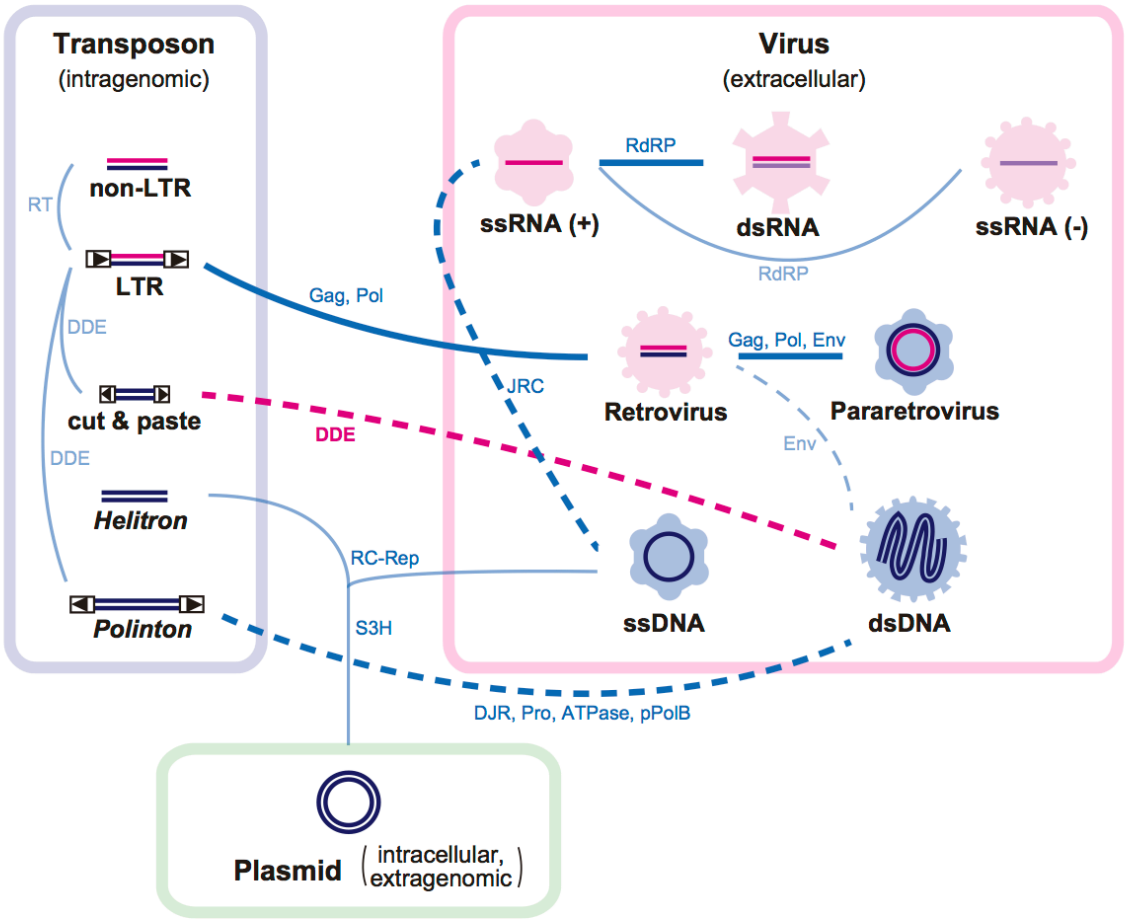


Figure 28 | Network-like evolutionary relationships in eukaryotic mobile genetic elements. Evolutionary relationships between different classes of mobile elements are connected by lines; Solid lines indicate that the evolutionary relationships are applied to almost all elements belonging to the class, while dashed lines indicate that the relationships are applied to only a part of elements belonging to the class. Bold lines indicate that the evolutionary relation is certain, while thin lines indicate that the evolutionary connection is tenuous. Shared genes are represented next to the lines. Abbreviations; ATPase, DNA packaging ATPase; DDE, DDE transposase; DJR, double jelly-roll capsid; Env, envelope protein; Gag, group-specific antigen; JRC, jelly-roll capsid; pPolB, protein-primed family B DNA polymerase; Pro, Ulp1-like cysteine protease; RC-Rep, rolling-circle replication initiation endonuclease; RdRP, RNA-dependent RNA polymerase; RT, reverse-transcriptase; S3H, superfamily 3 helicase.

Tables

Table 1 | Predicted ORFs inside *Teratorn* subtype 1

Gene	properties or putative functions	Evalue	start	end	direction	IcHV-1		CyHV-3		RaHV-1	
						ORF	Evalue	ORF	Evalue	ORF	Evalue
ORF1			3689	4048	+						
ORF2	myogenic factor MyoD2 (O. niloticus)	0.005	8022	8630	+						
ORF3			12667	13026	-						
ORF4			13388	13702	-						
ORF5			19344	20096	-						
ORF6	**ORF60**	**9.00E-22**	**20228**	**21406**	**+**	**ORF60**	**9.00E-22**	**ORF80L**	**6.00E-04**	**ORF84**	**5.00E-08**
ORF7	**ORF54**	**6.00E-05**	**21471**	**23450**	**+**	**ORF54**	**6.00E-05**	**ORF61**	**-**	**ORF75**	**-**
ORF8	***piggyBac*-like transposase**	**0**	**23875**	**26466**	**+**						
ORF9			29685	30569	-						
ORF10	**ORF56**	**4.00E-41**	**40287**	**44351**	**-**	**ORF56**	**1.00E-40**	**ORF107**	**2.00E-18**	**ORF73**	**2.00E-33**
ORF11	**DNA polymerase**	**2.00E-111**	**45181**	**50085**	**+**	**ORF57**	**2.00E-111**	**ORF79**	**7.00E-57**	**ORF72**	**6.00E-73**
ORF12			50203	51330	+						
ORF13	OTU-like cysteine protease domain (C. semilaevis)	2.00E-35	51603	53297	-						
ORF14			54098	54694	+						
ORF15			55227	56054	-						
ORF16	ORF73_protein kinase (IcHV-1)	4.00E-09	56053	58443	+	ORF73	4.00E-09				
ORF17			58410	59327	+						
ORF18	CD276 antigen-like (O. latipes)	8.00E-32	59157	61143	+						
ORF19	CD276 antigen-like (H. burtoni)	1.00E-19	62103	63705	+						
ORF20			64024	64353	-						
ORF21	HERV-H LTR-associating protein 2-like (P. nyererei)	1.00E-17	64751	65805	+						
ORF22			66376	70044	-						
ORF23	mcl-1 (O. mykiss)	1.00E-19	70917	71457	+						
ORF24	CDK2 (L. crocea)	2.00E-80	71862	72749	+						
ORF25	zinc finger protein 36, C3H1 type-like 2 (S. partitus)	2.00E-20	74275	75075	+						
ORF26	integrase / recombinase (Vibrio splendidus)	0.011	74362	74691	-						
ORF27			75467	76012	-						
ORF28	CAP-Gly domain containing linker protein 1-like (S. partitus)	0.02	76476	79142	-						
ORF29			78757	80472	-						
ORF30			81622	82077	-						
ORF31	ORF44	8.00E-18	82121	83569	-	ORF44	8.00E-18			ORF59	3.00E-05
ORF32			83417	86155	+						
ORF33			86700	87455	-						
ORF34			87418	88338	+						
ORF35	ORF34	7.00E-50	88079	89569	+	ORF34	4.00E-43			ORF49	5.00E-31
ORF36			89566	90429	+						
ORF37			91720	92091	-						
ORF38	**ORF37**	**6.00E-08**	**92640**	**94820**	**-**	**ORF37**	**3.00E-04**	**ORF90L**	**0.53**	**ORF52**	**-**
ORF39			94944	95378	+						
ORF40	**Major capsid protein**	**1.00E-22**	**95469**	**98927**	**+**	**ORF39**	**8.00E-21**	**ORF92**	**0.059**	**ORF54**	**2.00E-10**
ORF41			99000	99578	+						
ORF42	diguanylate cyclase (Firmicutes bacterium)	0.024	100296	101141	-						
ORF43			101152	101997	-						
ORF44			102065	103342	-						
ORF45	**Membrane glycoprotein**	**2.00E-89**	**104618**	**108688**	**-**	**ORF46**	**6.00E-71**	**ORF99**	**1.00E-07**	**ORF46**	**1.00E-08**
ORF46			108803	109411	-						
ORF47	myopalladin, partial, (O. latipes)	4.00E-20	109609	110190	+						
ORF48	**Capsid maturation protease**	**5.00E-13**	**110276**	**112294**	**+**	**ORF28**	**1.00E-08**	**ORF78**	**-**	**ORF63**	**0.006**
ORF49	unnamed protein (O. mykiss)	5.00E-29	112609	113673	+						
ORF50			113800	114390	+						
ORF51			114397	114924	+						
ORF52			114956	115129	+						
ORF53	thiopurine S-methyltransferase (Pseudomonas fluorescens)	0.027	115138	115542	+						
ORF54	zinc finger BED domain-containing protein 4 (L. crocea)	2.00E-27	115680	116837	+						

Table 1 Continued

Gene	properties or putative functions	Evalue	start	end	direction	IcHV-1		CyHV-3		RaHV-1	
						ORF	Evalue	ORF	Evalue	ORF	Evalue
ORF55	**Capsid triplex protein subunit 2**	**4.00E-05**	**117519**	**118388**	**-**	**ORF27**	**1.00E-04**	**ORF72**	**-**	**ORF95**	**-**
ORF56			118445	118855	-						
ORF57	**DNA helicase (UL9 homolog)**	**2.00E-47**	**118921**	**120687**	**-**	**ORF25**	**7.00E-25**	**ORF71**	**2.00E-35**	**ORF93**	**3.00E-26**
ORF58			120719	121948	-						
ORF59	putative permease YjgP/YjgQ family protein (E. coli)	2.00E-12	123657	124283	+						
ORF60	zinc finger and SCAN domain-containing protein 29 (A. platyrhynchos)	2.00E-18	124919	125347	-						
ORF61	type-2 angiotensin II receptor-like cxcr (O. afer)	3.00E-13	126419	127532	-						
ORF62			127937	129121	+						
ORF63	75-interferon_induced_dsRNA_activated protein kinase (H. microstoma)	1.00E-07	129414	130175	+						
ORF64			130819	131172	-						
ORF65			132599	133129	-						
ORF66	ORF57R (CyHV-3)	5.00E-10	133776	134714	-			ORF57R	1.00E-10		
ORF67			135058	135489	-						
ORF68			135683	137212	-						
ORF69			137209	140187	-						
ORF70	DNA methyltransferase (Lymphocystis disease virus 1)	7.00E-07	140334	141221	+						
ORF71	**Primase**	**1.00E-14**	**141798**	**144128**	**-**	**ORF63**	**1.00E-11**	**ORF46**	**9.00E-04**	**ORF87**	**1.00E-08**
ORF72			144317	144970	-						
ORF73	**DNA packaging terminase subunit 1**	**4.00E-41**	**145083**	**146207**	**-**	**ORF62**	**4.00E-41**	**ORF33**	**9.00E-27**	**ORF42**	**1.00E-21**
ORF74			146149	147051	+						
ORF75			147148	147954	+						
ORF76			148268	148777	+						
ORF77	**DNA packaging terminase subunit 1**	**2.00E-12**	**149071**	**149904**	**-**	**ORF62**	**1.00E-09**	**ORF33**	**5.00E-05**		
ORF78	cbp_p300-interacting transactivator (L. crocea)	9.00E-90	149947	150927	+						
ORF79	ORF70 (RaHV-2)	0.69	151279	152451	+						
ORF80			152882	153208	-						
ORF81	**ORF64**	**9.00E-13**	**153189**	**154802**	**-**	**ORF64**	**9.00E-13**	**ORF47**	**-**	**ORF88**	**-**
ORF82			154685	160591	-						
ORF83			160679	162142	+						
ORF84			162226	166377	+						
ORF85			166751	171127	-						
ORF86			171416	172927	-						
ORF87			173599	174402	+						
ORF88	serine/threonine-protein kinase pim3 (A. mexicanus)	8.00E-76	175073	176215	+						
ORF89	chloride channel CLIC-like, partial (S. salar)	6.00E-07	180402	180593	-						

Table 2 | Variants inside coding region of subtype 1 *Teratorn*.

#GeneName	frameshift	stop_gained	inframe_del	inframe_ins	nonsynonymous	synonymous
1	0	0	0	0	0	0
2-MyoD2	0	0	0	0	17	8
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	8	10
6-ORF60	1	0	0	0	32	24
7-ORF54	0	4	0	1	78	46
8-piggyBac	0	1	0	1	89	54
9	0	0	0	0	8	7
10_ORF56	0	1	0	0	119	126
11-DNApol	0	1	0	0	72	96
12	0	0	0	0	21	21
13-OTU	0	1	0	0	26	19
14	0	0	0	0	5	6
15	0	0	0	0	15	11
16-ORF73_kinase	1	0	0	0	37	38
17	0	0	0	0	16	20
18-CD276	0	1	0	0	26	11
19-CD276	0	2	0	0	29	11
20	0	0	0	0	6	2
21-HERV-H_LTR-ass_pro_2	0	0	0	0	19	7
22	0	0	0	0	54	32
23-Mcl	0	0	0	0	10	3
24-CDK2	0	0	0	0	11	9
25-ZFP36	0	0	0	0	23	10
27	0	1	0	0	8	5
28-CAP_Gly	0	0	0	0	49	22
29	3	0	0	0	23	20
30	0	0	0	0	6	8
31-ORF44	0	0	0	0	28	14
32	0	0	0	0	34	33
33	0	0	0	0	11	12
34	0	0	0	0	19	11
35-ORF34	0	0	0	0	17	19
36	1	1	0	0	15	12
37	0	1	0	0	7	3
38-ORF37	0	1	0	0	32	33
39	0	0	0	0	14	5
40-MajorCapsid	0	1	0	0	43	62
41	0	0	0	0	18	29
42-Diguanylate_cyclase	3	0	0	0	11	8
43	0	0	0	0	33	25
44	0	1	0	0	62	27
45-MemGly	0	1	0	0	53	69
46	0	0	0	0	10	15
47-myopalladin	0	0	0	0	9	6
48-CapMat	0	0	0	0	33	30
49	0	0	0	0	12	12
50	0	0	0	0	14	10
51	0	0	0	0	11	6
52	0	0	0	0	1	4
53-Thiopurine_S_methylase	0	0	0	0	6	2
54-ZF_BED_domain	3	3	0	0	23	11
55-triplex	0	1	0	0	3	4
57	0	0	0	0	12	8
58-DNAhel	0	1	0	0	13	21
59	0	0	0	0	21	12
60-permease_YjgP/YjgQ	0	1	0	0	16	14
61-ZF_SCA	0	1	0	0	25	5
62-CXCR	0	0	0	0	18	6
63	0	0	0	0	23	14
64-TARBP	0	2	0	0	7	7
65	0	0	0	0	15	5
66	0	0	0	0	11	6

Table 2 Continued

#GeneName	frameshift	stop_gained	inframe_del	inframe_ins	nonsynonymous	synonymous
67-ORF57R	0	0	0	0	14	13
68	0	0	0	0	4	2
69	0	0	0	0	31	23
70	2	3	0	0	73	26
71-methyltransferas	0	0	0	0	14	10
72-Primase	0	2	1	0	22	32
73	0	1	0	0	11	6
74-terminase	0	0	0	0	15	13
75	0	0	0	0	17	14
76	0	0	0	0	11	7
77	0	0	0	0	13	7
78-terminase	0	0	0	0	12	6
79-Cbp_p300	0	0	0	0	27	17
80-ORF70	0	0	0	0	23	28
81	0	0	0	0	4	3
82-ORF64	0	0	0	0	24	22
83	0	3	0	0	175	171
84	2	1	0	0	48	58
85	0	0	0	0	83	74
86	0	4	0	0	67	51
87	0	0	0	0	21	16
88	0	0	0	0	9	6
89-pim3	0	1	0	0	40	27
90-CLIC	0	0	0	0	6	4
TOTAL	**16**	**42**	**1**	**2**	**2251**	**1822**

Variant frequency <0.1 0.1-0.25 0.25-0.5 0.5-0.9 0.9<

Table 3 | Characteristics of *Teratorn*-like elements in teleosts.

Species	Intactness of ORFs	Copy No.	Validity of Genomic integration	*piggyBac*-herpes virus fusion	Subtypes
Salmo salar	intact	~8	integrated (contig-mediated link between *Teratorn*-like element and genomic region)	Some *piggyBac* copies exist near *Teratorn*-like element	
Oncorhynchus mykiss	degraded	–	integrated (ORF degradation)	–	
Esox lucius	partially degraded	~8	integrated (contig-mediated link between *Teratorn*-like element and genomic region)	–	
Periophthalmus magnuspinnatus	intact	–	unkonwn	–	
Pampus argenteus	unknown (contigs too short)	–	unknown	–	
Oryzias latipes	intact	~25, ~5	integrated (BAC sequencing)	Fused	2 subtypes
Austrofundulus limnaeus	intact	1-2, ~4	likely integrated (scaffold-mediated link between *Teratorn*-like element and genomic region)	some *piggyBac* copies exist near *Teratorn*-like element	2 subtypes
Nothobranchius furzeri	intact	1-2	unknown	Consistent link between *piggyBac* and *Teratorn*-like element	
Kryptolebias marmoratus	intact	–	unknown	–	
Amphilophus citrinellus	intact	–	likely integrated (scaffold-mediated link between *Teratorn*-like element and genomic region)	–	
Oreochromis niloticus	partially degraded	~12, ~2	integrated (fosmid sequencing)	Consistent link between *piggyBac* and *Teratorn*-like element for subtype 1	2 subtypes
Neolamprologus brichardi	intact	~2	likely integrated (scaffold-mediated link between *Teratorn*-like element and genomic region)	–	
Haplochromis burtoni	intact	~1	unknown	–	
Pundamilia nyererei	unknown (contigs too short)	–	unknown	–	
Melanochromis auratus	unknown (contigs too short)	–	unknown	–	

Table 3 Continued

Species	Intactness of ORFs	Copy No.	Validity of Genomic integration	*piggyBac*-herpes virus fusion	Subtypes
Cynoglossus semilaevis	degraded	–	integrated (ORF degradation)	some *piggyBac* copies exist near *Teratorn*-like element	
Cottus rhenanus	unknown (contigs too short)	–	unknown	–	
Larimichthys crocea	intact	~18	likely integrated (scaffold-mediated link between *Teratorn*-like element and genomic region)	Consistent link between *piggyBac* and *Teratorn*-like element	
Mola mola	intact	–	integrated (contig-mediated link between *Teratorn*-like element and genomic region, ORF degradation)	Consistent link between *piggyBac* and *Teratorn*-like element	

Table 4 | List of medaka strains used in this analysis.

Species	Collection site
Hd-rR*, O. latipes*	Aichi, Japan
HNI*, O. latipes*	Niigata, Japan
O. curvinotus	Sam A tsuen, Plover Cove Country Park, Hong Kong
O. luzonensis	Solsona, Ilocos Norte Province, Philippines
O. mekongensis	Udon Thani, Thailand
O. dancena	Chidambaram, Tamil Nadu, India
O. javanicus	Penang, Malaysia
O. celebensis	Malino river, Sulawesi, Indonesia
O. sarasinorum	Lake Lindu, Sulawesi, Indonesia
O. nigrimas	Tentena, Lake Poso, Sulawesi, Indonesia
O. matanensis	Soroako, Lake Matano, Sulawesi, Indonesia
O. marmoratus	Timampuu, Lake Towuti, Sulawesi, Indonesia
O. profundicola	Timampu, Lake Towuti, Sulawesi, Indonesia

Table 5 | Parameters used for phylogenetic analyses.

Analysis	*piggyBac* copies in medaka genome	*Teratorn* in *Herpesvirales*	*Teratorn* in *Alloherpesviridae*	*Teratorn* in medaka inbred strain	*Teratorn* in medaka related species
Sequence	Transposase	Terminase	Terminase, DNApol, DNAhel, Major capsid	16 *Teratorn* genes (consensus, 3rd codon)*1	six *Teratorn* genes*2
Figure	Fig. 6b	Fig. 8a	Fig. 8b	Fig. 20c	Fig. 21c
How to gain	blast	GenBank	GenBank	consensus calling & blast	PCR
Species	Hd-rR	Herpesvirales, T4 phage	Alloherpesviridae	Hd-rR, HNI, Kaga, HSOK, Nilan	medaka related species
DNA or Protein	Nucleotide	Protein	Protein	Nucleotide	Nucleotide
Trimming	–	trimAl -strictplus	trimAl -strictplus	–	–
Method *5	NJ	ML	ML	ML	ML
Substitution model *6	Jukes-Cantor	Le_Gascuel_2008	Le_Gascuel_2008	Tamura-Nei	Tamura-Nei
Evolutionary rate variance (No. categories)	uniform	InvGamma (5)	InvGamma (5)	InvGamma (5)	uniform
No. of bootstrap replicates (ML, NJ)	1000	1000	1000	1000	1000

Analysis	phylogeny of teleosts	*Teratorn*-like elements and alloherpesviruses	*piggyBac* copies in selected species	*piggyBac* in teleosts
Sequence	nine host genes*3	Terminase, DNApol, DNAhel, Major capsid	*piggyBac* copies	*piggyBac* transposase gene
Figure	Fig. 22a	Fig. 22b	Fig. 24c, 25b, 26	Fig. 27
How to gain	blast, GenBank	GenBank, blast	blast	GenBank
Species	teleost fishes	*Alloherpesviridae* + *Teratorn*-like elements	*L. crocea, N. furzeri, A. limnaeus, S. salar*	teleost fishes
DNA or Protein	Nucleotide	Protein	Nucleotide	Protein
Trimming	–	trimAl -strictplus	–	trimAl -strict
Method	Bayes	ML	NJ	NJ
Substitution model	General time reversible	Le_Gascuel_2008	Jukes-Cantor	Poisson-Correction
Evolutionary rate variance (No. categories)	InvGamma (4)	InvGamma (5)	uniform	uniform
No. of bootstrap replicates (ML, NJ)	–	1000	1000	1000
No. of generations (Bayes)	500,000			
Sampling rate (Bayes)	200			

*1 : *piggyBac*, DNApol, DNAhel, Primase, Terminase, Major capsid, Capsid triplex, Envelope, Protease, ORF34, ORF37, ORF44, ORF54, ORF56, ORF60, ORF64

*2 : *piggyBac*, DNApol, DNAhel, Major capsid, Envelope, Terminase

*3 : *glyt*, *myh6*, *plagl2*, *ptr*, *rag1*, *screb2*, *sh3px3*, *tbr*, *zic1* (Near T. J. et al., 2012[114])

*4 : Software; Maximum-likelihood (ML) and Neighbor-joining (NJ) analysis, MEGA6.0; Bayesian inference, MrBayes3.2

Table 6 | Whole-genome shotgun sequencing data used in this study.

Species	Accession
Hd-rR (*O. latipes*)	DRR002213
HNI (*O. latipes*)	DRR002216
Kaga (*O. latipes*)	DRR002226
HSOK (*O. latipes*)	DRR002222
Nilan (*O. latipes*)	DRR002230
L. crocea	SRR1258892, SRR1258893
S. salar	SRR1264541, SRR1264542, SRR1264544
E. lucius	SRR1931760, SRR1945106
O. niloticus	SRR071597, SRR071603
N. brichardi	SRR077327, SRR077332
H. burtoni	SRR077264, SRR077270, SRR077276
N. furzeri	ERR583467, ERR583468, SRR1246172, SRR1246176
A. limnaeus	SRR2006331

Table 7 | Primer sequences used in this analysis.

Experiment	Name	Primer sequence
BAC Screening	Tera_LeftR1	GTAAAACAGCGGAGGGAATGGGTTTCGTGC
	Tera_LeftR2	AGACCCCAAAGCACATGCCAACCTAAG
	Tera_RightF	GGTGTACCTGTCTGGGGTCAATACCAAAGG
	13I17_LeftF	AAGGCTGTGTCCCTCTGCAATGAGACTGTTA
	13I17_RightR	GACAGCTACAAGTCTTGGTATGAGAGAGTC
	14A10_LeftF	CTCTGGATGGAAAAGACTTCTAACCAAGTGG
	14A10_RightR	GTCCAGAAAAGCTGTTTTACACGGAGTCCTA
	6L21_LeftF	ACAGTGCTTCATAGAGAAACGCTTCCAC
	6L21_RightR	CAGTGTTTGGGATCTTTGAAAAGGTCAG
	11H24_LeftF	AGAATGAGCTCCAACAATCCCTTCATGC
	11H24_RightF	AAACCAGTGAGGCATTACCAGCTTTTGC
	11H24_RightR	TGACGAAGAAATACGGTATCCAGCTGTC
	73I9_LeftF	GGGAATGGTGTGAACAGTTTGGAGTTTC
	73I9_RightF	TCCACATAGTGCCATCTAGGATTTCAGG
	73I9_RightR	GGAGCGGAGTAAGCAGTTTTACATAACCC
	85H23_LeftF	GTTGTGTCACTACGCATTCTTTCTCACC
	85H23_RightR	GTCTTTCAAAATCCACACAGGCCAGGTC
Indicator plasmid	AcGFP_in_fusion_F	GGACTCAGATCTCGAGCATTGATGAATGAGACGGCTTTG
	AcGFP_in_fusion_R	GTCGACTGCAGAATTCTGTACTCATTGAACTTTGGATGGTC
	Sub1_LeftTIR_F	CATTGATGAATGAGACGGCTTTGATTTGGA
	Sub1_LeftTIR_R	TTATCGATCCCTGGTGGTCGTGCGCTGTAAG
	Sub1_puroGFP_F	CACCAGGGATCGATAACCGTATTACCGCCATGC
	Sub1_puroGFP_R	CGCTTTACCCGCCTTTGAGTGAGCTGATACCGC
	Sub1_RightTIR_F	AAAGGCGGGTAAAGCGAGACCCCAGGTGGTG
	Sub1_RightTIR_R	TGTACTCATTGAACTTTGGATGGTCACACA
	Sub1_internal_TIR_F	gaacatcaATCGATCTGGTGGTCGTGCGCTGTAAG
	Sub1_internal_TIR_R	gttacgttATCGATCGGATTACTGTGTCAAAGTGC
	Sub2_LeftTIR_F	CAGTAAGGCCAGTAGATGTAGCGGATGTG
	Sub2_LeftTIR_R	GTTATCGATTCGTAGAAGTAAGGTTCGGGTCGCGTG
	Sub2_puroGFP_F	CTTCTACGAATCGATAACCGTATTACCGCCATGC
	Sub2_puroGFP_R	ACCTGATGCCGCCTTTGAGTGAGCTGATACCGC
	Sub2_RightTIR_F	ACCTGATGCCGCCTTTGAGTGAGCTGATACCGC
	Sub2_RightTIR_R	ACCTGATGCCGCCTTTGAGTGAGCTGATACCGC
	Sub2_internal_TIR_F	GAACATCAATCGATCTCATAGAAGTAAGAGTGGGGTC
	Sub2_internal_TIR_R	GTTACGTTATCGATTAGATCACGTGAGGTGAATGAC
Helper plasmid	Sub1_TPase_CDS_Kozak_F	GGAATTCCGCCACCATGAACAAAGGCCGCAAAAGAAC
	Sub1_TPase_CDS_R	GTCGACTCACTGGGACGCGTCTGTG
	Sub2_Tpase_CDS_Kozak_F	GAATTCCGCCACCATGGGTCCCAAAAGACCTCAAAG
	Sub2_TPase_CDS_R	GGTCGACCTAGTCTGTGTCACTGTTGGAATCG
(EPTS)LM-PCR	LMPCR-adaptor-long-oligo	GACCCGGGAGATCTGAATTCAGTGGCACAGCAGTTAGG
	LMPCR-adaptor-short-oligo	(p)CCTAACTGCTGTGCCACTGAATTCAGATCTCCC
	Sub1-LMPCR-LeftTIR-biotin-oligo-F	(bio-) TGTAAAACAGCGGAGGGAATGGAGAG
	Sub2-LMPCR-LeftTIR-biotin-oligo-F	(bio-) TGACAATCGAGTGAACATTCTGACAG
	Sub1-LMPCR-LeftTIR-primer	AACCTAAGCGAGCGGAAG
	Sub2-LMPCR-LeftTIR-primer	ACCCTTCATTACAGCGTAGG
	LMPCR-common-primer	GACCCGGGAGATCTGAATTC
RT-PCR	piggyBac_F	TCAGAGAGGTGTGGGAAGAGTG
	piggyBac_R	GTCATTCTCCTGCAGCTGTACG
	DNA_polymerase_F	CACAGGTGCAGGATCACGCTCAACATAG
	DNA_polymerase_R	GGAGGATCAGCGTCGCATACTCAAAGC

Experiment	Name	Primer sequence
RT-PCR	DNA_helicase_F	CACCGTGTTGAACTGGAGGTAGGAGAGTG
	DNA_helicase_R	TTTGTCGAGCGCTACAGGAGATGCTACC
	Terminase_F	GTTGCTGCTGGACAGAAAGTGAGCGAAG
	Terminase_R	GATGGAGCTCTTGAGAGGAGTCGTGCTG
	Major_capsid_F	TTCAAGCAGCAGAAGGACACGAGC
	Major_capsid_R	TCAGCACAGAGTCCAGCTTCTCC
	Capsid_triplex_F	ACCGCTACGTTCCCACTCTACAC
	Capsid_triplex_R	GAATCTGTCGTTCACCAAAGACG
	Membrane_glycoprotein_F	GCGTCCCTGAGGAGTCCAAAGTTCTTG
	Membrane_glycoprotein_R	AGAGACTACGGCAAGCTGTGCGATTCAG
	ZFP36_F	TCTGAACACCCAACAGCTCTCA
	ZFP36_R	GGAAACGTCCATCTGTCGGTAG
	CXCR_F	GTTCAGTGTGGTACCCATGCTTC
	CXCR_R	CGACTTTGAAGGTTTCGTGATTG
	DNA_methyltransferase_F	CTCGTCCTCTTTCGGTACTTGG
	DNA_methyltransferase_R	GAGACACGGGGCTGATATAGTGA
	CDK_F	AGTTGTCTACAGGGTGCGAGATG
	CDK_R	CTCCCAAACTCCACATGTCTACG
	pim_F	GCAAAGTGTTTCTGGGAAAGAGG
	pim_R	TGTCGTTTCCAAATGGTATGTCC
	ZnSCAN_F	TGGAGCGATGATGAGGTTAAATG
	ZnSCAN_R	ATAAATTCCGTTTGCCCTTACCTG
	bactin_F	GATGAAGCCCAGAGCAAGAG
	bactin_R	AGGAAGGAAGGCTGGAAGAG
qRT-PCR	piggyBac_qPCR_F	GGTTCCTTTCAAAGGACGTTG
	piggyBac_qPCR_R	CCTCCAAGCGTAGCTCGTC
	DNA_polymerase_qPCR_F	CGATTTCGCCAGCATGTACC
	DNA_polymerase_qPCR_R	ACTTGTACACGACCCATCCG
	DNA_helicase_qPCR_F	CAGCAGGATCTTCGCTCGG
	DNA_helicase_qPCR_R	GTCATGAGCCCGTACTCGTC
	Terminase_qPCR_F	CTTGAGAGGAGTCGTGCTGG
	Terminase_qPCR_R	CGGGTCCTGGTAATCGTAGC
	Major_capsid_qPCR_F	GGTTCACGAGGCGTTACAAG
	Major_capsid_qPCR_R	GGTTCACGAGGCGTTACAAG
	Capsid_triplex_qPCR_F	CGCTACGTTCCCACTCTACA
	Capsid_triplex_qPCR_R	TTCAGGAACGGTCTGTCCAC
	Membrane_glycoprotein_qPCR_F	TCTGGCATTGCACGTGTCTC
	Membrane_glycoprotein_qPCR_R	GTTCTTGTTGGCGATCTCGG
	ZFP36_qPCR_F	GATATGTTCACCAGAGCTGACAC
	ZFP36_qPCR_R	GGTAGCCGTCCTTGTTCAGA
	CXCR_qPCR_F	GGTGATAACTTTCTGCCTGC
	CXCR_qPCR_R	GATATGTTCACCAGAGCTGACAC
	DNA_methyltransferase_qPCR_F	CTCCGTGCGAAAACTACAGC
	DNA_methyltransferase_qPCR_R	TGCGAGGGTTTTGTCCATGA
	CDK_qPCR_F	CAAAGCTATTCACCGCCAGC
	CDK_qPCR_R	AGCACGTTACTCGGAACCAG
	pim_qPCR_F	CTCCTACTGGGCGAGCTTAG
	pim_qPCR_R	TCAGCATCGGTCTCTCTTCG
	ZnSCAN_qPCR_F	GGTAGCCGTCCTTGTTCAGA
	ZnSCAN_qPCR_R	GCCACGAGCGCTTAAAACC
	bactin_qPCR_F	TGCCGCACTGGTTGTTGACAACG
	bactin_qPCR_R	CCAGGCACCAGGGTGTCATGG

Experiment	Name	Primer sequence
Antibodies	Major_capsid_GST_F	AAGAATTCGTGCGAGGGGGAGACGTG
	Major_capsid_GST_R	AAGTCGACCTACTGCATGATGTCCGTGGC
	Triplex_GST_F	AAGAATTCGACAGCAACCGCTACGTTC
	Triplex_GST_R	AAGTCGACCTACATGCTTGAGCGCCTGC
	Membrane_glycoprotein_GST_F	AAGAATTCAAGGAGCCCCACAGGGAG
	Membrane_glycoprotein_GST_R	AAGTCGACCTACGCCGGTTTCACGATCAC
Medaka related species	DNA_polymerase_deg_F1	GTAARCTNYDSAAGCTRRCSGRCGT
	DNA_polymerase_deg_F2	GCSGTGCTBSTSTACAACTTGGTBG
	DNA_polymerase_deg_R1	CCAKGACTTTRCACTCYMCCTCCTG
	DNA_polymerase_deg_R2	CKGTTGAGRTARTRCTCNACMGAGG
	Major_capsid_deg_F1	TCACSKTNGACGAMRTSGACTAYTG
	Major_capsid_deg_F2	CVGGCWTGTCSTAYCTRMTGTTCAG
	Major_capsid_deg_F3	GGCWTGTCSTAYCTRHTGTTCAG
	Major_capsid_deg_R1	CTGAACAKYAGRTASGACAWGCCBG
	Major_capsid_deg_R2	GTNGTGAARATGTAVGGYCTGCGGT
	Major_capsid_deg_R3	TAAGTACCRCAGAARGCMACYCCNA
	Major_capsid_deg_R4	AGTACCDCAGAARGCMACYCCNA
	Membrane_glycoprotein_deg_F1	GCNATTTGCCTSAGACAGTTYTCHA
	Membrane_glycoprotein_deg_F2	CRTCATYTGATGSGTCACGCTYTGC
	Membrane_glycoprotein_deg_F3	GCGTCYCTGAGGAGTCCRAAGTTYTTG
	Membrane_glycoprotein_deg_R1	GTGAACGACRGNTACRYGNTGGAYA
	Membrane_glycoprotein_deg_R2	CCCACSASNGAYMSCARTCACTAYA
	Membrane_glycoprotein_deg_R3	AGAGACTACGGCAAGCTGTGCGATTCRG
	DNA_helicase_deg_F1	AGTTCATCTCCACBCKTTTRCGRTA
	DNA_helicase_deg_F2	TARAAMNCMACMCGWCCSACNKWST
	DNA_helicase_deg_F3	RGTTCATCTCCWCBCKTTYRSGRKW
	DNA_helicase_deg_R1	GGCACATGGAYYMSAACGGNATACA
	DNA_helicase_deg_R2	TSTTTGTCAACAGRMGRCAGTTCSA
	DNA_helicase_deg_R3	TSTTTRTCAACAGRMGRCARTTCSA
	Terminase_deg_F1	ATCTCRTCTCTKGTGTGRCAHAGRA
	Terminase_deg_F2	TACATGWRCACGKCCATRKWGGAGC
	Terminase_deg_F3	AWCTCDTCYCTKGTGTGRCAHAAVA
	Terminase_deg_F4	TACATGHVCACRKYCATRKNSGANC
	Terminase_deg_F5	TYTCRTCTCTKGTGTGRCANAGVA
	Terminase_deg_R1	ATCTCRTCTCTKGTGTGRCAHAGRA
	Terminase_deg_R2	TACATGWRCACGKCCATRKWGGAGC
	Terminase_deg_R3	ARRTACATGYTVGGRAARCACGTNR
	Terminase_deg_R4	RGTACATGCTRGGRAARCACGTSR
	piggyBac_F	TCAGAGAGGTGTGGGAAGAGTG
	piggyBac_R	GTCATTCTCCTGCAGCTGTACG
	Major_capsid_sub2_F	CCGGCTTGTCCTATCTGTTGTTCAG
	Major_capsid_sub2_R	TAAGTACCACAGAAAGCCACCCCGA
	Membrane_glycoprotein_sub2_F	AGAGACTACGGCAAGCTGTGCGATTCAG
	Membrane_glycoprotein_sub2_R	GCGTCTCTGAGGAGTCCGAAGTTTTTG

References

1. Siefert, J. L. Defining the Mobilome. *Methods Mol. Biol.* **532,** 13–27 (2009).
2. Suttle, C. A. Marine viruses--major players in the global ecosystem. *Nat. Rev. Microbiol.* **5,** 801–812 (2007).
3. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
4. Schnable P. S. *et al.* The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Nature* **326,** 1112–1115 (2005).
5. Levin, H. L. & Moran, J. V. Dynamic interactions between transposable elements and their hosts. *Nat. Rev. Genet.* **12,** 615–627 (2011).
6. tenOever, B. R. The Evolution of Antiviral Defense Systems. *Cell Host Microbe* **19,** 142–9 (2016).
7. Koonin, E. V. & Krupovic, M. Evolution of adaptive immunity from transposable elements combined with innate immune systems. *Nat. Rev. Genet.* **16,** 184–192 (2015).
8. Shabalina, S. A. & Koonin, E. V. Origins and evolution of eukaryotic RNA interference. *Trends in Ecology and Evolution* **23,** 578–587 (2008).
9. Jacobs, F. M. J. *et al.* An evolutionary arms race between KRAB zinc finger genes *ZNF91/93* and SVA/L1 retrotransposons. *Nature* **516,** 242–245 (2014).
10. Imbeault, M., Helleboid, P. Y. & Trono, D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* 1–17 (2017). doi:10.1038/nature21683
11. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.* **advance on,** 71–86 (2016).
12. Duggal, N. K. & Emerman, M. Evolutionary conflicts between viruses and restriction factors shape immunity. *Nat. Rev. Immunol.* **12,** 687–95 (2012).
13. Aswad, A. & Katzourakis, A. Paleovirology and virally derived immunity. *Trends Ecol. Evol.* **27,** 627–636 (2012).
14. Roossinck, M. J. The good viruses: viral mutualistic symbioses. *Nat. Rev. Microbiol.* **9,** 99–108 (2011).

15. Frost, L. S., Leplae, R., Summers, A. O. & Toussaint, A. Mobile genetic elements: the agents of open source evolution. *Nat.Rev.Microbiol.* **3,** 722–732 (2005).
16. Feschotte, C. Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.* **9,** 397–405 (2008).
17. Lynch, V. J., Leclerc, R. D., May, G. & Wagner, G. P. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nat. Genet.* **43,** 1154–1159 (2011).
18. Kunarso, G. *et al.* Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.* **42,** 631–634 (2010).
19. Koonin, E. V. Viruses and mobile elements as drivers of evolutionary transitions. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **371,** 555–571 (2016).
20. Huff, J. T., Zilberman, D. & Roy, S. W. Mechanism for DNA transposons to generate introns on genomic scales. *Nature* **538,** 533–536 (2016).
21. Strand, M. R. & Burke, G. R. Polydnavirus-wasp associations: Evolution, genome organization, and function. *Curr. Opin. Virol.* **3,** 587–594 (2013).
22. Ono, R. *et al.* Deletion of Peg10, an imprinted gene acquired from a retrotransposon, causes early embryonic lethality. *Nat. Genet.* **38,** 101–106 (2006).
23. Feschotte, C. & Gilbert, C. Endogenous viruses: insights into viral evolution and impact on host biology. *Nat. Rev. Genet.* **13,** 283–296 (2012).
24. Koonin, E. V., Dolja, V. V. & Krupovic, M. Origins and evolution of viruses of eukaryotes: The ultimate modularity. *Virology* **479–480,** 2–25 (2015).
25. Koonin, E. V. & Dolja, V. V. A virocentric perspective on the evolution of life. *Curr. Opin. Virol.* **3,** 546–557 (2013).
26. Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8,** 973–982 (2007).
27. King, A. M. Q., Lefkowitz, E., Adams, M. J. & Carstens, E. B. (eds) *Virus Taxonomy: ninth report of the International Committee on Taxonomy of Viruses.* 1338 (Elsevier–Academic, London, 2011). doi:10.1016/B978-0-12-384684-6.00115-4
28. Koonin, E. V & Dolja, V. V. Virus world as an evolutionary network of viruses and capsidless selfish elements. *Microbiol. Mol. Biol. Rev.* **78,** 278–

303 (2014).

29. Krupovic, M., Prangishvili, D., Hendrix, R. W. & Bamford, D. H. Genomics of Bacterial and Archaeal Viruses: Dynamics within the Prokaryotic Virosphere. *Microbiol. Mol. Biol. Rev.* **75,** 610–635 (2011).
30. Malik, H. S., Henikoff, S. & Eickbush, T. H. Poised for contagion: Evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome Res.* **10,** 1307–1318 (2000).
31. Stoye, J. P. Studies of endogenous retroviruses reveal a continuing evolutionary saga. *Nat. Rev. Microbiol.* **10,** 395–406 (2012).
32. Jern, P. & Coffin, J. M. Effects of Retroviruses on Host Genome Function. *Annu. Rev. Genet.* **42,** 709–732 (2008).
33. Kapitonov, V. V & Jurka, J. Self-synthesizing DNA transposons in eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 4540–4545 (2006).
34. Krupovic, M. & Koonin, E. V. Polintons: a hotbed of eukaryotic virus, transposon and plasmid evolution. *Nat. Rev. Microbiol.* **13,** 105–115 (2015).
35. Yutin, N., Shevchenko, S., Kapitonov, V., Krupovic, M. & Koonin, E. V. A novel group of diverse Polinton-like viruses discovered by metagenome analysis. *BMC Biol.* **13,** 95 (2015).
36. Yutin, N., Raoult, D. & Koonin, E. V. Virophages, polintons, and transpovirons: a complex evolutionary network of diverse selfish genetic elements with different reproduction strategies. *Virol. J.* **10,** 158 (2013).
37. Fischer, M. G. & Suttle, C. A. A virophage at the origin of large DNA transposons. *Science* **332,** 231–234 (2011).
38. Feschotte, C. & Pritham, E. J. DNA transposons and the evolution of eukaryotic genomes. *Annu. Rev. Genet.* **41,** 331–368 (2007).
39. Kapitonov, V. V & Jurka, J. Rolling-circle transposons in eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 8714–9 (2001).
40. Pritham, E. J., Putliwala, T. & Feschotte, C. *Mavericks*, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. *Gene* **390,** 3–17 (2007).
41. Moriyama, Y. *et al.* The Medaka *zic1*/*zic4* Mutant Provides Molecular Insights into Teleost Caudal Fin Evolution. *Curr. Biol.* **22,** 601–607 (2012).
42. Kondo, S. *et al.* The medaka *rs-3* locus required for scale development

encodes ectodysplasin-A receptor. *Curr. Biol.* **11,** 1202–1206 (2001).

43. Hashimoto, H. *et al.* Polycystic kidney disease in the medaka (*Oryzias latipes*) pc mutant caused by a mutation in the Gli-similar3 (glis3) gene. *PLoS One* **4,** e6299 (2009).
44. Kamura, K. *et al.* Pkd1l1 complexes with Pkd2 on motile cilia and functions to establish the left-right axis. *Development* **138,** 1121–1129 (2011).
45. Kasahara, M. *et al.* The medaka draft genome and insights into vertebrate genome evolution. *Nature* **447,** 714–719 (2007).
46. Keith, J. H., Schaeper, C. A., Fraser, T. S. & Fraser, M. J. Mutational analysis of highly conserved aspartate residues essential to the catalytic core of the *piggyBac* transposase. *BMC Mol. Biol.* **9,** 73 (2008).
47. Neumann, P., Koblizkova, A., Navratilova, A. & Macas, J. Significant Expansion of *Vicia pannonica* Genome Size Mediated by Amplification of a Single Type of Giant Retroelement. *Genetics* **173,** 1047–1056 (2006).
48. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78–94 (1997).
49. Besemer, J., Lomsadze, A. & Borodovsky, M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.* **29,** 2607–2618 (2001).
50. Arvin A. et al. (eds) *Human Herpesviruses: Biology, Therapy, and Immunoprophylaxis.* (Cambridge Univ. Press, 2007).
51. Aoki, T. *et al.* Genome sequences of three koi herpesvirus isolates representing the expanding distribution of an emerging disease threatening koi and common carp worldwide. *J. Virol.* **81,** 5058–65 (2007).
52. Philippe, N. *et al.* Pandoraviruses : Amoeba Viruses with Genomes Up to 2.5 Mb Reaching That of Parasitic Eukaryotes. *Science* **341,** 281–286 (2013).
53. Magel, G. D. & Tyring, S. (eds) *Herpesviridae - A Look Into This Unique Family of Viruses.* (InTech, 2012).
54. Hanson, L., Dishon, A. & Kotler, M. Herpesviruses that infect fish. *Viruses* **3,** 2160–2191 (2011).
55. Koga, A. *et al.* The medaka fish *Tol2* transposable element can undergo excision in human and mouse cells. *J. Hum. Genet.* **48,** 231–235 (2003).

56. Koga, A. *et al.* The *Tol1* transposable element of the medaka fish moves in human and mouse cells. *J. Hum. Genet.* **52,** 628–635 (2007).
57. Ding, S. *et al.* Efficient Transposition of the *piggyBac* (*PB*) Transposon in Mammalian Cells and Mice. *Cell* **122,** 473–483 (2005).
58. Yergeau, D. A., Kuliyev, E. & Mead, P. E. Injection-mediated transposon transgenesis in *Xenopus tropicalis* and the identification of integration sites by modified extension primer tag selection (EPTS) linker-mediated PCR. *Nat. Protoc.* **2,** 2975–2986 (2007).
59. Rao, V. B. & Feiss, M. The bacteriophage DNA packaging motor. *Annu. Rev. Genet.* **42,** 647–81 (2008).
60. Selvarajan Sigamani, S., Zhao, H., Kamau, Y. N., Baines, J. D. & Tang, L. The structure of the herpes simplex virus DNA-packaging terminase pUL15 nuclease domain suggests an evolutionary lineage among eukaryotic and prokaryotic viruses. *J. Virol.* **87,** 7140–8 (2013).
61. Cheng, H., Shen, N., Pei, J. & Grishin, N. V. Double-stranded DNA bacteriophage prohead protease is homologous to herpesvirus protease. *Protein Sci.* **13,** 2260–9 (2004).
62. Huet, A. *et al.* Extensive subunit contacts underpin herpesvirus capsid stability and interior-to-exterior allostery. *Nat. Struct. Mol. Biol.* **23,** 531–539 (2016).
63. Booy, F. P., Trus, B. L., Davidson, A. J. & Steven, A. C. The capsid Architecture of Channel Catfish Virus, an Evolutionarily Distant Herpesvirus, Is Largely Conserved in the Absence of Dicernible Sequence Homology with Herpes Simplex Virus. *Virology* **215,** 134–141 (1996).
64. Spivakov, M. *et al.* Genomic and phenotypic characterization of a wild medaka population: towards the establishment of an isogenic population genetic resource in fish. *G3 (Bethesda).* **4,** 433–45 (2014).
65. Yu, Y. *et al.* Induction of human herpesvirus-8 DNA replication and transcription by butyrate and TPA in BCBL-1 cells. *J. Gen. Virol.* **80,** 83–90 (1999).
66. Chen, J. *et al.* Activation of latent Kaposi's sarcoma-associated herpesvirus by demethylation of the promoter of the lytic transactivator. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 4119–24 (2001).

67. Juranic Lisnic, V. *et al.* Dual Analysis of the Murine Cytomegalovirus and Host Cell Transcriptomes Reveal New Aspects of the Virus-Host Cell Interface. *PLoS Pathog.* **9,** (2013).
68. Nakamura, R. *et al.* Large hypomethylated domains serve as strong repressive machinery for key developmental genes in vertebrates. *Development* **141,** 2568–80 (2014).
69. Morissette, G. & Flamand, L. Herpesviruses and chromosomal integration. *J. Virol.* **84,** 12100–12109 (2010).
70. Aswad, A. & Katzourakis, A. The First Endogenous Herpesvirus, Identified in the Tarsier Genome, and Novel Sequences from Primate Rhadinoviruses and Lymphocryptoviruses. *PLoS Genet.* **10,** e1004332 (2014).
71. Kaufer, B. B. & Flamand, L. Chromosomally integrated HHV-6: Impact on virus, cell and organismal biology. *Curr. Opin. Virol.* **9,** 111–118 (2014).
72. Arbuckle, J. H. *et al.* The latent human herpesvirus-6A genome specifically integrates in telomeres of human chromosomes in vivo and in vitro. *Proc Natl Acad Sci U S A* **107,** 5563–5568 (2010).
73. Huang, Y. *et al.* Human telomeres that carry an integrated copy of human herpesvirus 6 are often short and unstable, facilitating release of the viral genome from the chromosome. *Nucleic Acids Res.* **42,** 315–327 (2014).
74. Krupovic, M., Bamford, D. H. & Koonin, E. V. Conservation of major and minor jelly-roll capsid proteins in Polinton (Maverick) transposons suggests that they are bona fide viruses. *Biol. Direct* **9,** 6 (2014).
75. Blanc, G., Gallot-Lavallée, L. & Maumus, F. Provirophages in the Bigelowiella genome bear testimony to past encounters with giant viruses. *Proc. Natl. Acad. Sci.* **112,** E5318–E5326 (2015).
76. Fischer, M. G. & Hackl, T. Host genome integration and giant virus-induced reactivation of the virophage mavirus. *Nature* **540,** 288–291 (2016).
77. Macfarlan, T. S. *et al.* Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* **487,** 57–63 (2012).
78. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351,** 1083–1087 (2016).
79. Wang, J. *et al.* Primate-specific endogenous retrovirus-driven transcription

defines naive-like stem cells. *Nature* **516,** 405–9 (2014).

80. Magiorkinis, G., Gifford, R. J., Katzourakis, A., De Ranter, J. & Belshaw, R. Env-less endogenous retroviruses are genomic superspreaders. *Proc Natl Acad Sci U S A* **109,** 7385–7390 (2012).
81. Ribet, D. *et al.* An infectious progenitor for the murine IAP retrotransposon: Emergence of an intracellular genetic parasite from an ancient retrovirus. *Genome Res.* **18,** 597–609 (2008).
82. Katzourakis, A. & Gifford, R. J. Endogenous viral elements in animal genomes. *PLoS Genet.* **6,** e1001191 (2010).
83. Geering, A. D. W. *et al.* Endogenous florendoviruses are major components of plant genomes and hallmarks of virus evolution. *Nat. Commun.* **5,** 5269 (2014).
84. Liu, H. *et al.* Widespread Horizontal Gene Transfer from Circular Single-stranded DNA Viruses to Eukaryotic Genomes. *BMC Evol. Biol.* **11,** 276 (2011).
85. Krupovic, M. & Forterre, P. Single-stranded DNA viruses employ a variety of mechanisms for integration into host genomes. *Ann. N. Y. Acad. Sci.* **1341,** n/a-n/a (2015).
86. Staginnus, C. & Richert-Pöggeler, K. R. Endogenous pararetroviruses: two-faced travelers in the plant genome. *Trends Plant Sci.* **11,** 485–491 (2006).
87. Takeda, H. & Shimada, A. The art of medaka genetics and genomics: what makes them so unique? *Annu. Rev. Genet.* **44,** 217–241 (2010).
88. Kofler, R., Pandey, R. V. & Schlötterer, C. PoPoolation2: Identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics* **27,** 3435–3436 (2011).
89. 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).
90. Ao, J. *et al.* Genome Sequencing of the Perciform Fish *Larimichthys crocea* Provides Insights into Molecular and Genetic Mechanisms of Stress Adaptation. *PLOS Genet.* **11,** e1005118 (2015).

91. Waltzek, T. B. *et al.* Phylogenetic relationships in the family *Alloherpesviridae*. *Dis. Aquat. Organ.* **84,** 179–194 (2009).
92. Fraser, M. J., Smith, G. E. & Summers, M. D. Acquisition of Host Cell DNA Sequences by Baculoviruses: Relationship Between Host DNA Insertions and FP Mutants of *Autographa californica* and *Galleria mellonella* Nuclear Polyhedrosis Viruses. *J. Virol.* **47,** 287–300 (1983).
93. Gilbert, C. *et al.* Population genomics supports baculoviruses as vectors of horizontal transfer of insect transposons. *Nat. Commun.* **5,** 3348 (2014).
94. Piskurek, O. & Okada, N. Poxviruses as possible vectors for horizontal transfer of retroposons from reptiles to mammals. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 12046–12051 (2007).
95. Sun, C., Feschotte, C., Wu, Z. & Mueller, R. L. DNA transposons have colonized the genome of the giant virus *Pandoravirus salinus*. *BMC Biol.* **13,** 38 (2015).
96. Feschchotte, C. & Pritham, E. J. Non-mammalian c-integrases are encoded by giant transposable elements. *Trends Genet.* **21,** 551–552 (2005).
97. Krupovic, M. & Koonin, E. V. Evolution of eukaryotic single-stranded DNA viruses of the *Bidnaviridae* family from genes of four other groups of widely different viruses. *Sci. Rep.* **4,** 5347 (2014).
98. Roux, S. *et al.* Chimeric viruses blur the borders between the major groups of eukaryotic single-stranded DNA viruses. *Nat. Commun.* **4,** 2700 (2013).
99. Matsuda, M. *et al.* Construction of a BAC library derived from the inbred Hd-rR strain of the teleost fish, *Oryzias latipes*. *Genes Genet. Syst.* **76,** 61–3 (2001).
100. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics* **30,** 2114–2120 (2014).
101. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–2158 (2011).
102. McWilliam, H. *et al.* Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Res.* **41,** 597–600 (2013).
103. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25,** 1972–1973 (2009).

104. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30,** 772–780 (2013).

105. Mii, Y. & Taira, M. Secreted Frizzled-related proteins enhance the diffusion of Wnt ligands and expand their signalling range. *Development* **4088,** 4083–4088 (2009).

106. Hirt, B. Selective extraction of polyoma DNA from infected mouse cell cultures. *J. Mol. Biol.* **26,** 365–369 (1967).

107. Pei, J., Kim, B. H. & Grishin, N. V. PROMALS3D: A tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.* **36,** 2295–2300 (2008).

108. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).

109. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).

110. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}; *iso*-2; *iso*-3. *Fly (Austin).* **6,** 80–92 (2012).

111. Betancur, R. R. *et al.* The Tree of Life and a New Classification of Bony Fishes. *PLOS Curr. Tree Life* 1–54 (2013). doi:10.1371/currents.tol.53ba26640df0ccaee75bb165c8c26288.

112. Okonechnikov, K. *et al.* Unipro UGENE: A unified bioinformatics toolkit. *Bioinformatics* **28,** 1166–1167 (2012).

113. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).

114. Near, T. J. *et al.* Resolution of ray-finned fish phylogeny and timing of diversification. *Proc. Natl. Acad. Sci.* **109,** 13698–13703 (2012).

Acknowledgements

First of all, I would like to explain my deepest gratitude for my supervisor, Professor Hiroyuki Takeda, for providing me with the opportunity to study in a splendid environment. I owe my completion of my Ph.D. course to his substantial supports, encouragement, and patients.

I would like to express my gratitude to Dr. Akihiko Koga for helpful discussion and experimental suggestions, especially about transposition assay. Without his help, I would not have achieved to obtain Ph.D.

I would also like to thank Dr. Shinichi Morishita and Dr. Wei Qu for sequencing BAC clones and providing the newly assembled medaka genome data, Dr. Kiyoshi Naruse, Dr. Tetsuaki Kimura and Dr. Yusuke Takehana for helping me with BAC screening and providing medaka related species fish, Dr. Yasushi Kawaguchi, Dr. Akihisa Kato and Dr. Jun Arii for the help of experiments to examine reactivation of *Teratorn*.

I am truly grateful to all members of Takeda laboratory for all they have done for my life in the laboratory. Particularly, I'd like to express my appreciation to Dr. Atsuko Shimada for deep discussion and helpful feedbacks, Dr. Masahiko Kumagai for the generous help about computational analyses and evolutionary studies, Mr. Tomonori Saga and Mr. Takumi Aikawa for the help of my study and fruitful discussion.

Finally, I am indebt to my family for their generous affection, heartfelt

support and encouragement all the time. I dedicate my doctoral thesis to them.