Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

An extraordinary retrotransposon family encoding dual endonucleases

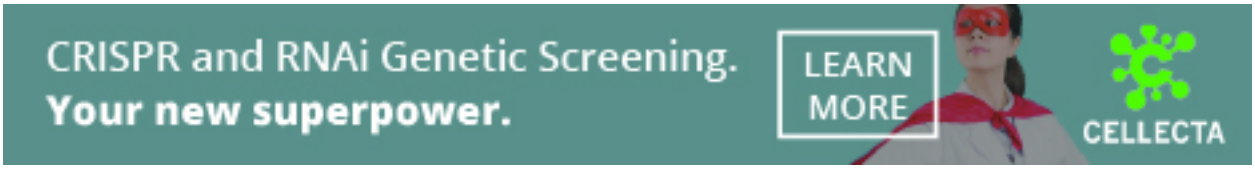
Kenji K. Kojima and Haruhiko Fujiwara

Genome Res. 2005 15: 1106-1117
Access the most recent version at doi:10.1101/gr.3271405

References This article cites 32 articles, 11 of which can be accessed free at:
http://genome.cshlp.org/content/15/8/1106.full.html#ref-list-1

License

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**



To subscribe to *Genome Research* go to:
https://genome.cshlp.org/subscriptions

Cold Spring Harbor Laboratory Press

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

Letter

An extraordinary retrotransposon family encoding dual endonucleases

Kenji K. Kojima and Haruhiko Fujiwara[1]

Department of Integrated Biosciences, Graduate School of Frontier Sciences, University of Tokyo, Kashiwa, Japan

Retrotransposons commonly encode a reverse transcriptase (RT), but other functional domains are variable. The acquisition of new domains is the dominant evolutionary force that brings structural variety to retrotransposons. Non-long-terminal-repeat (non-LTR) retrotransposons are classified into two groups by their structure. Early branched non-LTR retrotransposons encode a restriction-like endonuclease (RLE), and recently branched non-LTR retrotransposons encode an apurinic/apyrimidinic endonuclease-like endonuclease (APE). In this study, we report a novel non-LTR retrotransposon family Dualen, identified from the *Chlamydomonas reinhardtii* genome. Dualen encodes two endonucleases, RLE and APE, with RT, ribonuclease H, and cysteine protease. Phylogenetic analyses of the RT domains revealed that Dualen is positioned at the midpoint between the early-branched and the recently branched groups. In the APE tree, Dualen was branched earlier than the I group and the Jockey group. The ribonuclease H domains among the Dualen family and other non-LTR retrotransposons are monophyletic. Phylogenies of three domains revealed the monophyly of the Dualen family members. The domain structure and the phylogeny of each domain imply that Dualen is a retrotransposon conserving the domain structure just after the acquisition of APE. From these observations, we discuss the evolution of domain structure of non-LTR retrotransposons.

[Supplemental material is available online at www.genome.org and http://www.biol.s.u-tokyo.ac.jp/users/animal/kojima/sequence.html. The following individuals and institute kindly provided reagents, samples, or unpublished information as indicated in the paper: M. Hirono, M. Kurosawa, K. Sonoike, and the US Department of Energy Joint Genome Institute.]

Retrotransposons are mobile genetic elements found in a wide range of eukaryotes (Arkhipova and Meselson 2000). Retrotransposons have a reverse transcriptase (RT) in common, but other functional domains are quite variable. Retroviruses are considered to be retrotransposons that have acquired a domain for extracellular function (Malik et al. 2000). The acquisition of new domains has provided variable life styles to retroelements. Non-long-terminal-repeat (non-LTR) retrotransposons are one major group of retrotransposons and are considered to be the ancestors of long-terminal-repeat (LTR) retrotransposons and retroviruses (Malik and Eickbush 2001). Non-LTR retrotransposons are classified into two groups by their structure (Malik et al. 1999; Yang et al. 1999). The early branched non-LTR retrotransposons, such as the insect R2, include only one open-reading-frame (ORF) encoding an RT and a restriction-like endonuclease (RLE). In contrast, the recently branched non-LTR retrotransposons, such as the human L1 (long interspersed nuclear element-1, LINE-1), include two ORFs, and the second ORF encodes an RT and an apurinic/apyrimidinic endonuclease-like endonuclease (APE). Several families of the recently branched non-LTR retrotransposons have only the second ORF. Several recently branched non-LTR retrotransposons, such as the I and the TRAS families, have a ribonuclease H (RNH) domain immediately after the RT domain, similar to LTR retrotransposons. The RT domain is the only common structure among all non-LTR retrotransposons (see Fig. 7B, below).

[1]Corresponding author.
E-mail haruh@k.u-tokyo.ac.jp; fax 81-4-7136-3659.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.3271405.

The existence of the endonuclease domain is the most remarkable feature of non-LTR retrotransposons when compared with LTR retrotransposons, which use integrases, but not endonucleases for genome integration. Endonuclease defines a transposition mechanism peculiar to non-LTR retrotransposons that is called target-primed reverse transcription (TPRT). In the nucleus, endonuclease nicks the target DNA, and the free 3′-hydroxyl end of DNA is used as primer for reverse transcription. This contrasts with LTR retrotransposons and retroviruses that use tRNA as primer and are reverse transcribed in the cytoplasm. LTR retrotransposons and retroviruses import their cDNA into the nucleus, and integrate it into the genomic DNA by integrases. Both non-LTR retrotransposons with RLE, and those with APE, were shown to transpose by TPRT (Yang et al. 1999; Cost et al. 2002). Inactivation of endonuclease dramatically reduces the transposition efficiency of non-LTR retrotransposons (Feng et al. 1996; Takahashi and Fujiwara 2002). Because all early branched non-LTR retrotransposons have an RLE and all recently branched retrotransposons encode an APE, it is certain that non-LTR retrotransposons once exchanged their endonuclease type from RLE to APE. Until now, however, we did not have any evidence for this evolutionary event.

In this study, we report a novel non-LTR retrotransposon family that encodes both RLE and APE. These elements, which we named Dualen, are positioned phylogenetically at the midpoint between the early branched group and the recently branched group. In addition, Dualen also encodes an RNH domain after the RT domain. We discuss the origin and the evolutionary implication of the extraordinary domain structure of Dualen.

15:1106–1117 ©2005 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/05; www.genome.org

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

Results and Discussion

Dualen, a new family of non-LTR retrotransposons, has dual endonuclease domains and ribonuclease H

While we screened early branched non-LTR retrotransposons from genomic databases (Kojima and Fujiwara 2004), we identified a novel non-LTR retrotransposon which was apparently distinct from other non-LTR elements, in the *Chlamydomonas reinhardtii* genomic database at the US Department of Energy Joint Genome Institute (JGI, http://aluminum.jgi-psf.org/prod/bin/runBlast.pl?db=chlre1/). This novel retrotransposon (DualenCr1) (Fig. 1A) encodes an ORF longer than 9 kb. BLAST analysis of the protein sequence of DualenCr1 to conserved domain database (CDD) (Marchler-Bauer et al. 2003) at NCBI revealed that it includes four domains, namely, josephin (*E*-value 0.002), apurinic/apyrimidinic endonuclease-like endonuclease (APE, *E*-value 2e-13), reverse transcriptase (RT, *E*-value 6e-14), and ribonuclease H (RNH, *E*-value 4e-5). Since the similarity between josephin domains and DualenCr1 is weak, we further analyzed the significance of this similarity (discussed later). We manually characterized an additional two domains by alignment with other non-LTR retrotransposons, CCHC zinc-finger motif (ZF) and restriction-like endonuclease (RLE) (Fig. 1A). We named this non-LTR retrotransposon family Dualen, because it encodes dual endonucleases, APE and RLE, each of which is specific to the early branched and the recently branched non-LTR retrotransposons, respectively. RT and ZF were reported in both the early branched and the recently branched non-LTR retrotransposons. RNH was shown in only some recently branched non-LTR retrotransposons (see Fig. 7B, below).

It was shown that Dualen constituted a family including several elements that have the same protein domain composition, but their nucleotide sequences were less conserved. From the *C. reinhardtii* genomic database, we identified three complete Dualen elements (DualenCr1, DualenCr3, DualenCr4) and one related element (DualenCr2), which has 5′-truncation because of partial sequencing and/or incomplete retrotransposition (Fig. 1A). We also found two Dualen elements (DualenU1 and DualenU2) from an unassembled *Arabidopsis thaliana* HTGS (high-throughput genomic sequence) (accession no. AC109923). However, there were no corresponding sequences for the two Dualen elements in the *A. thaliana* complete genome sequence database.

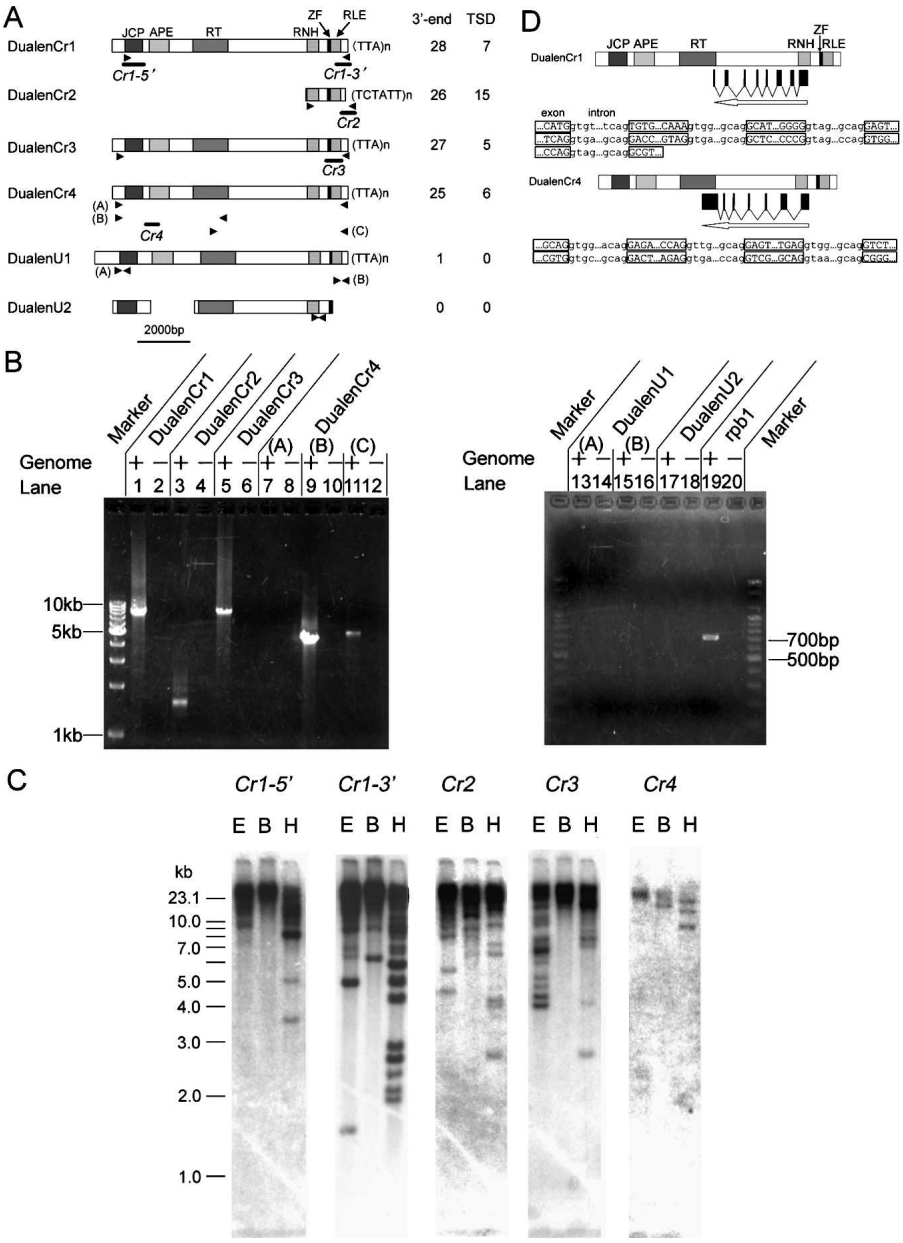


Figure 1. Characterization of Dualen. (*A*) Structure of six Dualen elements. Motifs and domains are schematically shown as boxes. The numbers of identified 3′-ends and target site duplications (TSDs) are listed at *right*. Primers used for PCR (*B*) are shown as arrowheads, and regions used for probes in Southern hybridization (*C*) are shown as bold lines *below* the structure. (JCP) Josephin-related cysteine protease; (APE) apurinic/apyrimidinic endonuclease-like endonuclease; (RT) reverse transcriptase; (RNH) ribonuclease H; (ZF) zinc-finger motif; (RLE) restriction-like endonuclease. (*B*) Confirmation of the existence of Dualen elements by PCR. Primers used for PCR and expected sizes of PCR products are as follows: (lanes *1*,*2*) DualenCr1_F2 and DualenCr1_R1, 8740 bp; (lanes *3*,*4*) DualenCr2_F2 and DualenCr2_R1, 1659 bp; (lanes *5*,*6*) DualenCr3longF1 and DualenCr3longR1, 9233 bp; (lanes *7*,*8*) DualenCr4longF1 and DualenCr4longR1, 10,075 bp; (lanes *9*,*10*) DualenCr4longF1 and DualenCr4longR3, 4940 bp; (lanes *11*,*12*) DualenCr4longF3 and DualenCr4longR1, 5701 bp; (lanes *13*,*14*) DualenAt1_F1 and DualenAt1_R1, 610 bp; (lanes *15*,*16*) DualenAt1_F2 and DualenAt2_R2, 462 bp; (lanes *17*,*18*) DualenAt2_F and DualenAt2_R, 467 bp; (lanes *19*,*20*) Atrpb1F and Atrpb1R, 707 bp. Primers used for PCR are listed in Supplemental Table 1. (*C*) Southern hybridization. Probes are shown in *A*. (E) EcoRI; (B) BglII; (H) HindIII. (*D*) Diagrams of spliced antisense RNAs. Exon–intron boundaries of antisense RNAs are shown *below*. Exon sequences are in uppercase and boxed, while intron sequences are in lowercase.

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

AC109923 was derived from a BAC clone of the strain Colombia. This strain was also used for the genome sequencing project (*Arabidopsis* Genome Initiative 2000).

To investigate whether the Dualen elements truly exist in the genome of *C. reinhardtii* and *A. thaliana*, and to exclude the possibility of assembly error, we tried to detect genomic copies of the Dualen elements by polymerase chain reaction (PCR) (Fig. 1B). We designed primer pairs to amplify almost full-length Dualen elements (Fig. 1A, arrowheads). The PCR bands of DualenCr1, DualenCr2, and DualenCr3 were detected with the expected sizes (Fig. 1B, lanes 1, 3, and 5), suggesting that these Dualen elements exist in the *C. reinhardtii* genome. As we could identify only a fragmental DualenCr2 sequence from the genomic database and by PCR, it is uncertain whether complete DualenCr2 copies exist in the *C. reinhardtii* genome.

We could not amplify the PCR product corresponding to the full-length DualenCr4 (Fig. 1B, lane 7). We reconstructed the full-length DualenCr4 using two scaffold sequences, scaffold 255 and scaffold 581. Scaffold 581 corresponds to 1–5382 bp and scaffold 255 corresponds to 2596–10,346 bp of the reconstructed DualenCr4. However, scaffold 581 contains only a 5′-half of DualenCr4. The Dualen copy in scaffold 581 is 3′-truncated. There are no other 5′-half sequences of DualenCr4 in the *C. reinhardtii* genomic sequence database at present (the draft release version 2.0 of the *Chlamydomonas reinhardtii* genome at JGI). We designed two sets of primers to amplify two overlapping regions, the region 190–5129 bp (Fig. 1A, DualenCr4 [B]) and the region 4564–10,264 bp (Fig. 1A, DualenCr4 [C]). We could amplify these two overlapping PCR products at the expected sizes (Fig. 1B, lanes 9 and 11). These data indicate that there are 5′-truncated and 3′-truncated copies of DualenCr4 in the *C. reinhardtii* genome, but there is no full-length DualenCr4 copy.

Table 1. Target site duplications (TSDs) of the Dualen elements

5′ TSD	Element (5′–3′)[a]	3′ TSD[b]
DualenCr1 7 TSDs		
ATTGCGGGTGTC	TTCAT...GTAGAtattattattattatt	attGCGGGTGTC
TGTCAACCATGTC	CCCAG...GTAGAtattattattattttt	TGTCAACCATGTC
GGGGTG	AGCTT...GTAGAtattatttcatca	GGGGTG
TGTAGCAAGT	TATGG...GTAGAtattattattattattat	tGTAGCAAGT
CCGTGGAGCTGCC	TGGTG...GTAGAtattattattattatt	CCGTGGAGCTGCC
CTACCGTCCTCCTCG	ATGGT...GTAGAtattattattatta	CTACCGTCCTCCTCG
TTCTCCA	TCCGT...GTAGAtattattattatta	ttCTCCA
DualenCr2 15 TSDs		
CATCTGGATGAC	TCTAC...GTGTCtctatttctatttctattt	cATCTGGATGAC
TTCACTCTGCCTGG	TCGTG...GTGTCtctatttctatttctat	ttcACTCTGCCTGG
CAGCCTGCCTGT	ATTGC...GTGTCtctatttctatttctattt	cAGCCTGCCTGT
TACAGCTAGAT	TCACA...GTGTCtctatttctatttctatttc	taCAGCTAGAT
CGTGGTGAGGT	AGTGG...GTGTAtctatttctatttctattt	cGTGGTGAGGT
TCACTCGCCTGGT	CGGTA...GTGTAtctatttctatttctatt	tcACTCGCCTGGT
CCCGGTT	GCTCG...GTGTAtctatctctatttctattctt	CCCGGTT
CAAACAATATCAC	GCTCG...GTGTAtctatttctatttctatttct	CAAACAATATCAC
TTCCAGATATTC	GGTAG...GTGTGtctatttctatttctgt	ttcCAGATATTC
TGCGCGCCACCG	GTAAT...GTGTCtctatttctatttctattct	TGCGCGCCACCG
CTAAACTTGTGGAT	GCTCG...GTGTAtctatttctatttctattt	ctaAACTTGTGGAT
TTTGGTTGGGCT	GAGTG...GTGTAtctatttctatttctatgt	TTTGGTTGGGCT
TTGCTCTTTGC	CTCGG...GTGTAtctgttttctatttcctatt	tTGCTCTTGC
TACCGTGTG	CCTGG...GTGTCtctatttctatttctatttc	taCCGTGTG
CGCCCTGCTGACCACTTGCCT	GCCTG...GTGTAtctatttctattt	cGCCCTGCTGACCACTTGCC
GCCACTCCACCGTGCTG		TGCCACTCCACCGTGCTG
DualenCr3 5 TSDs		
TTGCCACCGTTGTAG	ACATT...TTTGAattattattatta	ttGCCACCGTTGTAG
TGCAACTGCG	GTTAC...TTTGAattattattattatta	tGCAACTGCG
ATCAGCAACAGCAGCA	CTGCC...TTTGAattattattattatt	atCAGCAACAGCAGCA
TTCTTGCGGTCAGG	ATCCT...TTTGAattattattatta	ttCTTGCGGTCAGG
CCCTATGAAACACCA	CAGCT...TTTGAattattattatcat	CCCTATGAAACACCA
DualenCr4 6 TSDs		
CCATCTACAAGATACCGGTAC	CGAAG...ACC-Attattattattatta	CCATCTACAAGATACCGGTA
GGTACTGCACCGGGTGCCGTC		CGGTACTGCACCGGGTGCCG
CGG		TCCGG
TTGCCCAGGACC	AACCG...AAC-Attaatatta	ttGCCCAGGACC
CTCCTGCTCCTGCT	CAGCG...ACC-Attatcattattattattattat	CTCCTGCTCCTGCT
TGCTGGACC	ATGCG...ACCTAttattattattat	tGCTGGACC
TCACAGCCCCAA	TTTGG...ACCTAttattattatca	tCACAGCCCCAA
TGACGCACAGGT	GCAGG...ACCTAttattattattattta	TGACGCACAGGT

[a]3′-terminal repeats are shown in lowercase.
[b]Nucleotides that can be interpreted as either 3′-terminal repeats or TSDs are in lowercase.

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

We could not detect either the PCR products for the two proposed Dualen elements (DualenU1 and DualenU2) in the *A. thaliana* genome (Fig. 1B, lanes 13, 15, and 17). The PCR product for the RNA polymerase II (*rpb1*) gene, which is a single-copy gene in the *A. thaliana* genome and used as positive control, could be amplified (Fig. 1B, lane 19). The most probable explanation is that the two Dualen elements found in AC109923 originated from a contaminating organism, not from the *A. thaliana* genome. We also searched other genomic databases, including arthropods, chordates, plants, fungi, and protozoan parasites, but we could not find any other Dualen elements. At present, there are very little genomic sequences other than *C. reinhardtii* in green algae. Thus, it is possible that other green algae, such as *Volvox* and *Chlorella*, have Dualen elements on the genome. The six Dualen elements identified here hold the same domain structure in common (Fig. 1A). The similarity throughout ORF among Dualen elements indicates that they were branched from the common ancestral retrotransposon that had the same structure as Dualen, and that the structure of Dualen is a functional unit for retrotransposition, not a "nonfunctional" chimera of two non-LTR retrotransposons.

Genomic structures of Dualen indicate the recent activity for retrotransposition

We performed Southern hybridization for further characterization of Dualen in the *C. reinhardtii* genome (Fig. 1C). We could observe at least three separate bands below a mass at high molecular weight when *Cr1-5′* was used as a probe, 14 bands with *Cr1-3′*, seven bands with *Cr2*, 12 bands with *Cr3*, and three bands with *Cr4*. Similar intensity of several bands indicates that each of them represents a single copy. We calculated the copy number of each Dualen element using densitometry. The whole-copy number of DualenCr1, which was indicated by Southern hybridization with *Cr1-3′*, was calculated at about 50, and the full-length DualenCr1 was calculated at five to 10 copies. Since non-LTR retrotransposons are frequently truncated at the 5′-terminus because of incomplete reverse transcription, the copy number of the 3′-region is higher than that of the 5′-region. Actually, there were more bands when the 3′-region of DualenCr1 (*Cr1-3′*) was used as a probe than when the 5′-region of DualenCr1 (*Cr1-5′*) was used as a probe. The whole-copy numbers of DualenCr2 and DualenCr3 were presumed ~60 (Fig. 1C). The copy number of the 5′-region of DualenCr4 was three, all of which were visible in Southern hybridization (Fig. 1C). We concluded that these bands represent 3′-truncated copies, combining with the results of long range PCR (Fig. 1B).

To obtain further genomic information of Dualen, such as copy number, target sequence preference, and the length of target-site duplication (TSD), we searched Dualen copies from the genomic database. We identified 28 copies of DualenCr1, 26 of DualenCr2, 27 of DualenCr3, and 25 of DualenCr4 from the *C. reinhardtii* genomic database (Fig. 1A), all of which integrated into different sequences (data not shown). These copy numbers are consistent with the results of Southern hybridization. We also identified several target-site duplications (TSDs) of all four Dualen elements in *C. reinhardtii* (Table 1). The existence of TSDs shows that dual endonucleases function for nicking the target sites, since endonuclease-independent transposition events do not make TSDs (Morrish et al. 2002). Because ancient copies of retrotransposons and TSDs are subject to accumulation of mutations, which makes it difficult to distinguish TSDs, the existence of recognizable TSDs at both sides of retrotransposons supports the recent retrotransposition activity. DualenCr4 seems to be inactive now as described above, but the existence of several TSDs without any mutations indicates that DualenCr4 was active recently. We observed frequent 5′-truncations, variable length of 3′-terminal repeats (TTA or TCTATT; mainly 10–25 bp), and relatively long TSDs (mainly 10–15 bp), all of which are common features of non-LTR retrotransposons, suggesting that Dualen transposes according to the TPRT mechanism like other non-LTR retrotransposons. There is neither clear sequence similarity nor consensus sequence upstream from TSDs among target sequences (Table 1), which indicates that Dualen has no sequence specificity. Both the early branched and the recently branched groups contain sequence-specific and nonsequence-specific non-LTR retrotransposons (Kojima and Fujiwara 2003, 2004). In conclusion, we could not judge the activity of RLE and APE based on the target sequence.

Antisense RNA of Dualen is transcribed and spliced

We next investigated the transcription of Dualen using the EST (expressed sequence tags) database. BLAST search to EST databases at NCBI revealed the transcription of all Dualen elements in *C. reinhardtii*. We identified 22 (DualenCr1), two (Cr2), six (Cr3), and 12 (Cr4) EST clones that showed >90% nucleotide identity throughout the sequences (Table 2). To our surprise, we found spliced antisense transcripts of DualenCr1 and DualenCr4 in EST sequences (Table 2; Fig. 1D). Spliced antisense transcripts of non-LTR retrotransposons have been reported only in Tad, a non-LTR retrotransposon of *Neurospora* (Sewell and Kinsey 1996). We characterized seven introns in antisense DualenCr1 and six introns in antisense DualenCr4 (Fig. 1D). The spliced antisense transcripts of DualenCr1 and DualenCr4 have no sequence similarity to each other. They contain no long ORFs. The function of spliced antisense transcripts is unknown, but it is possible that they serve a regulatory role.

Dualen is an intermediate retrotransposon between RLE-encoding and APE-encoding retrotransposon groups

Because Dualen has both an RLE domain that is considered to be specific for the early branched non-LTR retrotransposons and an APE domain that is considered to be specific for the recently branched retrotransposons, we analyzed the phylogenetic position of Dualen. In the Bayesian phylogenetic inference based on the RT domains, the Dualen family is a monophyletic group positioned at the midpoint between the early branched and the recently branched non-LTR retrotransposons (Fig. 2A). Malik et al.

Table 2. EST clones of Dualen

Element	Accession number
DualenCr1	BI993891, BU646135, BM002694, BU645902, BU647193, BI728709, BG854433, BI875335, BM002695, BI728708, BI875545, BU647852, BM002198, BU647192, AV628877, BM002199, AV627144
	(spliced antisense transcripts) BI724416, AV629322, BE337758, AV641246, AV641941
DualenCr2	BE441359, BU650121
DualenCr3	BU654516, BU654235, BE352316, BU654236, BE352323, BU647532
DualenCr4	BM003214, BI994690, BI995499, BG860704
	(spliced antisense transcripts) BM003215, BI997140, BI720006, BI720007, BG860705, BI718488, BG856815, AV635046

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

(1999) proposed the use of the term "clade" to represent non-LTR retrotransposons that (1) share the same structural features, (2) are grouped together with ample phylogenetic support, and (3) date back to the Precambrian era. The Dualen family satisfies these three points; thus, we propose the Dualen clade including only the Dualen family. Eickbush and Malik (2002) also proposed an additional classification scheme "group," in which the various clades are grouped on the basis of both the phylogenetic relationship and the nature and arrangement of their protein domains. In their classification, the Dualen family also composes a distinct group.

Figure 2B shows the 50% consensus phylogenetic trees of 15 clades. It was reported that bootstrap proportions and Bayesian posterior probabilities cannot be directly compared (Douady et al. 2003). The topologies of the Bayesian inference tree and the Neighbor-Joining (NJ) tree are basically consistent. We previously reported that the HERO family belongs to the NeSL clade (Kojima and Fujiwara 2004), and Bouneau et al. (2003) (in their paper, HERO was named Zebulon) also reached the same inference, but the HERO family is positioned at internal of the R2 clade in the Bayesian tree. The clade branching just before the Dualen clade is the R4 clade, and the clade branching just after the Dualen clade is the L1 clade. The R4 clade elements have an RT and an RLE, and the L1 clade elements have an APE and an RT (see Fig. 7B, below). Thus, the domain structure of Dualen is consistent with the phylogenetic position.

Malik et al. (1999) showed that the "thumb" region of the RT domain appears to be divided into two subtypes, CRE/R2/R4/L1/RTE (corresponding to the R2, the L1, and the RTE groups) and Tad/R1/LOA/Jockey/CR1/I (the I and the Jockey groups). However, the monophyly among the RTE and the latter group was highly supported in their study and in our analysis (98% in the Bayesian tree and 84% in the NJ tree; Fig. 2B). The RT phylogeny clearly shows that the Dualen group is grouped with the R2 and the L1 groups, neither with the RTE group nor with the I and the Jockey groups. Thus, we propose that the RT domains of non-LTR retrotransposons are classified into three subtypes, the eldest subtype (the R2, the Dualen, and the L1 groups), the middle-aged sub-

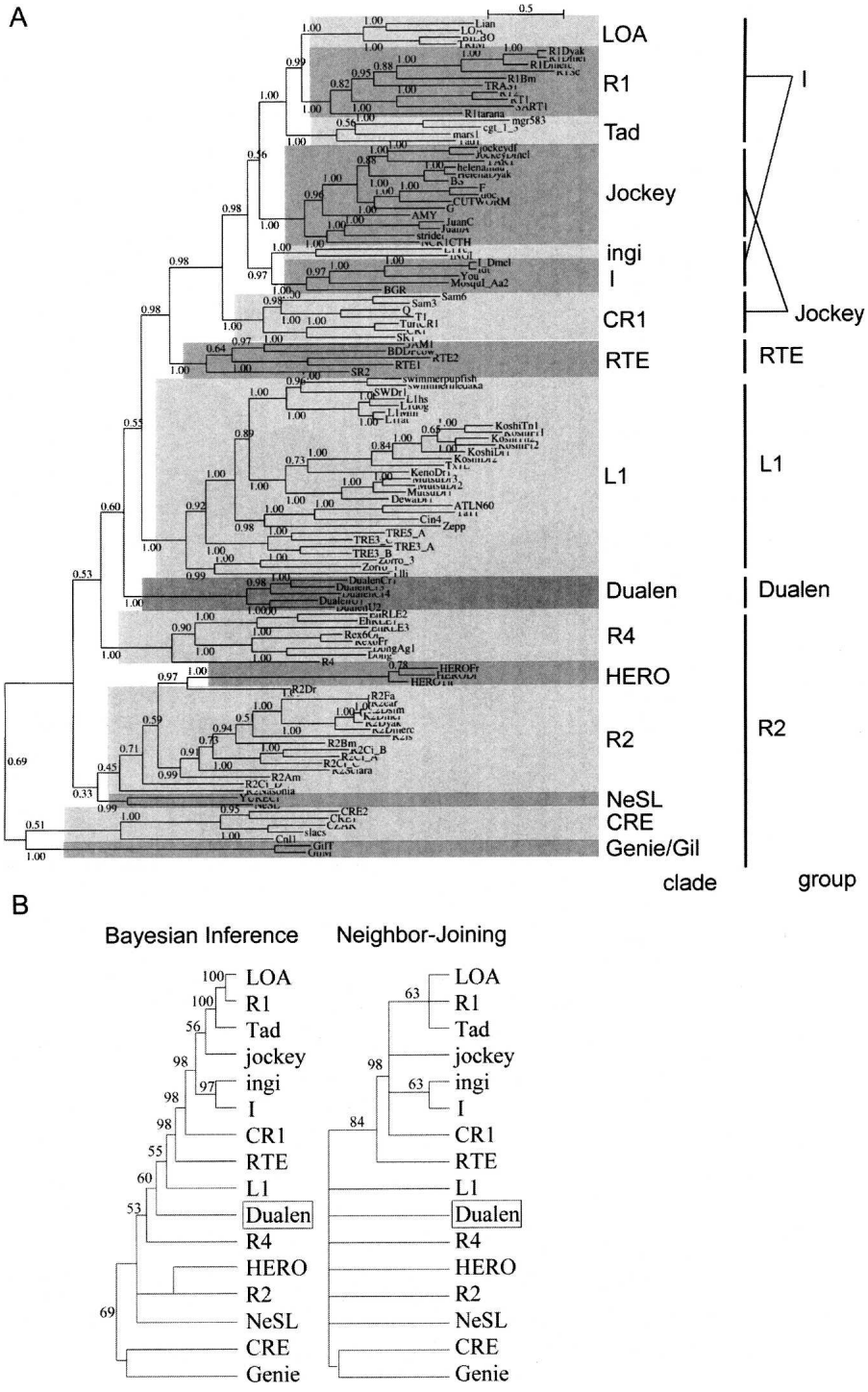


Figure 2. Phylogenetic analyses of reverse transcriptase (RT). (*A*) The Bayesian phylogenetic inference. Posterior probabilities are indicated. Markov chain Monte Carlo (MCMC) chain length was 500,000 generations with trees sampled every 10 generations; the first 3300 trees were discarded as burn-in. (*B*) The 50% consensus trees of the Bayesian inference and the Neighbor-Joining methods. Retrotransposons belonging to the same clade are compressed. Posterior probabilities (Bayesian inference) and bootstrap values (Neighbor-Joining) are indicated as a percentage. HERO is not recognized as clade, but treated independently here due to its indefinite position. The sources of sequences in the L1, R4, HERO, R2, NeSL, CRE, and Genie/Gil clades are given in our previous study (Kojima and Fujiwara 2004). Others are given in the report by Malik et al. (1999). The alignment used to generate this tree is based on two previous alignments, ds36752 (Malik et al. 1999) and ALIGN 00231 (Burke et al. 2002), and available as Supplemental Figure 1.

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

type (the RTE group), and the youngest subtype (the I and the Jockey groups).

Conservation of the catalytic residues in dual endonucleases (RLE and APE) suggests that both endonucleases are functional in retrotransposition

The most extraordinary feature of Dualen is its dual endonuclease domains. We identified the complete RLE of five elements except DualenU2 (Fig. 3). The K/R-P-D-X_{12-19}-D/E motif is conserved among all early branched non-LTR retrotransposons as well as among the type IIS restriction enzymes such as FokI (Yang et al. 1999). The P-D and D/E residues are essential for the endonuclease activity of the type IIS restriction enzymes (Waugh and Sauer 1993), and the former D residue was also reported to be essential for endonuclease activity of two early branched non-LTR retrotransposons, R2Bm and EhLINE1 (Yang et al. 1999; Mandal et al. 2004). These essential residues for cleavage activity of RLE are also conserved among the Dualen elements (Fig. 3, indicated by asterisks). Among the early branched non-LTR retrotransposons, two additional motifs are conserved in the RLE region. One is an upstream K/R-H-D/N motif and the other is a downstream K/R-X_2-K/R-Y motif. These motifs are peculiar to RLEs of non-LTR retrotransposons. In Dualen, the downstream motif is changed to K-X_2-Q-H (Fig. 3). Q, K, and R have an amino group (-NH_2) or a guanidinium group (-NH-C$(NH_2)_2^+$) at the terminus of their side chains. H and Y are aromatic amino acids. Considering the similarity of the features of these amino acids, the substitutions at this motif would have minor effects. The conservation of catalytic residues and substitutions on some conserved residues indicate that the endonuclease activity of the RLE domain of Dualen could be weakened, but is not lost. There is another possibility that the substituted residues only reflect Dualen-specific mutations and do not change the endonuclease activity at all. Although biochemical analyses are necessary to characterize the catalytic activity of the RLE domain of Dualen correctly, it is certain that RLEs of Dualen and of early branched retrotransposons share the common ancestor.

With regard to APE, we identified the complete APE sequences in four elements (Fig. 4). All catalytic residues whose activity was confirmed by experiments (Feng et al. 1996) are not mutated in the Dualen elements (Fig. 4, indicated by asterisks). However, the highly conserved triplet S-D-H is substituted to F-D-H (DualenCr1, 3), T-D-H (DualenCr4), or L-D-H (DualenU1). Although serine is highly conserved among the endonuclease/exonuclease/phosphatase family (pfam03372.7), which includes apurinic/apyrimidinic endonucleases, DNase I, and inositol-1, 4, 5-trisphosphate phosphatases, there are a few exceptions, such as the APE domain of Zepp, an active non-LTR retrotransposon in *Chlorella*, in which the triplet S-D-H is substituted to G-D-H (Yamamoto et al. 2003). In CgT-1, a non-LTR retrotransposon in the fungal phytopathogen *Colletotrichum gloeosporioides*, S-D-H is substituted to A-D-H. Nonconserved residues are variable among the Dualen subfamilies. This indicates that there has been enough time to be mutated since Dualen acquired APE. If the APE domain of Dualen was inactive, other conserved residues would be mutated. Thus, the substitution from S-D-H to F-D-H, T-D-H, L-D-H would not eliminate the activity completely.

Although some substitutions at conserved residues are observed in both RLE and APE, the conservation throughout the domains indicates that both endonuclease activities are required for efficient retrotransposition. Both RLE and APE were reported to cleave the bottom (first, primer) strand (Feng et al. 1996; Yang et al. 1999), and the top (second, nonprimer) strand (Yang et al. 1999; Anzai et al. 2001). Sequence alignments of both endonucleases indicated that both endonuclease activities could be weakened. It is possible that dual endonucleases in Dualen compensate for their weakened activities to each other. Another possibility is that mutations of both endonucleases do not affect the

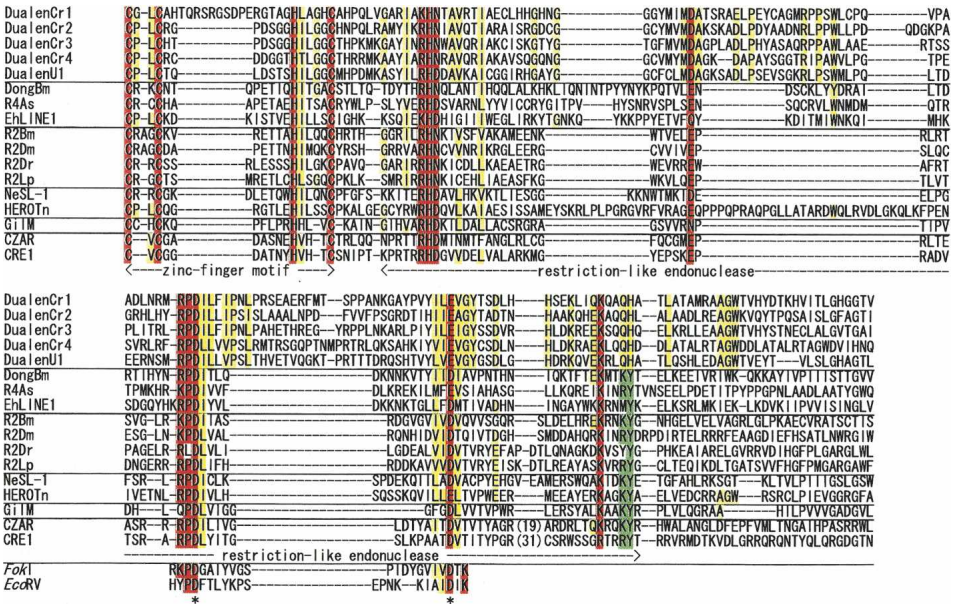


Figure 3. Sequence alignments of zinc-finger motif (ZF) and restriction-like endonuclease (RLE). Five Dualen elements and representatives of five clades of early branched non-LTR retrotransposons were aligned. Catalytic residues of restriction enzymes (FokI and EcoRV) are also shown. Residues conserved among all sequences are red, and residues conserved among Dualen elements are yellow. Residues conserved among other retrotransposons, but not among Dualen elements are green. Catalytic residues are indicated by asterisks.

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

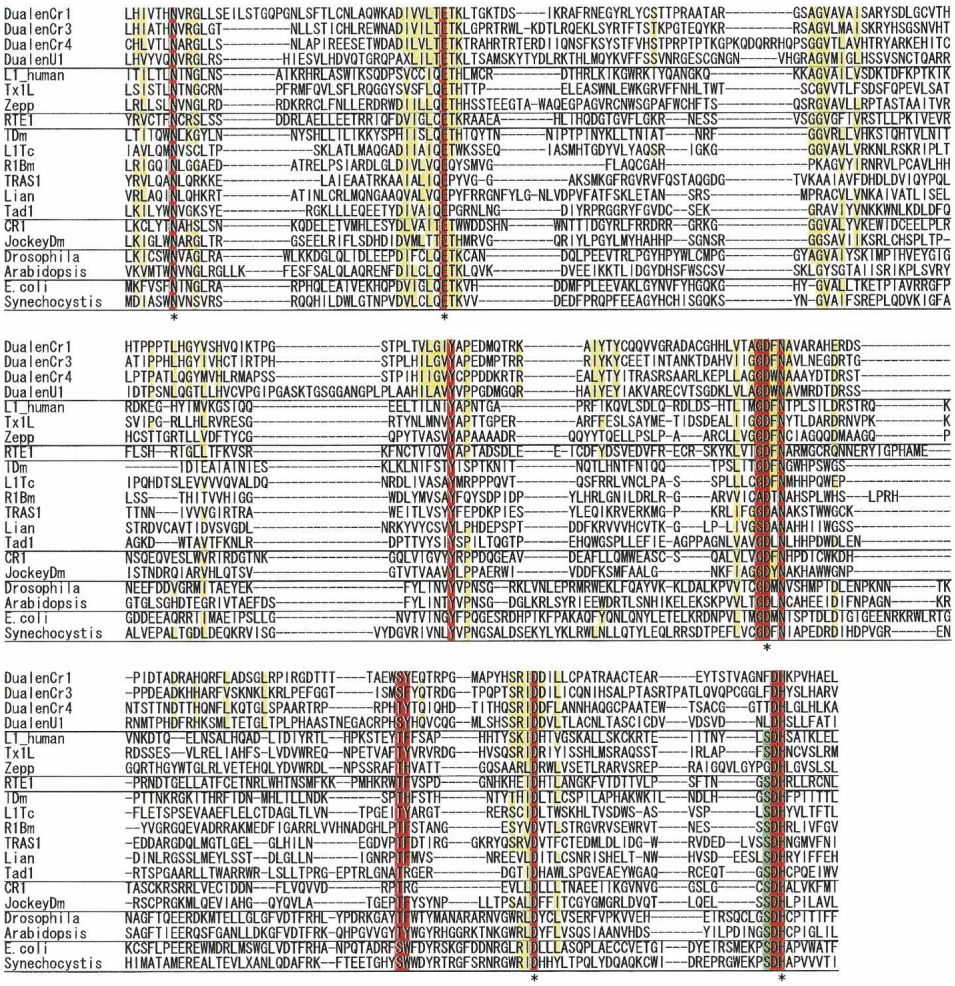


Figure 4. Sequence alignment of apurinic/apyrimidinic endonuclease-like endonuclease (APE). Four Dualen elements and representatives of four groups of recently branched non-LTR retrotransposons, eukaryotic apurinic/apyrimidinic endonucleases (*Drosophila melanogaster* and *Arabidopsis thaliana*), and prokaryotic exonuclease III (*Escherichia coli* K12 and *Synechocystis* sp. PCC 6803) were aligned. Residues conserved among all sequences are red, and residues conserved among Dualen elements are yellow. Residues conserved among other retrotransposons, but not among Dualen elements, are green. Catalytic residues are indicated by asterisks.

endonuclease activity in the least, and the persistence of two endonucleases is simply a result of selection for the ability to transpose into a wide range of sequences. Even if both endonucleases of Dualen had only a weak activity now, it is likely that the common ancestor of Dualen was more active due to dual endonucleases than retrotransposons that had a single endonuclease.

The Bayesian phylogenetic inference of the APE domains is shown in Figure 5A, and the 50% consensus trees of the Bayesian inference and the NJ methods are shown in Figure 5B. The APE trees have less resolution than the RT trees, because the APE domain is smaller and less conserved than the RT domain. The Dualen family is monophyletic like the RT phylogeny, but positioned internal of the recently branched non-LTR retrotransposons. The paraphyly of the L1 clade and that of the R1 clade are corresponding to the previous report (Malik et al. 1999), which is contradictory to the phylogeny of the RT domains. Since RLE is short and not well conserved, the RLE phylogeny brought no useful information (data not shown).

Ribonuclease H (RNH) of Dualen has the same origin as other non-LTR retrotransposons

The other functional domain conserved among several non-LTR retrotransposons and the Dualen family is RNH (Fig. 6A). The RNH domains are found only in the I group (including the I, the ingi, the R1, the LOA, and the Tad clades; see Figs. 2 and 7B), except the Dualen family. The RNH domain of Dualen contains five conserved catalytic D, E, D, H, D residues (Fig. 6A, asterisks) and is more similar to cellular RNH and the RNH of non-LTR retrotransposons than to those of LTR retrotransposons. The RNH domain of Dualen is located distant from the RT domain, which is quite different from other retrotransposons. The region between the RT and the RNH domains in Dualen is about 1000 residues, with no conserved functional domain in the interval. RNH is positioned immediately after the RT domain in all known RNH-containing non-LTR retrotransposons, LTR retrotransposons, and retroviruses (Malik and Eickbush 2001).

The Bayesian phylogenetic inference of the RNH domains is shown in Figure 5C, and the 50% consensus trees are shown in

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

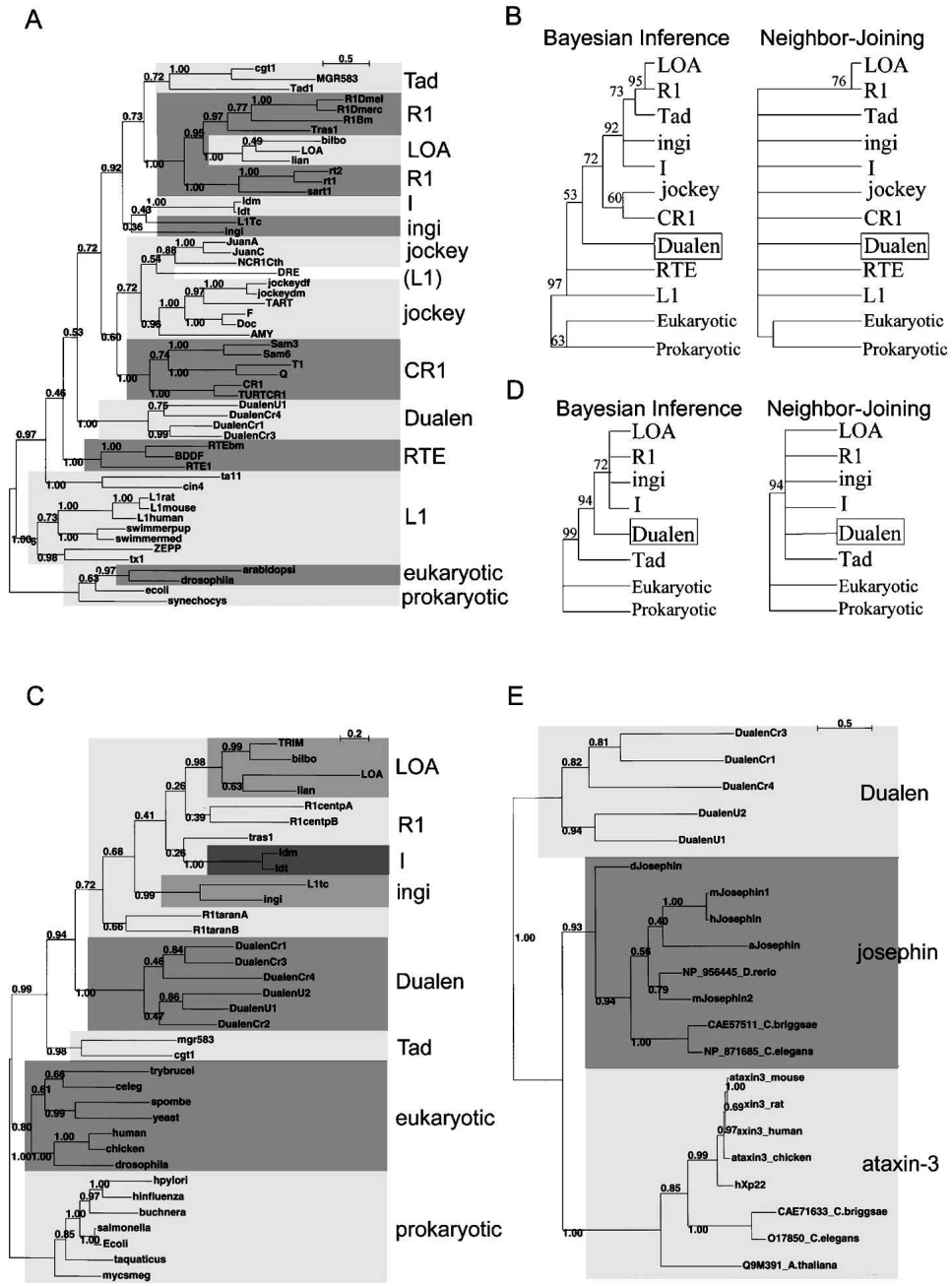


Figure 5. Phylogenetic analyses of apurinic/apyrimidinic endonuclease-like endonuclease (APE), ribonuclease H (RNH), and josephin-related cysteine protease (JCP). (*A*) The Bayesian phylogenetic inference tree of APE. Posterior probabilities are indicated. MCMC chain length was 150,000 generations with trees sampled every 10 generations; the first 1000 trees were discarded as burn-in. (*B*) The 50% consensus APE trees of the Bayesian inference and the Neighbor-Joining methods. Retrotransposons belonging to the same clade are compressed. Posterior probabilities (Bayesian inference) and bootstrap values (Neighbor-Joining) are indicated as a percentage. (*C*) The Bayesian phylogenetic inference of RNH. MCMC chain length was 100,000 generations with trees sampled every 10 generations; the first 700 trees were discarded as burn-in. (*D*) The 50% consensus RNH trees of the Bayesian inference and the Neighbor-Joining methods. (*E*) The Bayesian phylogenetic inference of JCP and josephin. MCMC chain length was 100,000 generations with trees sampled every 10 generations; the first 400 trees were discarded as burn-in.

Figure 5D. The Dualen family is also monophyletic in the RNH tree. The R1 clade is not monophyletic, and the Tad clade is the deepest branch among the I group, both of which are consistent with the previous report (Malik et al. 1999). Since the R1 clade is monophyletic in the RT tree, the paraphyly of the R1 clade is possibly due to the less conservation of RNH. The Tad clade is the sister clade of the R1 and the LOA clades in both the RT and the APE trees. The position of the Tad clade can be also due to the less conservation of RNH, although there is a possibility of domain swapping. Because the RNH phylogeny is not consistent with the RT phylogeny, we could not determine the position of RNH of the Dualen clade.

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

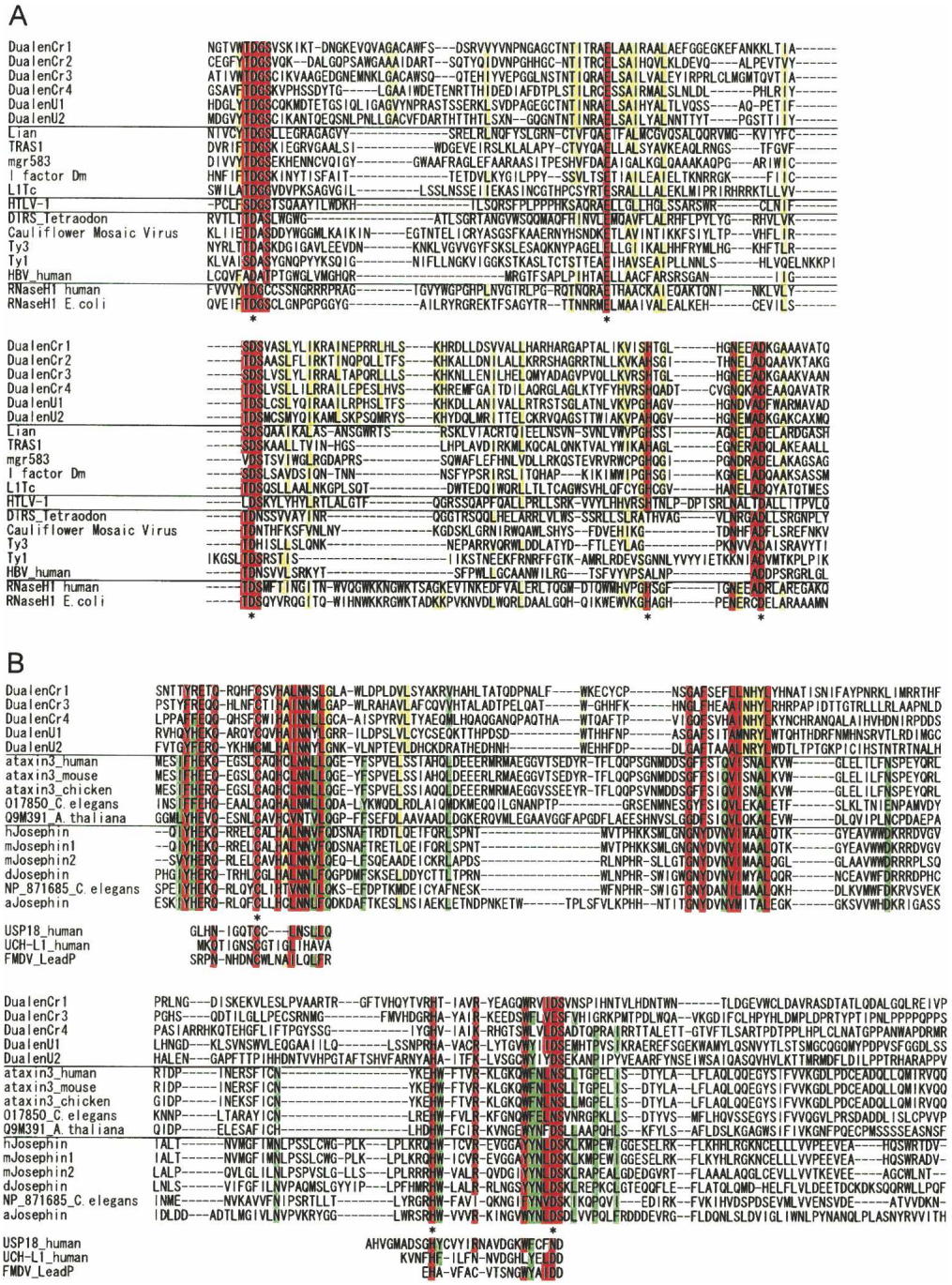


Figure 6. Sequence alignments of ribonuclease H (RNH) and josephin-related cysteine protease (JCP). Residues conserved among all sequences are red, and residues conserved among Dualen elements are yellow. Catalytic residues are indicated by asterisks. (*A*) Alignment of RNH. Six Dualen elements and representatives of five clades of recently branched non-LTR retrotransposons, six groups of LTR retroelements, and cellular RNH were aligned. (*B*) Alignment of JCP and josephin. Five Dualen elements and representatives of the ataxin-3 family and the josephin family were aligned. Catalytic residues of distant cysteine proteases (the human USP18, the human UCH-L1, and foot-and-mouth disease virus leader protease (FMDV_LeadP)) are also shown. Residues conserved among ataxin-3 and josephin, but not among JCP of Dualen elements are green.

The N-terminal josephin-like domain is a cysteine protease

The upstream region of the APE domain shows weak similarity to josephin domains, which was recently characterized in ataxin-3, also called Machado-Joseph disease gene, and its related proteins (Albrecht et al. 2003). Since the *E*-value of the BLAST search to conserved domain database (CDD) was marginal, we performed a PSI-BLAST search of the nonredundant database at NCBI seeded with the putative josephin domain of DualenU1. A total of 53 sequences belonging to the ataxin-3 family and the josephin family were found with statistical significance (*E*-values <E-5) within two iterations (data not shown). We aligned the presumed

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

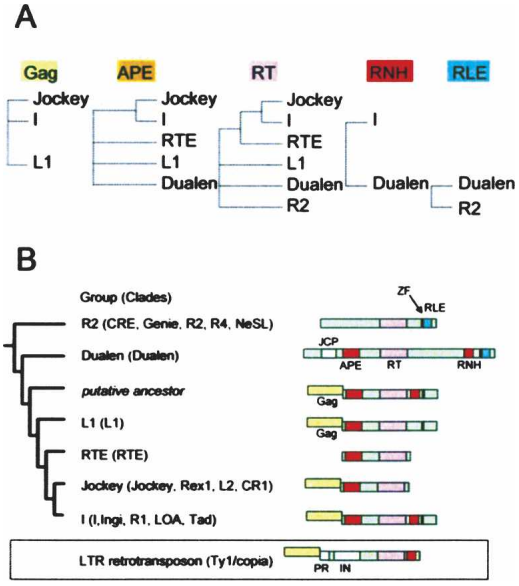


Figure 7. Summary of domain structure of non-LTR retrotransposons. (*A*) Reliable phylogeny of three conserved domains (APE, RT, and RNH) and the existence of two domains (Gag and RLE). (*B*) Evolution of domain structure of non-LTR retrotransposons. (PR) Protease; (IN) integrase.

josephin domains of Dualen with these two protein family members (Fig. 6B). Recent studies revealed that the josephin domain of ataxin-3 has highly conserved amino acids reminiscent of the catalytic residues of cysteine proteases known as deubiquitinating enzymes (Scheel et al. 2003), and this activity was confirmed experimentally (Burnett et al. 2003). Ubiquitin proteases function for editing and disassembly of polyubiquitin chains by removing ubiquitin from them. Two families, ubiquitin C-terminal hydrolases (UCHs) and ubiquitin-specific proteases (USPs), are distant relatives of josephin domains (Burnett et al. 2003; Scheel et al. 2003). The catalytic triads (C, H, D/N) conserved among three families (josephin, UCHs, and USPs) are observed in the putative josephin domains of Dualen, with the exception of DualenCr3, in which D/N is replaced to E (Fig. 6B, asterisks). These catalytic triads are also conserved among a wide variety of cysteine proteases as C, H, and a polar residue (data not shown) (Anantharaman and Aravind 2003). The putative josephin domains of Dualen are substituted at many sites that are conserved among the ataxin-3 family and the josephin family (Fig. 6B, green). However, the putative josephin domains of Dualen have residues conserved among wide cysteine proteases (Fig. 6B, red). These observations support that the putative josephin domain of Dualen is a cysteine protease related to josephin domains. Based on the above discussion, we named the putative josephin domain of Dualen, josephin-related cysteine protease (JCP) domain. Although JCP domains have not been identified in other non-LTR retrotransposons, another type of cysteine proteases was reported to exist in an early branched non-LTR retrotransposon NeSL-1 (Malik and Eickbush 2000). NeSL-1 encodes a cysteine protease similar to the yeast Ulp1 (ubiquitin-like protein-specific protease 1), which cleaves proteins from ubiquitin-like motifs, Smt3 or SUMO-1 (Li and Hochstrasser 1999). Two retrotransposons encode cysteine proteases related to ubiquitin proteasome, which is suggestive of the interaction between ubiquitin proteasome system and non-LTR retrotransposons. One possible function of JCP is cleaving ubiquitin chains from the ORF protein of Dualen to obstruct degradation. Otherwise, JCP processes the polyprotein of Dualen itself, like one of its distant relatives, foot-and-mouth disease virus leader protease (Guarné et al. 1998).

The phylogeny of the josephin and the JCP domains supports the monophyly of Dualen, as well as the josephin family and the ataxin-3 family (Fig. 5E). The origin of the JCP domain of Dualen remains unclear, but both the josephin family and the ataxin-3 family are identified in various organisms including animals and plants, which shows that Dualen acquired a JCP domain at the early stage of the evolution of eukaryotes. It is also unclear whether this domain is a feature of ancestral non-LTR retrotransposons or only exists in the Dualen branch.

Evolution of non-LTR retrotransposons implicated by the extraordinary domain structure of Dualen

Figure 7A shows the reliable phylogeny of three domains, RT, APE, and RNH, and the existence of other two less conserved domains, Gag and RLE. The CCHC zinc-finger and RLE of the Dualen family and the R2 group have the common structure. The highly reliable results in phylogenetic analyses are that (1) RT of the Dualen clade belongs to the eldest subtype including the R2 group (including the CRE, the Genie, the R2, the R4, and the NeSL clades) and the L1 group (including only the L1 clade), (2) APE of the Dualen clade is the outgroup of the I group (including the I, the ingi, the R1, the LOA, and the Tad clades) and the Jockey group (including the Jockey, the Rex1, the L2, and the CR1 clades), and (3) RNH among the Dualen clade and the I group is monophyletic.

There are two possible origins of each domain of Dualen, from cellular genes or from other non-LTR retrotransposons. The simplest event that originated the domain structure of Dualen is that an I group element, which is the only non-LTR retrotransposon group having RNH except Dualen, transposed into an R2 group retrotransposon; however, it is unlikely, because the APE and the RT domains of Dualen are clearly phylogenetically distant from those of the I group elements.

RT and endonuclease activities are essential for non-LTR retrotransposons because of their retrotransposition mechanism (Feng et al. 1996; Yang et al. 1999; Cost et al. 2002; Takahashi and Fujiwara 2002). Retrotransposons cannot survive without endonucleases. The dual endonuclease structure of Dualen is the only way to change endonuclease domains without loss of retrotransposition ability. Since early branched retrotransposons have an RLE, the ancestor of Dualen could have newly acquired an APE. Dualen could have acquired an APE from a cellular gene, not from other non-LTR retrotransposons, because Dualen is the most ancient non-LTR retrotransposon having an APE. One of the possible mechanisms of acquiring a cellular APE is that an early branched retrotransposon transposed to just downstream of a cellular APE gene was cotranscribed and was comobilized. Comobilization of 5′-flanking sequences was experimentally demonstrated in the human L1 and the silkworm SART1 (Symer et al. 2002; Takahashi and Fujiwara 2002). RLE could have been lost after the branch between Dualen and recently branched non-LTR retrotransposons.

The RNH domain is considered to have been acquired independent of the APE domain, because the RNH domain of Dualen is positioned between the RT and the RLE domains, both of which are related to those of the early branched non-LTR retrotransposons. The RNH domains of non-LTR retrotransposons are monophyletic; thus, either the Dualen group or the I group could

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

have acquired a cellular RNH gene. If the I group had acquired a cellular RNH, Dualen would be a chimeric retrotransposon whose RNH was acquired from the I group retrotransposons. But, there have been no obvious reports of chimeras or domain swapping between two non-LTR retrotransposons. If Dualen had acquired a cellular RNH, the L1, the RTE, and the Jockey groups would have lost their RNH domains secondarily (Fig. 7B). The loss of the RNH domains actually occurred in the I group, as in the R1 clade and the Tad clade (Malik et al. 1999). Thus, it is more likely that Dualen is the most ancient retrotransposon that acquired cellular RNH. If it is true, the common ancestor of the recently branched non-LTR retrotransposons corresponding to the four groups, the L1, the RTE, the Jockey and the I groups, had Gag, APE, and RNH, in addition to RT and ZF (Fig. 7B, putative ancestor).

In addition, since LTR retrotransposons and retroviruses are considered to have evolved from non-LTR retrotransposons having both Gag and RNH (Malik and Eickbush 2001), LTR retrotransposons could have branched from the common ancestor of the recently branched non-LTR retrotransposons. The most ancient LTR retrotransposon group is the Ty1/copia group, which retains protease and integrase domains, instead of the APE domain of non-LTR retrotransposons (Fig. 7B). If LTR retrotransposons had branched from the common ancestor of the recently branched non-LTR retrotransposons, APE, RNH, Gag, protease, and integrase would have acquired sequentially. It is a quite exciting possibility.

Methods

Computer-based nucleotide and protein searches were performed using different BLAST search programs (Altschul et al. 1997) at NCBI (http://www.ncbi.nlm.nih.gov/BLAST) and JGI (http://aluminum.jgi-psf.org/prod/bin/runBlast.pl?db=chlre1). Protein sequences of non-LTR retrotransposons previously described (Kojima and Fujiwara 2004) were used as queries for database searches. As the *Chlamydomonas reinhardtii* genomic sequences were in draft format and there were no single sequences containing complete retrotransposons, we constructed representative retrotransposon sequences from several sequences derived from different genomic positions. Sequences more than 90% identical to each other were connected in order to include longer ORFs. The reconstructed sequences of the Dualen elements from *C. reinhardtii* are available from the authors' Web site (http://www.biol.s.u-tokyo.ac.jp/users/animal/kojima/sequence.html). DualenU1 corresponds to bases 20173–30250 of AC109923 and DualenU2 corresponds to bases 133667–134928 joined to 89898–95300 of AC109923.

Amino acid sequences of elements were aligned using CLUSTAL X (Thompson et al. 1997) on the basis of previous reports (Malik et al. 1999; Burke et al. 2002; Kojima and Fujiwara 2004; Weichenrieder et al. 2004). Bayesian phylogenetic trees were constructed using MrBayes 3 (Ronquist and Huelsenbeck 2003). Neighbor-joining (NJ) trees were constructed using CLUSTAL X. Nonparametric bootstrap analyses were performed with 1000 replicates.

Primers used for PCR are listed in Supplemental Table 1. Probes used in Southern hybridization were amplified by PCR with pairs of primers as follows: *Cr1-5′*, DualenCr1_F2 and DualenCr1_R2; *Cr1-3′*, DualenCr1_F1 and DualenCr1_R1; *Cr2*, DualenCr2_F1 and DualenCr2_R1; *Cr3*, DualenCr3_F1 and DualenCr3_R1; *Cr4*, DualenCr4_F1 and DualenCr4_R1. PCR conditions were as follows: 35 cycles of 96°C for 30 sec, 60°C for 30 sec, and 72°C for 1min.

Approximately 5 µg of genomic DNA was digested with respective restriction enzymes (EcoRI, BglII, and HindIII), separated on 1.0% agarose gel and blotted onto Hybond-N^+ nylon membrane (Amersham) in 0.4N NaOH. Radioactive probes were obtained by using *Bca*BEST Labeling Kit (TaKaRa) with [α-^{32}P]dCTP (ICN). Hybridization was performed at 45°C in 50% formamide, 10× Denhardt's solution (1× Denhardt's solution is 0.2% each of BSA, Ficoll, and polyvinylpyrrolidone), 50 mM sodium phosphate (pH 7.0), and 25 µg/mL sonicated salmon sperm DNA in 5× SSC, and the ^{32}P-labeled DNA probe. Post-hybridization washes were carried out in 2× SSC with 0.1% SDS for 15 min at 65°C and 0.2× SSC with 0.1% SDS for 15 min at 65°C.

Acknowledgments

Chlamydomonas reinhardtii sequence data were produced by the US Department of Energy Joint Genome Institute, http://www.jgi.doe.gov/ and are provided for use in this publication/correspondence only. Genomic DNA of *Chlamydomonas reinhardtii* was kindly provided by Masafumi Hirono, and genomic DNA of *Arabidopsis thaliana* by Mari Kurosawa and Kintake Sonoike. We thank Hiroyuki Toh, Mizuko Osanai, and Hideyuki Aoyagi for discussions and critical reading of the manuscript. This work was supported by grants from the Ministry of Education, Science and Culture of Japan (MESCJ), by a Grant-in-Aid from the Research for the Future Program of the Japan Society for the Promotion of Science (JSPS), and by the JSPS Research Fellowships for Young Scientists.

References

Albrecht, M., Hoffmann, D., Evert, B.O., Schmitt, I., Wullner, U., and Lengauer, T. 2003. Structural modeling of ataxin-3 reveals distant homology to adaptins. *Proteins* **50:** 355–370.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Anantharaman, V. and Aravind, L. 2003. Evolutionary history, structural features and biochemical diversity of the NlpC/P60 superfamily of enzymes. *Genome Biol.* **4:** R11.

Anzai, T., Takahashi, H., and Fujiwara, H. 2001. Sequence-specific recognition and cleavage of telomeric repeat (TTAGG)*n* by endonuclease of non-long terminal repeat retrotransposon TRAS1. *Mol. Cell. Biol.* **21:** 100–108.

Arabidopsis Genome Initiative 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796–815.

Arkhipova, I. and Meselson, M. 2000. Transposable elements in sexual and ancient asexual taxa. *Proc. Natl. Acad. Sci.* **97:** 14473–14477.

Bouneau, L., Fischer, C., Ozouf-Costaz, C., Froschauer, A., Jaillon, O., Coutanceau, J.P., Korting, C., Weissenbach, J., Bernot, A., and Volff, J.N. 2003. An active non-LTR retrotransposon with tandem structure in the compact genome of the pufferfish *Tetraodon nigroviridis*. *Genome Res.* **13:** 1686–1695.

Burke, W.D., Malik, H.S., Rich, S.M., and Eickbush, T.H. 2002. Ancient lineages of non-LTR retrotransposons in the primitive eukaryote, *Giardia lamblia*. *Mol. Biol. Evol.* **19:** 619–630.

Burnett, B., Li, F., and Pittman, R.N. 2003. The polyglutamine neurodegenerative protein ataxin-3 binds polyubiquitylated proteins and has ubiquitin protease activity. *Hum. Mol. Genet.* **12:** 3195–3205.

Cost, G.J., Feng, Q., Jacquier, A., and Boeke, J.D. 2002. Human L1 element target-primed reverse transcription in vitro. *EMBO J.* **21:** 5899–5910.

Douady, C.J., Delsuc, F., Boucher, Y., Doolittle, W.F., and Douzery, E.J. 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Mol. Biol. Evol.* **20:** 248–254.

Eickbush, T.H. and Malik, H.S. 2002. Origins and evolution of retrotransposons. In *Mobile DNA II* (eds. N.L. Craig et al.), pp. 1111–1144. American Society of Microbiology Press, Washington, DC.

Feng, Q., Moran, J.V., Kazazian Jr., H.H., and Boeke, J.D. 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87:** 905–916.

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

Guarné, A., Tormo, J., Kirchweger, R., Pfistermueller, D., Fita, I., and Skern, T. 1998. Structure of the foot-and-mouth disease virus leader protease: A papain-like fold adapted for self-processing and eIF4G recognition. *EMBO J.* **17:** 7469–7479.

Kojima, K.K. and Fujiwara, H. 2003. Evolution of target specificity in R1 clade non-LTR retrotransposons. *Mol. Biol. Evol.* **20:** 351–361.

———. 2004. Cross-genome screening of novel sequence-specific non-LTR retrotransposons: Various multicopy RNA genes and microsatellites are selected as targets. *Mol. Biol. Evol.* **21:** 207–217.

Li, S.J. and Hochstrasser, M. 1999. A new protease required for cell-cycle progression in yeast. *Nature* **398:** 246–251.

Malik, H.S. and Eickbush, T.H. 2000. NeSL-1, an ancient lineage of site-specific non-LTR retrotransposons from *Caenorhabditis elegans*. *Genetics* **154:** 193–203.

———. Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. *Genome Res.* **11:** 1187–1197.

Malik, H.S., Burke, W.D., and Eickbush, T.H. 1999. The age and evolution of non-LTR retrotransposable elements. *Mol. Biol. Evol.* **16:** 793–805.

Malik, H.S., Henikoff, S., and Eickbush, T.H. 2000. Poised for contagion: Evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome Res.* **10:** 1307–1318.

Mandal, P.K., Bagchi, A., Bhattacharya, A., and Bhattacharya, S. 2004. An *Entamoeba histolytica* LINE/SINE pair inserts at common target sites cleaved by the restriction enzyme-like LINE-encoded endonuclease. *Eukaryot. Cell* **3:** 170–179.

Marchler-Bauer, A., Anderson, J.B., DeWeese-Scott, C., Fedorova, N.D., Geer, L.Y., He, S. Hurwitz, D.I., Jackson, J.D., Jacobs, A.R., Lanczycki, C.J., et al. 2003. CDD: A curated Entrez database of conserved domain alignments. *Nucleic Acids Res.* **31:** 383–387.

Morrish, T.A., Gilbert, N., Myers, J.S., Vincent, B.J., Stamato, T.D., Taccioli, G.E., Batzer, M.A., and Moran, J.V. 2002. DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nat. Genet.* **31:** 159–165.

Ronquist, F. and Huelsenbeck, J.P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19:** 1572–1574.

Scheel, H., Tomiuk, S., and Hofmann, K. 2003. Elucidation of ataxin-3 and ataxin-7 function by integrative bioinformatics. *Hum. Mol. Genet.* **12:** 2845–2852.

Sewell, E. and Kinsey, J.A. 1996. Tad, a *Neurospora* LINE-like retrotransposon exhibits a complex pattern of transcription. *Mol. Gen. Genet.* **252:** 137–145.

Symer, D.E., Connelly, C., Szak, S.T., Caputo, E.M., Cost, G.J., Parmigiani, G., and Boeke, J.D. 2002. Human L1 retrotransposition is associated with genetic instability in vivo. *Cell* **110:** 327–338.

Takahashi, H. and Fujiwara, H. 2002. Transplantation of target site specificity by swapping the endonuclease domains of two LINEs. *EMBO J.* **21:** 408–417.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. 1997. The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25:** 4876–4882.

Waugh, D.S. and Sauer, R.T. 1993. Single amino acid substitutions uncouple the DNA binding and strand scission activities of *Fok* I endonuclease. *Proc. Natl. Acad. Sci.* **90:** 9596–9600.

Weichenrieder, O., Repanas, K., and Perrakis, A. 2004. Crystal structure of the targeting endonuclease of the human LINE-1 retrotransposon. *Structure* **12:** 975–986.

Yamamoto, Y., Fujimoto, Y., Arai, R., Fujie, M., Usami, S., and Yamada, T. 2003. Retrotransposon-mediated restoration of *Chlorella* telomeres: Accumulation of Zepp retrotransposons at termini of newly formed minichromosomes. *Nucleic Acids Res.* **31:** 4646–4653.

Yang, J., Malik, H.S., and Eickbush, T.H. 1999. Identification of the endonuclease domain encoded by R2 and other site-specific, non-long terminal repeat retrotransposable elements. *Proc. Natl. Acad. Sci.* **96:** 7847–7852.

Web site references

http://www.ncbi.nlm.nih.gov/BLAST; The BLAST server at the National Center for Biotechnology Information.

http://aluminum.jgi-psf.org/prod/bin/runBlast.pl?db=chlre1; *Chlamydomonas reinhardtii* BLAST server at the US Department of Energy Joint Genome Institute.

http://www.jgi.doe.gov/; the US Department of Energy Joint Genome Institute.

http://www.biol.s.u-tokyo.ac.jp/users/animal/kojima/sequence.html; authors' Web site.

Received September 21, 2004; accepted in revised form May 10, 2005.