Bioinformatic Annotation of Transposon DNA Processing Genes on the Long-Read Genome Assembly of *Caenorhabditis elegans*

Bioinformatics and Biology Insights
Volume 18: 1–16
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/11779322241304668

Yukinobu Arata[1], Peter Jurica[1], Nicholas Parrish[2] and Yasushi Sako[1]

[1]Cellular Informatics Laboratory, Cluster for Pioneering Research (CPR), RIKEN, Saitama, Japan.
[2]Genome Immunobiology RIKEN Hakubi Research Team, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan.

ABSTRACT: Transposable elements (TEs) or transposons are thought to play roles in animal physiological processes, such as germline, early embryonic, and brain development, as well as aging. However, their roles have not been systematically investigated through experimental studies. In this study, we created a catalog of genes directly involved in replication, excision, or integration of transposon-coding DNA, which we refer to as transposon DNA processing genes (TDPGs). Specifically, to bridge the gap to experimental studies, we sought potentially functional TDPGs which maintain intact open reading frames and the amino acids at their catalytic cores on the latest long-read genome assembly of *Caenorhabditis elegans*, VC2010. Among 52 519 TE loci, we identified 145 potentially functional TDPGs encoded in long terminal repeat elements, long interspersed nuclear elements, terminal inverted repeat elements, Helitrons, and Mavericks/Polintons. Our TDPG catalog, which contains a feasible number of genes, allows for the experimental manipulation of TE mobility *in vivo*, regardless of whether the TEs are autonomous or non-autonomous, thereby potentially promoting the study of the physiological functions of TE mobility.

KEYWORDS: *C. elegans*, transposon, transposable element, transposon DNA processing gene

RECEIVED: June 14, 2024. **ACCEPTED:** November 13, 2024.

TYPE: Research Article

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by JSPS KAKENHI grant number 20K20321 (Grant-in-Aid for Challenging Research (Exploratory)) (Y.A.).

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Yukinobu Arata, Cellular Informatics Laboratory, Cluster for Pioneering Research (CPR), RIKEN, S356, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan. Email: arata@riken.jp

Introduction

Transposable element (TE)-related repetitive DNA elements occupy significant proportions of eukaryotic genomes: 12% in *Caenorhabditis elegans*, 4% to 17% in *Drosophila melanogaster*, 37% in mouse, 46% in human, and ~85% in corn.[1,2] Transposable elements are thought to play roles in animal physiological processes, such as germline and early embryonic and brain development,[3] as well as aging.[4] However, experimental investigations through the systematic manipulation of TE activity have rarely been performed. Due to the large number of TEs present in any given genome, it is necessary to devise new approaches for their systematic manipulation.

TEs transpose through enzymes encoded within their own regions. Depending on the medium required for transposition, TEs are primarily classified into retrotransposons (class I) and DNA transposons (class II), and these include both autonomously mobile and non-autonomously mobile TEs.[5-7] Autonomously mobile class I and II TEs are further subdivided by the transposition mechanism[6] into the 4 retrotransposon orders: the long terminal repeat (LTR) elements, long interspersed nuclear elements (LINEs), *Dictyostelium* intermediate repeat sequences (DIRS), and Penelope-like elements (PLEs) as well as the 4 DNA transposon orders: the terminal inverted repeat (TIR) elements, Helitrons, Maverick/Polintons (MP), and Cryptons. Regarding retrotransposons, LTR elements transpose by generating DNA copies from a RNA transcript via reverse transcriptase (RT). Integrase (IN) is required to efficiently insert LTR element DNA fragments into new genomic loci. Similarly, LINEs transpose by replicating DNA copies from a transcript via LINE-specific RT. In this case, reverse transcription starts at the 3′-end of the nick site on the genomic DNA generated via the Endonuclease (EN) domain at the N-terminus of the RT. *Dictyostelium* intermediate repeat sequence transposition occurs when free circular double-stranded DNAs are generated via reverse transcription. The circular double-stranded DNAs are then integrated into other loci by Tyrosine-Recombinase (Tyr-REC).[8-11] In PLE transposition, the DNA fragment of PLE is reverse-transcribed from the 3′ end of the nick site on the genome generated by the GIY-YIG EN activity in the PLE-specific RT.[12] As for DNA transposons, the TIR elements transpose via the EN activity of Transposase (TP). The TP recognizes TIRs at both ends of TIR element and cleaves double-stranded DNA (dsDNA) to generate a free DNA fragment. This fragment is integrated into another genomic locus by the EN activity of TP.[13-15] Helitron transposition occurs when the Helitron DNA element is nicked via the EN activity of the REP domain in REP-Helicase (REP-HEL) and is then unwound into single-stranded DNA (ssDNA) through HEL activity. Although it remains controversial whether the Helitron ssDNA is replicated before integration into another genomic locus, a free Helitron fragment is integrated into another locus by EN activity of the REP domain.[16-20] In MPs, free DNA copies are replicated by DNA Polymerase B encoded within MP. These MP DNA copies are inserted into other genomic loci by MP-encoded INs.[21-24] Cryptons transpose after being excised as free circular DNAs (circDNAs) by Tyr-REC. The free circDNAs are integrated into another genome locus by Tyr-REC.[11,25]

As described above, the transposition of TEs can be conceptualized as occurring through 2 DNA processing reactions: (1) the production of a free TE DNA and (2) its integration into another genomic locus. In retrotransposons LTR element and DIRS, and DNA transposon MP, the free TE DNA is generated and integrated by different enzymes. In retrotransposons LINE and PLE, and DNA transposon Helitron, these processes are mediated by different enzymatic activities associated with each functional domain in the same enzyme. In DNA transposons TIR element and Crypton, these processes are mediated by the same functional domains in the same enzyme. Hereafter, we refer to the genes involved in the production and/or integration of free TE DNAs as transposon DNA processing genes (TDPGs).

Detailed analyses of the open reading frame (ORF) structure and amino acid sequence of TDPGs have been primarily focused on individual TE orders.[26-33] These sequence analyses alone offer limited insights into their roles in TE mobile activity. For instance, a TDPG containing numerous mutations or deletions may still retain enough catalytic activity to facilitate transposition *in vivo*. In this study, we have cataloged potentially functional TDPGs, which maintain the ORF structure and the conserved amino acids at the catalytic core across all representative TE orders in a single species, *C. elegans*. This work provides a foundation for systematically investigating the physiological roles of TE mobility in *C. elegans*, paving the way to seek conserved mechanisms across different organisms.

Methods

Bioinformatic analysis

The VC2010 genome assembly was downloaded from https://www.ebi.ac.uk/ena/browser/view/UNSB01. To identify TEs, RepeatMasker version 4.1.0 (http://www.repeatmasker.org/) was used with options -no_is for skipping bacterial insertion element check, -s: slow search for more sensitivity, and -pa 8 for sequencing batch jobs to run in parallel. For RepeatMasker analysis, we used Dfam_3.1 library.[34] To identify ORFs, SNAP (Semi-HMM-based Nucleic Acid Parser), version 2006-07-28,[35] was used to analyze a FASTA file containing all the repeat sequences identified by RepeatMasker. The command used was snap HMM/C.elegans.hmm TEgenomeseq.fasta. Here, C.elegans.hmm contained parameters optimized for the *C. elegans* genome, and TEgenomeseq.fasta contained all the repeat sequences identified by RepeatMasker. To infer the function of proteins encoded in ORFs, DIAMOND (Double Index AlignMent Of Next-generation sequencing Data) (v2.0.2.140)[36] blastx tool was used to compare our ORF sequences identified by SNAP against the UniRef50 protein database. Default parameters were employed, with no additional sequence masking or complexity adjustments. The search sensitivity was set to the default "fast" mode. The following command line was used for the search: diamond blastx –db uniref50.fasta –query TDPGsORF.fasta –out blastx.txt –outfmt 6 qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore qlen slen. Here, TDPGsORF.fasta contained ORF sequences identified by SNAP. No specific e-value threshold was set; hence, the default cutoff of 0.001 was applied. UniRef50 protein ID produced by DIAMOND was translated into functional protein IDs using the website Uniprot (https://www.uniprot.org/uploadlists/). To align amino acid sequences, MAFFT v7.453[37,38] was used. A sequence alignment viewer was downloaded from https://github.com/dmnfarrell/teaching/blob/master/pyviz/bokeh_sequence_align.ipynb.

Results

Bioinformatic identification of TDPGs in the C. elegans *genome*

We first applied RepeatMasker, an algorithm that identifies TEs (https://www.repeatmasker.org/), to the latest *C. elegans* VC2010 genome assembly (for the detail, see Materials and Methods). The VC2010 genome assembly was constructed using long-read sequencing technology.[39] Long-read sequencing helps to accurately determine the genomic locations of repeat elements, such as TEs, and thereby results in more accurate estimates of the copy numbers of TEs in the genome, which is challenging for short-read sequencing. Our RepeatMasker analysis showed that repeat elements occupy 13.86% of the *C. elegans* genome (Supplementary Table 1). This result is consistent with previous reports that repetitive elements comprise 12% to 17% of the *C. elegans* genome.[28,40,41] After excluding simple, satellite, and low-complexity repeat elements, we identified 52 519 TE loci (Supplementary Table 1). To identify protein-coding elements among these 52 519 TE copies, we used an *ab initio* gene finder, SNAP.[35] The SNAP analysis identified 808 potential gene-coding regions. Among them, 428 genes conserved the complete ORF structure. To infer the function of genes encoded in these complete ORFs, we used DIAMOND, an algorithm for fast and sensitive protein alignment.[36] DIAMOND identified 80 RT, 11 IN, 189 TP, and 5 REP-HEL genes, for a total of 285 TDPGs (for the detail, see Materials and Methods). In following sections, we systematically dissected these genes to create a curated list of potentially functional genes.

TDPGs in LTR elements in C. elegans

Among the 80 RT genes, 19 genes were encoded in LTR elements (*rtz_LTRs*) and 61 genes were encoded in LINEs (*rtz_LINEs*) (Supplementary Tables 2 and 4). Reverse transcriptases, ie, RNA-dependent DNA polymerases, have 5 evolutionarily conserved motifs (A, B′, C, D, and E).[42,43] Asp residues in Motifs A and C are widely conserved in all RNA/DNA-dependent DNA polymerases and DNA/RNA-dependent RNA polymerases.[44] Previous crystal structure analyses showed that 1 Asp in Motif A and 2 Asp residues in Motif C form a catalytic triad for holding 2 bivalent metal ions for conjugating the alpha phosphate of a new dNTP to the OH group to the 3′ end of the DNA strand.[45-47] Mutations in Asp residues of Motifs A or C reportedly abolish RT activity.[48-51] By aligning RTZ_LTRs with the reference RTs, we identified 8 RTZ_LTRs with conserved Asp residues in Motifs A and C (red asterisks in Figure 1A and Table 1). In addition, these 8 RTZ_LTRs preserved (1) the conserved Gly residue in Motif B′,[52] which interacts with the incoming nucleotide and template strand,[53] and (2) the Leu and Gly residues in Motif E, which fixes the primer strand and positions it toward the active site[53] (Figure 1C and Table 1).

In RTZ_LTR-11, RTZ_LTR-17, and RTZ_LTR-19, the second Asp residue in Motif C was substituted with Asn (black asterisk in Figure 1A and Table 1). As described earlier, the second Asp residue is involved in a catalytic triad. Site-directed mutagenesis at the second Asp residue, including mutagenesis to Asn as found in RTZ_LTR-11, RTZ_LTR-17, and RTZ_LTR-19, significantly reduces RT activity in human immunodeficiency virus (HIV).[48-51] However, amino acid substitution at the second Asp to Asn has been found in functional RNA-dependent RNA polymerases in negative-strand RNA viruses.[42,44,54] Therefore, we considered the possibility that RTZ_LTR-11, RTZ_LTR-17, and RTZ_LTR-19 might still be functional.

In RTZ_LTR-11 and RTZ_LTR-19, the conserved Gly and Lys in Motif D were substituted. In RTZ_LTR-3, RTZ_LTR-8, and RTZ_LTR-17, the conserved Gly in Motif D was substituted. Amino acid substitutions at Gly in Motif D are often observed in RNA-dependent RNA polymerases in negative-strand RNA viruses.[42] Therefore, we considered the possibility that RTZ_LTR-3, RTZ_LTR-8, and RTZ_LTR-17 might still be functional. On the other hand, Lys in Motif D is highly conserved throughout polymerases.[42,43] Motif D functions for forming a phosphodiester bond with dNTPs with the 3′-OH (hydroxyl) of the primer.[55,56] Motif D in RTZ_LTR-11 and RTZ_LTR-19 may be nonfunctional. Nevertheless, given the conservation of the 2 critical amino acids holding 2 bivalent metal ions, and the fact that we did not perform functional testing, we include RTZ_LTR-11 and RTZ_LTR-19 as potentially functional TDPGs. Taken together, these results led us to classify all 8 *rtz_LTR*s as potentially functional genes (Table 1 and Supplementary Table 2).

Next, we identified 11 IN genes encoded in LTR elements (*inz_LTR*s) (Supplementary Table 3). The catalytic core domain of IN has an evolutionarily conserved DD35E motif that is required for EN activity.[57-59] DD35E holds 2 metal ions required for the catalysis involved in integrating a free DNA fragment into the genome.[60,61] Amino acid substitution at the 3 critical amino acid residues in the DD35E motif abolishes EN activity of INs in Rous sarcoma virus (RSV) and HIV.[57-59,62] By aligning the 11 INZ_LTRs with the reference INs, we identified 7 INZ_LTRs with conserved DD35E triads (Figure 1B, Table 1 and Supplementary Table 3). Therefore, we considered these 7 *inz_LTR*s as potentially functional genes. Potentially functional RT and IN genes generally co-occurred in LTR element loci, except for the loci encoding *rtz_LTR-11* and *rtz_LTR-17*, which lacked the corresponding *inz_LTR*, and the locus encoding *inz_LTR-10*, which lacked the corresponding *rtz_LTR* (Figure 1C and Table 1).

TDPGs in LINEs in C. elegans

By aligning the 61 remaining RTs (from the initial 80 RTs, excluding the 19 RTs analyzed as RTZ_LTRs in the previous section) with reference RTs, we identified 28 RTZ_LINEs that had Asp residues conserved in Motifs A and C (red asterisks in Figure 2A and Table 2). In addition, these 28 RTZ_LINEs had residues conserved in Motifs B′ and C (black asterisks in Figure 2A and Table 2). These 28 RTZ_LINEs had Lys but not Gly residues conserved in Motif D. Amino acid substitution of Gly in Motif D is often observed in RNA-dependent RNA polymerases in negative-strand RNA viruses.[42] Therefore, we held out the possibility that the RTZ_LTRs with amino acid substitution of Gly in Motif D are functional. In addition, in RTZ_LINE-61, Gly residues in Motif E were substituted. The substitution in this residue was found in RT in HIV-1 and Hepatitis B virus (HepB),[42] leading us to consider the possibility that RTZ_LINE-61 is potentially functional (black asterisks in Figure 2A and Table 2). In summary, analysis of RT domains suggests that these 28 *rtz_LINE*s are potentially functional.

Next to study the EN domain, we aligned the 61 RTZ_LINEs with reference ENs. A previous experimental test of transposition activity in 138 human L1 copies revealed that the Asp and His residues are essential, but conserved amino acid residues in other motifs tolerate multiple mutations.[63] We identified 14 RTZ_LINEs in which the critical Asp and His residues were conserved in their EN domains (red asterisks in Figure 2B and Table 2). Among the 14 RTZ_LINEs, RTZ_LINE-19 and RTZ_LINE-34 lacked a large N-terminal portion of the EN domain, whereas RTZ_LINE-44 and RTZ_LINE-61 lacked about half of this portion (Figure 2B). Nevertheless, in the absence of experimental tests of the function of N-terminal deletion EN mutants, we held out the possibility that all the 14 RTZ_LINEs have potentially functional EN domains. Taken together, we concluded that 6 *rtz_LINE*s encoded both potentially functional RT and EN domains (*rtz_LINE-5*, *rtz_LINE-13*, *rtz_LINE-22*, *rtz_LINE-57*, *rtz_LINE-59*, and *rtz_LINE-61*; Figure 2A and B, Table 2 and

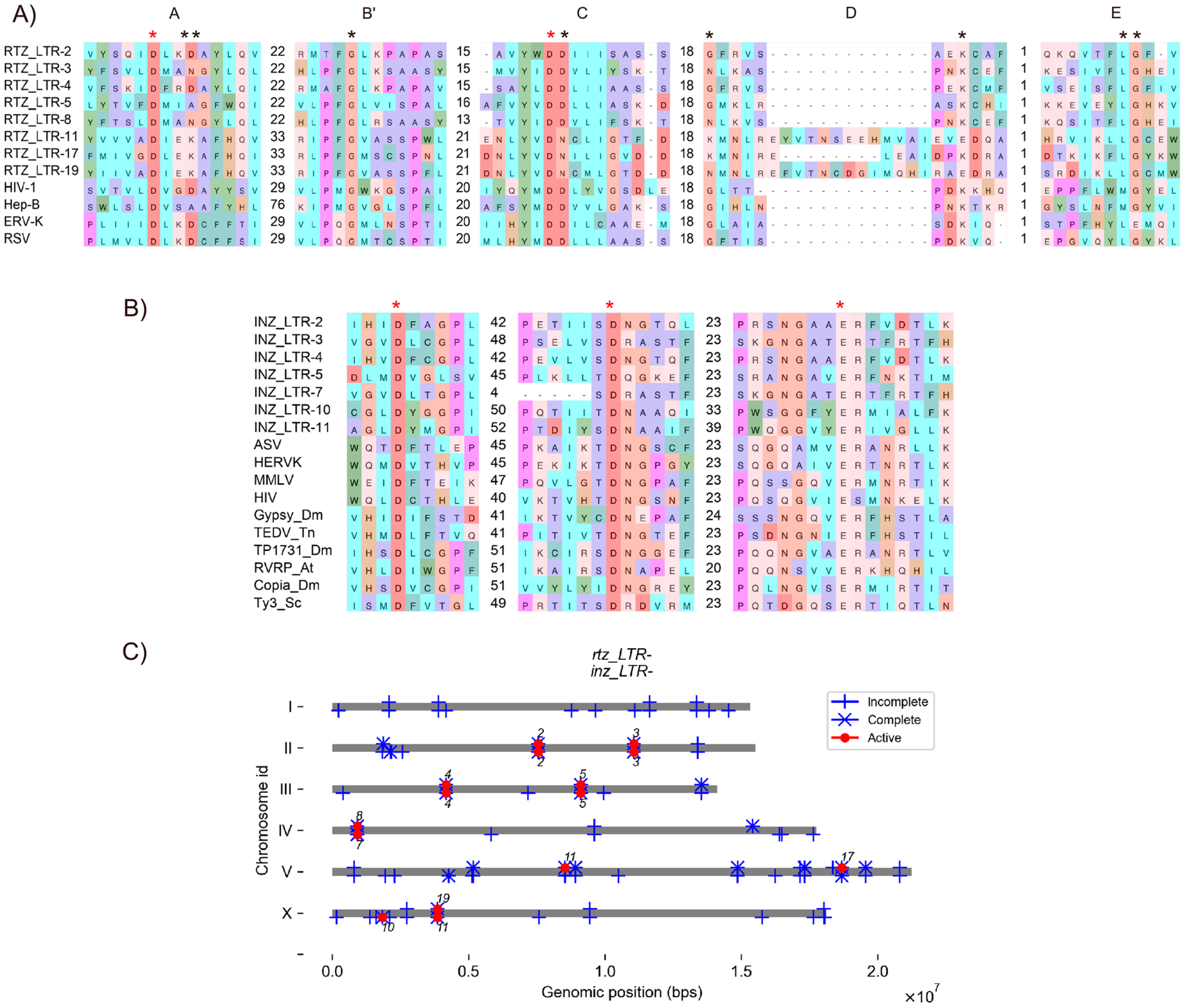


Figure 1. Potentially functional *rtz_LTR*s and *inz_LTR*s encoded in long terminal repeat (LTR) elements. (A) Alignment of catalytic core domains of RTZ_LTRs with reference Reverse transcriptases (RTs). Abbreviations: HIV-1: Chain C, HIV-1 RT P66 subunit of human immunodeficiency virus type 1 [5TXO_C], Hep-B: Hepatitis B virus RT_like family [QFR04538], ERV-K: Pol protein of human endogenous retrovirus K [CAA76882], RSV: Pol of Rous sarcoma virus [CAA48535]. (B) Alignment of catalytic core domains of INZ_LTRs with reference Integrases (INs). Abbreviations: ASV: IN in avian sarcoma virus [1ASU_A], HERVK: Pol protein in human endogenous retrovirus K [CAA76885], MMLV: p46 IN in Moloney murine leukemia virus (MoMLV) [NP_955592.1], HIV: IN in human immunodeficiency virus type 1 [1BIZ_A], Gypsy_Dm: IN, Gypsy endogenous retrovirus in *Drosophila melanogaster* [CAB69645], TEDV_Tm: ORFB in TED virus in *Trichoplusia ni* [YP_009507248], TP1731: Pol polyprotein in transposon_1731 in *D. melanogaster* [S00954], RVRP_At: retrovirus-related like polyprotein in *Arabidopsis thaliana* [CAB78488_1], Copia_Dm: Gag-Int-Pol protein in COPIA in *D. melanogaster* [P04146], Ty-3_Sc: Gag-Pol polyprotein in Ty3-G in *Saccharomyces cerevisiae* [GFP69998.1]. Asterisks indicate conserved residues. Red asterisks indicate residues for identifying potentially functional TDPGs. (C) Genomic positions of 8 potentially functional *rtz_LTR*s and 7 potentially functional *inz_LTR*s in the *Caenorhabditis elegans* genome. Gray lines represent chromosomes. Red circles over and under chromosome indicate positions of potentially functional genes of *rtz_LTR*-n and *inz_LTR*-n, respectively. Numbers over and under chromosome indicate numbers of *rtz_LTR* and *inz_LTR* genes, respectively. Vertical ticks and × marks with vertical ticks over and under chromosome indicate incomplete and complete ORFs, respectively, of *rtz_LTR*s and *inz_LTR*s, respectively.

Supplementary Table 4). However, it has been noted that human L1 can transpose via an EN-independent, RT-dependent mechanism.[64] Therefore, we concluded that the 28 *rtz_LINE*s with conserved RT domains (Figure 2A and Table 2) are potentially functional genes for LINE transposition. The 28 *rtz_LINE*s distributed across each chromosome (Figure 2C) and amino acid sequences of these genes could be grouped into 4 homologous clusters (Supplementary Figure 1 and Supplementary Table 4).

TDPGs in TIR elements in C. elegans

Terminal inverted repeat element is composed of 19 super families: hAT, Tc1/mariner, CACTA (En/Spm), Mutator (MuDR),

Table 1. List of potentially functional *rtz_LTR*s and *inz_LTR*s.

RTZ ID	INZ ID	CLASS	CHR	POSITION	RT	IN
rtz_LTR-2	*inz_LTR-2*	CER5-I_CELTR/Gypsy	II	(7556985, 7561601)	D G DD G K LG	D D E
rtz_LTR-3	*inz_LTR-3*	CER2-I_CELTR/Gypsy	II	(11067565, 11073355)	D G DD N K LG	D D E
rtz_LTR-4	*inz_LTR-4*	CER6-I_CELTR/Gypsy	III	(4178552, 4183161)	D G DD G K LG	D D E
rtz_LTR-5	*inz_LTR-5*	CER1LTR/Gypsy	III	(9116926, 9123745)	D G DD G K LG	D D E
rtz_LTR-8	*inz_LTR-7*	CER3-I_CELTR/Gypsy	IV	(923461, 926380)	D G DD N K LG	D D E
rtz_LTR-11	-	CER10-I_CELTR/Pao	V	(8540078, 8541860)	D G DN K E LG	- - -
rtz_LTR-17	-	CER8-I_CELTR/Pao	V	(18702294, 18708087)	D G DN K K LG	- - -
-	*inz_LTR-10*	CER13-I_CELTR/Pao	X	(1841438, 1849184)	- - -- - - --	D D E
rtz_LTR-19	*inz_LTR-11*	CER11-I_CELTR/Pao	X	(3867459, 3871374)	D G DN N E LG	D D E

Evolutionarily conserved amino acids within the Reverse transcriptase ("RT") and Integrase ("IN") within potentially functional RTZ_LTRs and INZ_LTRs. The consecutive amino acid single letters without spaces indicate conserved amino acids within the same domain. See amino acids indicated by asterisks in Figure 1A and B. A dash (-) indicates the absence of an ID or conserved amino acid. The "Class" denotes TE identity as determined by RepeatMasker analysis. Chromosome ID ("Chr") and genomic positions ("Position") are provided for each gene.

P, PiggyBac, PIF/Harbinger, Mirage, Merlin, Transib, Novosib, Rehavkus, ISL2EU, Kolobok, Chapaev, Sola, Zator, Ginger, and Academ.[65] The transposition of TIR elements is mediated by Transposase (TP), which has a conserved DDD/E motif at the catalytic core.[65] Structural analysis suggests that the DDD/E motif holds 2 metal ions to cleave and integrate dsDNA of the TIR element.[66-68] Mutation of the conserved DDD/E motif abolishes TP activity.[69,70] By aligning the 189 TPZs with reference TPs, we identified 94 TPZs that had DDD/E motifs conserved that were homologous to those of Tc1/mariner family TPs (Figure 3A and Supplementary Table 5). No TP aligned to DDD/E motifs of other TP super families. Thus, we considered these 94 *tpzs* to be potentially functional genes (Supplementary Table 5). *tpzs* were distributed across each chromosome (Figure 3B) and amino acid sequences of these genes predominantly formed 3 main homologous clusters (Supplementary Figure 2 and Supplementary Table 5).

TDPGs in Helitrons in C. elegans

Helitron transposition is mediated by a SF1 family Helicase (HEL) with REP domain (REP-HEL). The 4 SF family HELs (SF1, SF2, SF3, and SF4) have conserved Motif I and Motif II domains. Motifs I and II correspond to the Walker A and Walker B domains, respectively, which are widely conserved among NTP-binding proteins.[71,72] The Walker A/Motif I and Walker B/Motif II domains exhibit conservation of the Lys and Asp-Glu residues, respectively.[72-75] Crystallographic analysis showed that the Lys in Motif I contacts a magnesium ion and β phosphate of ATP and functions to stabilize the transition state during ATP hydrolysis. Asp-Glu residues in Motif II are also involved in ATP hydrolysis.[76,77] Mutations at the Lys in Motif I or Asp-Glu in Motif II abolish HEL activity.[78-81]

By aligning the 5 REP-HELs (RHZs) with reference HEL domains, we identified 5 RHZs that had Lys residues conserved in Motif I and had Asp-Glu residues conserved in Motif II (red asterisks in Figure 4A and Table 3). In addition, these 5 RHZs had (1) a conserved residue in Motif Ia (black asterisks in Figure 4A and Table 3), which functions for ssDNA binding and energy transfer from the ATP-binding site to the DNA-binding site[76]; (2) conserved Gly-Asp and other resides in Motif III (black asterisk in Figure 4A and Table 3), which is involved in contacting nucleotide γ-phosphates[76]; (3) a conserved Arg residue in Motif IV, which may be involved in NTP hydrolysis[76]; (4) a conserved residue in Motif IV/V, the function of which is not well understood[18]; (5) conserved residues in Motif V, which interacts with the sugar-phosphate backbone of DNA[76]; and (6) conserved residues in Motif VI, which may form part of the ATP-binding cleft to couple ATPase activity to HEL activity[76] (Figure 4A and Table 3). Therefore, we considered these 5 *rhzs* to encode potentially functional HEL domains.

In the REP domain, the HUH Y2 motif (in which U is a hydrophobic residue) is evolutionarily conserved.[16] HUH holds divalent metal ions to form nicks in the DNA strand in EN activity, whereas Tyr residues form a transient covalent bond with the cleaved DNA strand to generate phosphotyrosine for DNA strand transfer.[16] The EN activity is abolished by mutation of either 2 His or 2 Tyr residues.[17] By aligning the 5 RHZs with reference REP domains, we found that the 5 RHZs conserved the 2 His and 2 Tyr residues (red asterisks in Figure 4B and Table 3). Taken together, our results suggest that the 5 *rhzs* encode both potentially functional

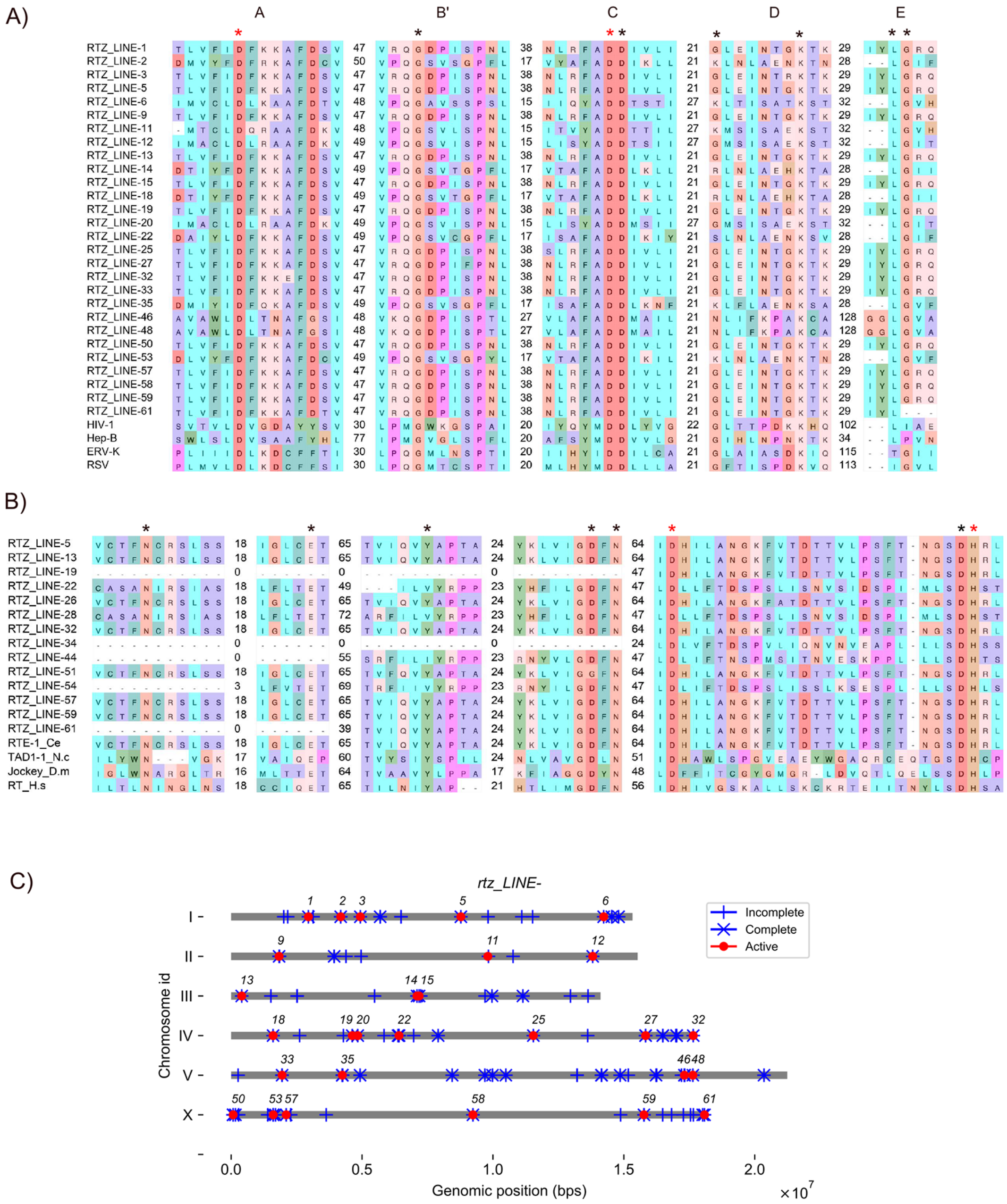


Figure 2. Potentially functional *rtz_LINE*s encoded in long interspersed nuclear elements (LINEs). (A) Alignment of reverse transcriptase (RT) domains of RTZ_LINEs with reference RTs. Abbreviations: HIV-1: Chain C, HIV-1 RT P66 subunit of human immunodeficiency virus type 1 [5TXO_C], Hep-B: hepatitis B virus RT_like family [QFR04538], ERV-K: Pol protein of human endogenous retrovirus K [CAA76882], RSV: Pol of Rous sarcoma virus [CAA48535]. (B) Alignment of the Endonuclease (EN) domains of RTZ_LINEs with reference ENs. Abbreviations: RTE-1_Ce: apurinic-apyrimidic EN domain containing RT of non-LTR retrotransposon in *Caenorhabditis elegans* [AAC72298.1], AP_End_Hs: DNA-(apurinic or apyrimidinic site) EN in *Homo sapiens* [NP_542379.1], TAD1-1_N.c: Exonuclease-Endonuclease-Phosphatase (EEP) domain containing Pol protein *Neurospora crassa* [AAA21781.1], Jockey_D.m: EEP domain containing RT in *Drosophila melanogaster* [AAA28675.1], RT_H.s: EN domain containing RT of L1, *H. sapiens* [AAB59368.1]. Asterisks indicate conserved residues. Red asterisks indicate residues for identifying potentially functional TDPGs. (C) Genomic positions of 28 potentially functional *rtz_LINE*s in the *C. elegans* genome. Gray lines represent chromosomes. Red circles indicate positions of potentially functional genes of *rtz_LINE-n*. Numbers indicate numbers of *rtz_LINE-n* genes. Vertical ticks and × marks indicate incomplete and complete ORFs, respectively, of *rtz_LINE*s.

HEL and EN domains (Table 3 and Supplementary Table 6). These 5 *rhzs* are located only on chromosome II (Figure 4C).

TDPGs in MPs in C. elegans

Our RepeatMasker analysis did not identify MP elements. Dfam_3.1 library contains 9 copies of Maverick and 20 copies of Polinton, excluding the absence of these elements from the repeat library as a potential cause. One possible cause is that MP is more than 10 kbp in length, longer than other TEs identified here; optimizing the parameters of RepeatMasker's algorithm may be necessary for efficient identification. Notably, MPs located on chromosomes I, II, III, IV, and X of the *C. elegans* genome assembly (PRJNA907379) have previously

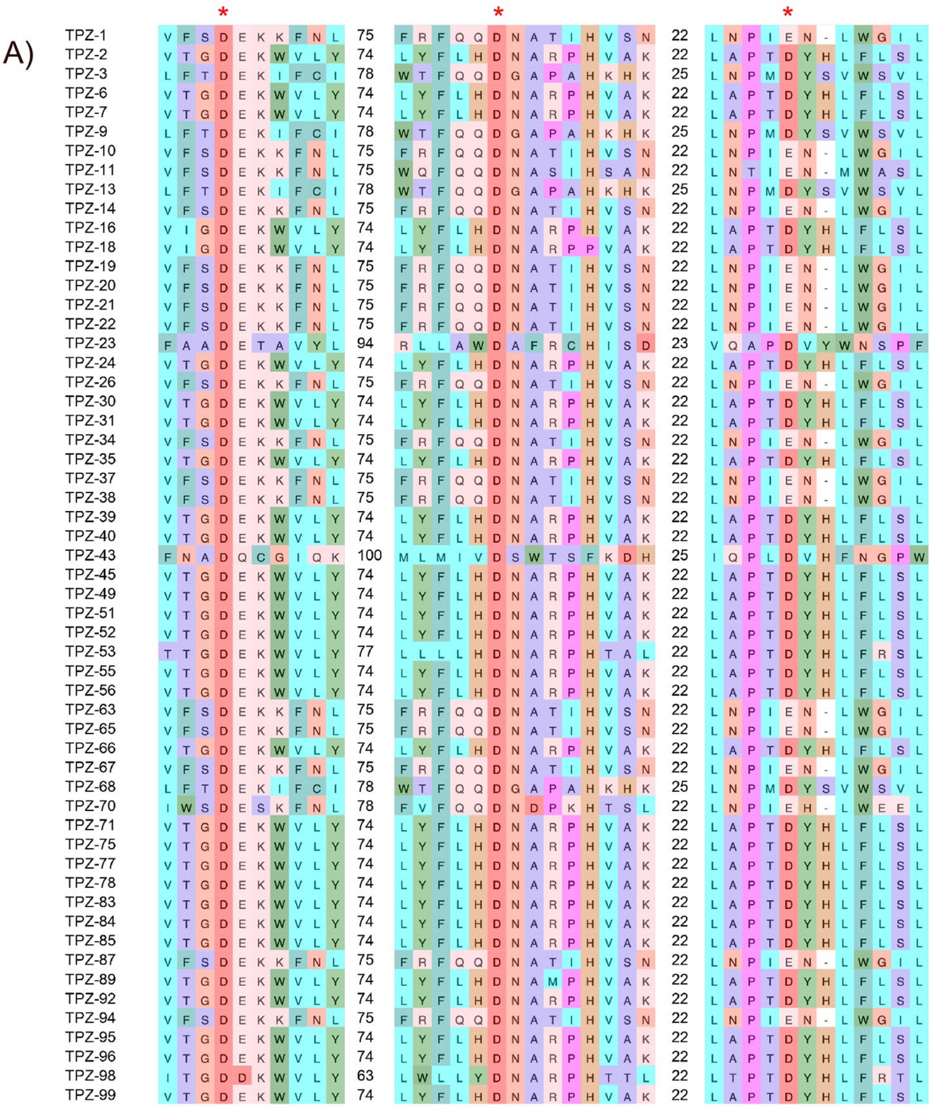
Table 2. List of potentially functional *rtz_LINE*s.

GENE ID	CLASS	CHR	POSITION	RT	EN	PC
rtz_LINE-1	RTE1LINE/RTE-RTE	I	(2950852, 2952124)	D G DD G K LG	- - - - - ---	PC1
rtz_LINE-2	LINE2E_CELINE/CR1	I	(4185973, 4187716)	D G DD K K LG	- - - - - ---	PC2
rtz_LINE-3	RTE1LINE/RTE-RTE	I	(4935419, 4937297)	D G DD G K LG	- - - - - ---	PC1
rtz_LINE-5	RTE1LINE/RTE-RTE	I	(8775178, 8778372)	D G DD G K LG	N E Y D N DDH	PC1
rtz_LINE-6	Vingi-2_CELINE/I-Jockey	I	(14229294, 14230404)	D G DD K K LG	- - - - - ---	PC3
rtz_LINE-9	RTE1LINE/RTE-RTE	II	(1831404, 1833282)	D G DD G K LG	- - - - - ---	PC1
rtz_LINE-11	Vingi-2_CELINE/I-Jockey	II	(9804786, 9806178)	D G DD K K LG	- - - - - ---	PC3
rtz_LINE-12	Vingi-1_CELINE/I-Jockey	II	(13809694, 13811320)	D G DD G K LG	- - - - - ---	PC3
rtz_LINE-13	RTE1LINE/RTE-RTE	III	(404229, 407423)	D G DD G K LG	N E Y D N DDH	PC1
rtz_LINE-14	LINE2C1_CELINE/CR1	III	(7100465, 7102105)	D G DD R K LG	- - - - - ---	PC2
rtz_LINE-15	RTE1LINE/RTE-RTE	III	(7180182, 7183418)	D G DD G K LG	- - - - - ---	PC1
rtz_LINE-18	LINE2C1_CELINE/CR1	IV	(1602836, 1604476)	D G DD R K LG	- - - - - ---	PC2
rtz_LINE-19	RTE1LINE/RTE-RTE	IV	(4625912, 4628285)	D G DD G K LG	- - - - - DDH	PC1
rtz_LINE-20	Vingi-1_CELINE/I-Jockey	IV	(4835118, 4837566)	D G DD G K LG	- - - - - ---	PC3
rtz_LINE-22	LINE2A_CELINE/CR1	IV	(6411447, 6414339)	D G DD S K LG	N E Y D N DDH	PC2
rtz_LINE-25	RTE1LINE/RTE-RTE	IV	(11542654, 11544867)	D G DD G K LG	- - - - - ---	PC1
rtz_LINE-27	RTE1LINE/RTE-RTE	IV	(15832895, 15833927)	D G DD G K LG	- - - - - ---	PC1
rtz_LINE-32	RTE1LINE/RTE-RTE	IV	(17649917, 17653111)	D G DD G K LG	N E Y D N DDH	PC1
rtz_LINE-33	RTE1LINE/RTE-RTE	V	(1946891, 1948145)	D G DD G K LG	- - - - - ---	PC1
rtz_LINE-35	LINE2A_CELINE/CR1	V	(4246672, 4247554)	D G DD K K LG	- - - - - ---	PC2
rtz_LINE-46	NeSL-1LINE/R2	V	(17312793, 17316369)	D G DD N K LG	- - - - - ---	PC4
rtz_LINE-48	NeSL-1LINE/R2	V	(17637370, 17642434)	D G DD N K LG	- - - - - ---	PC4
rtz_LINE-50	RTE1LINE/RTE-RTE	X	(75209, 76481)	D G DD G K LG	- - - - - ---	PC1
rtz_LINE-53	LINE2E_CELINE/CR1	X	(1608458, 1610063)	D G DD K K LG	- - - - - ---	PC2
rtz_LINE-57	RTE1LINE/RTE-RTE	X	(2102788, 2105982)	D G DD G K LG	N E Y D N DDH	PC1
rtz_LINE-58	RTE1LINE/RTE-RTE	X	(9235621, 9236893)	D G DD G K LG	- - - - - ---	PC1
rtz_LINE-59	RTE1LINE/RTE-RTE	X	(15766961, 15770155)	D G DD G K LG	N E Y D N DDH	PC1
rtz_LINE-61	RTE1LINE/RTE-RTE	X	(18078630, 18081306)	D G DD G K L-	- - Y D N DDH	PC1

Evolutionarily conserved amino acids within the Reverse transcriptase ("RT") and Endonuclease ("EN") domains within potentially functional RTZ_LINEs. The consecutive amino acid single letters without spaces indicate conserved amino acids within the same domain. See amino acids indicated by asterisks in Figure 2A and B. A dash (-) indicates the absence of a conserved amino acid. The "Class" denotes TE identity as determined by RepeatMasker analysis. Chromosome ID ("Chr") and genomic positions ("Position") are provided for each gene.

been reported.[21,22,82] Using the representative AL110478.1 sequence of MP in *C. elegans*,[82] we searched for MP copies in the VC2010 assembly by using the NCBI nucleotide BLAST program (https://blast.ncbi.nlm.nih.gov/). We identified short homologous regions >100 bps that were densely scattered on 3 distinct regions on chromosome I (MP Ia, MP Ib, and MP Ic) and 2 regions on chromosome X (MP Xa and MP Xb) (Figure 5A and B). Thus, the reference MP copy used in our homology search was aligned discontinuously in these putative MP copies in the VC2010 genome assembly (Figure 5B), and these MP copies were located on different chromosomes from previous reports.

The VC2010 genome assembly provides substantial advantages over its predecessors in both precision and completeness, but at the current stage, any assembly is imperfect.[39] Compared with the N2 genome assembly, the VC2010 assembly contains



```
TPZ-1   VFSDEKKFNL  75  FRFQQDNATIHVSN  22  LNPIEN-LWGIL
TPZ-2   VTGDEKWVLY  74  LYFLHDNARPHVAK  22  LAPTDYHLFLSL
TPZ-3   LFTDEKIFCI  78  WTFQQDGAPAHKHK  25  LNPMDYSVWSVL
TPZ-6   VTGDEKWVLY  74  LYFLHDNARPHVAK  22  LAPTDYHLFLSL
TPZ-7   VTGDEKWVLY  74  LYFLHDNARPHVAK  22  LAPTDYHLFLSL
TPZ-9   LFTDEKIFCI  78  WTFQQDGAPAHKHK  25  LNPMDYSVWSVL
TPZ-10  VFSDEKKFNL  75  FRFQQDNATIHVSN  22  LNPIEN-LWGIL
TPZ-11  VFSDEKKFNL  75  WQFQQDNASIHSAN  22  LNTIEN-MWASL
TPZ-13  LFTDEKIFCI  78  WTFQQDGAPAHKHK  25  LNPMDYSVWSVL
TPZ-14  VFSDEKKFNL  75  FRFQQDNATIHVSN  22  LNPIEN-LWGIL
TPZ-16  VIGDEKWVLY  74  LYFLHDNARPHVAK  22  LAPTDYHLFLSL
TPZ-18  VIGDEKWVLY  74  LYFLHDNARPPVAK  22  LAPTDYHLFLSL
TPZ-19  VFSDEKKFNL  75  FRFQQDNATIHVSN  22  LNPIEN-LWGIL
TPZ-20  VFSDEKKFNL  75  FRFQQDNATIHVSN  22  LNPIEN-LWGIL
TPZ-21  VFSDEKKFNL  75  FRFQQDNATIHVSN  22  LNPIEN-LWGIL
TPZ-22  VFSDEKKFNL  75  FRFQQDNATIHVSN  22  LNPIEN-LWGIL
TPZ-23  FAADETAVYL  94  RLLAWDAFRCHISD  23  VQAPDVYWNSPF
TPZ-24  VTGDEKWVLY  74  LYFLHDNARPHVAK  22  LAPTDYHLFLSL
TPZ-26  VFSDEKKFNL  75  FRFQQDNATIHVSN  22  LNPIEN-LWGIL
TPZ-30  VTGDEKWVLY  74  LYFLHDNARPHVAK  22  LAPTDYHLFLSL
TPZ-31  VTGDEKWVLY  74  LYFLHDNARPHVAK  22  LAPTDYHLFLSL
TPZ-34  VFSDEKKFNL  75  FRFQQDNATIHVSN  22  LNPIEN-LWGIL
TPZ-35  VTGDEKWVLY  74  LYFLHDNARPHVAK  22  LAPTDYHLFLSL
TPZ-37  VFSDEKKFNL  75  FRFQQDNATIHVSN  22  LNPIEN-LWGIL
TPZ-38  VFSDEKKFNL  75  FRFQQDNATIHVSN  22  LNPIEN-LWGIL
TPZ-39  VTGDEKWVLY  74  LYFLHDNARPHVAK  22  LAPTDYHLFLSL
TPZ-40  VTGDEKWVLY  74  LYFLHDNARPHVAK  22  LAPTDYHLFLSL
TPZ-43  FNADQCGIQK  100 MLMIVDSWTSFKDH  25  IQPLDVHFNGPW
TPZ-45  VTGDEKWVLY  74  LYFLHDNARPHVAK  22  LAPTDYHLFLSL
TPZ-49  VTGDEKWVLY  74  LYFLHDNARPHVAK  22  LAPTDYHLFLSL
TPZ-51  VTGDEKWVLY  74  LYFLHDNARPHVAK  22  LAPTDYHLFLSL
TPZ-52  VTGDEKWVLY  74  LYFLHDNARPHVAK  22  LAPTDYHLFLSL
TPZ-53  TTGDEKWVLY  77  LLLLHDNARPHTAL  22  LAPTDYHLFRSL
TPZ-55  VTGDEKWVLY  74  LYFLHDNARPHVAK  22  LAPTDYHLFLSL
TPZ-56  VTGDEKWVLY  74  LYFLHDNARPHVAK  22  LAPTDYHLFLSL
TPZ-63  VFSDEKKFNL  75  FRFQQDNATIHVSN  22  LNPIEN-LWGIL
TPZ-65  VFSDEKKFNL  75  FRFQQDNATIHVSN  22  LNPIEN-LWGIL
TPZ-66  VTGDEKWVLY  74  LYFLHDNARPHVAK  22  LAPTDYHLFLSL
TPZ-67  VFSDEKKFNL  75  FRFQQDNATIHVSN  22  LNPIEN-LWGIL
TPZ-68  LFTDEKIFCI  78  WTFQQDGAPAHKHK  25  LNPMDYSVWSVL
TPZ-70  IWSDESKFNL  78  FVFQQDNDPKHTSL  22  LNPIEH-LWEEL
TPZ-71  VTGDEKWVLY  74  LYFLHDNARPHVAK  22  LAPTDYHLFLSL
TPZ-75  VTGDEKWVLY  74  LYFLHDNARPHVAK  22  LAPTDYHLFLSL
TPZ-77  VTGDEKWVLY  74  LYFLHDNARPHVAK  22  LAPTDYHLFLSL
TPZ-78  VTGDEKWVLY  74  LYFLHDNARPHVAK  22  LAPTDYHLFLSL
TPZ-83  VTGDEKWVLY  74  LYFLHDNARPHVAK  22  LAPTDYHLFLSL
TPZ-84  VTGDEKWVLY  74  LYFLHDNARPHVAK  22  LAPTDYHLFLSL
TPZ-85  VTGDEKWVLY  74  LYFLHDNARPHVAK  22  LAPTDYHLFLSL
TPZ-87  VFSDEKKFNL  75  FRFQQDNATIHVSN  22  LNPIEN-LWGIL
TPZ-89  VTGDEKWVLY  74  LYFLHDNAMPHVAK  22  LAPTDYHLFLSL
TPZ-92  VTGDEKWVLY  74  LYFLHDNARPHVAK  22  LAPTDYHLFLSL
TPZ-94  VFSDEKKFNL  75  FRFQQDNATIHVSN  22  LNPIEN-LWGIL
TPZ-95  VTGDEKWVLY  74  LYFLHDNARPHVAK  22  LAPTDYHLFLSL
TPZ-96  VTGDEKWVLY  74  LYFLHDNARPHVAK  22  LAPTDYHLFLSL
TPZ-98  ITGDDKWVLY  63  LWLLYDNARPHTTL  22  LTPTDYHLFRTL
TPZ-99  VTGDEKWVLY  74  LYFLHDNARPHVAK  22  LAPTDYHLFLSL
```

Figure 3. *(Continued)*

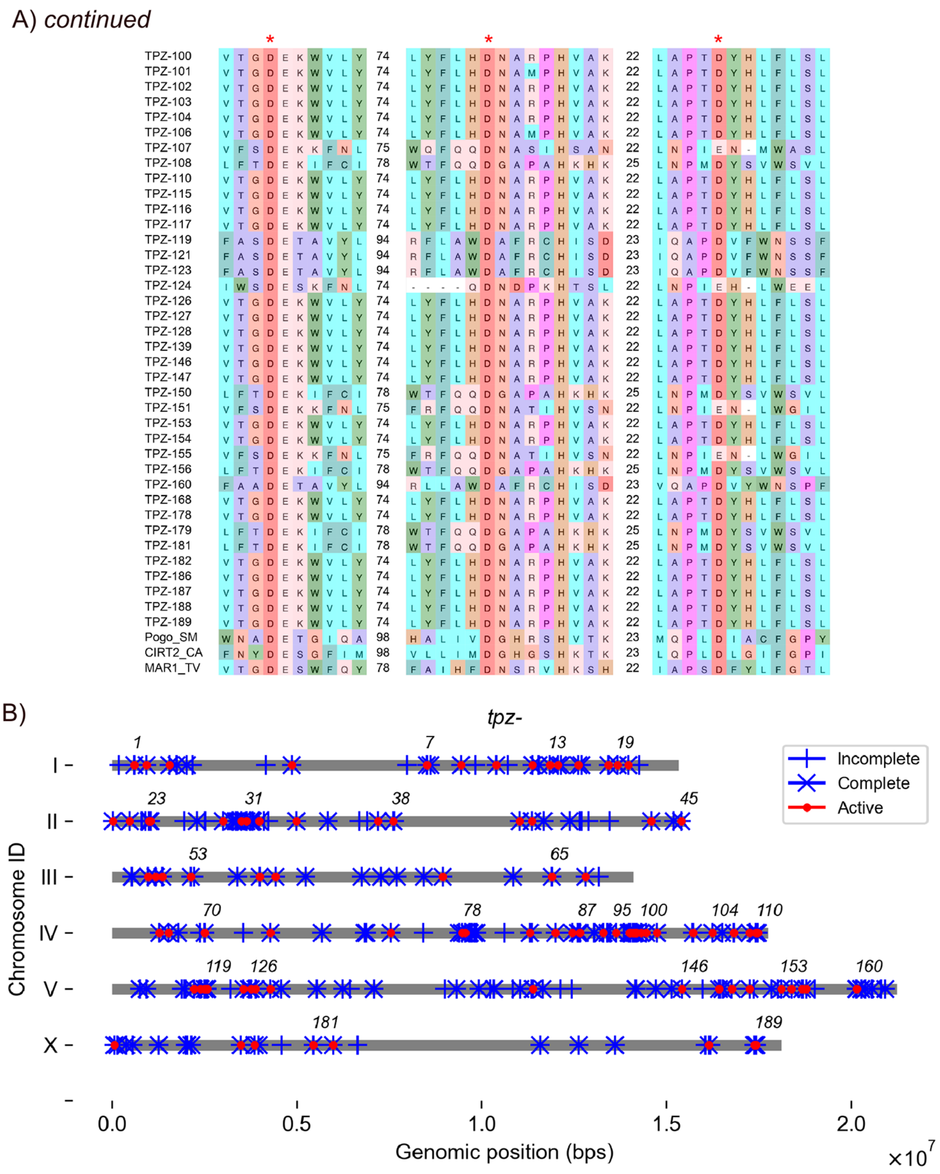


Figure 3. Potentially functional *tpzs* encoded in terminal inverted repeat (TIR) elements. (A) Alignment of catalytic core domains of TPZs with reference Transposases (TPs). Amino acid sequences of Pogo_SM, CIRT2_CA, and Mar1_TV as examples of Tc1/mariner class TPs were arbitrarily obtained from Yuan and Wessler.[65] Red asterisks indicate conserved residues used to identify potentially functional TDPGs. (B) Genomic positions of 94 potentially functional *tpzs* in the *Caenorhabditis elegans* genome. Gray lines represent chromosomes. Red circles indicate positions of potentially functional genes of *tpz-n*. Numbers indicate numbers of *tpz* genes. Vertical ticks and × marks indicate incomplete and complete ORFs, respectively, of *tpzs*.

short insertions, deletions, and duplications ranging from tens to thousands of base pairs, which are distributed in all chromosomes.[39] These differences could be due to polymorphism in the VC2010 strain or could be due to errors arising during sequencing and/or assembly of the N2 genome assembly. In the comparison between recent *de novo* assemblies of the N2 and Hawaiian genomes, indels with about 50 bases are widely detected across the genomes,[83] suggesting that such small indels may be inserted at a relatively rapid rate in evolution. We interpret that the discontinuity within the MP copies (Figure 5A and B) reflects the actual sequence present in the VC2010 strain. On the other hand, as mentioned in previous reports,[39] artifactual large structural variations in VC2010 could still remain even after careful correction in the VC2010 assembly. At this point, we remain agnostic as to whether the different chromosomal locations of the MP copies in VC2010 are actually present in the strain or reflect errors in the assembly process to obtain the VC2010 assembly.

To identify TDPGs encoded in these MP loci, ie, DNA polymerase B (DNA POLB) and IN, we applied SNAP and

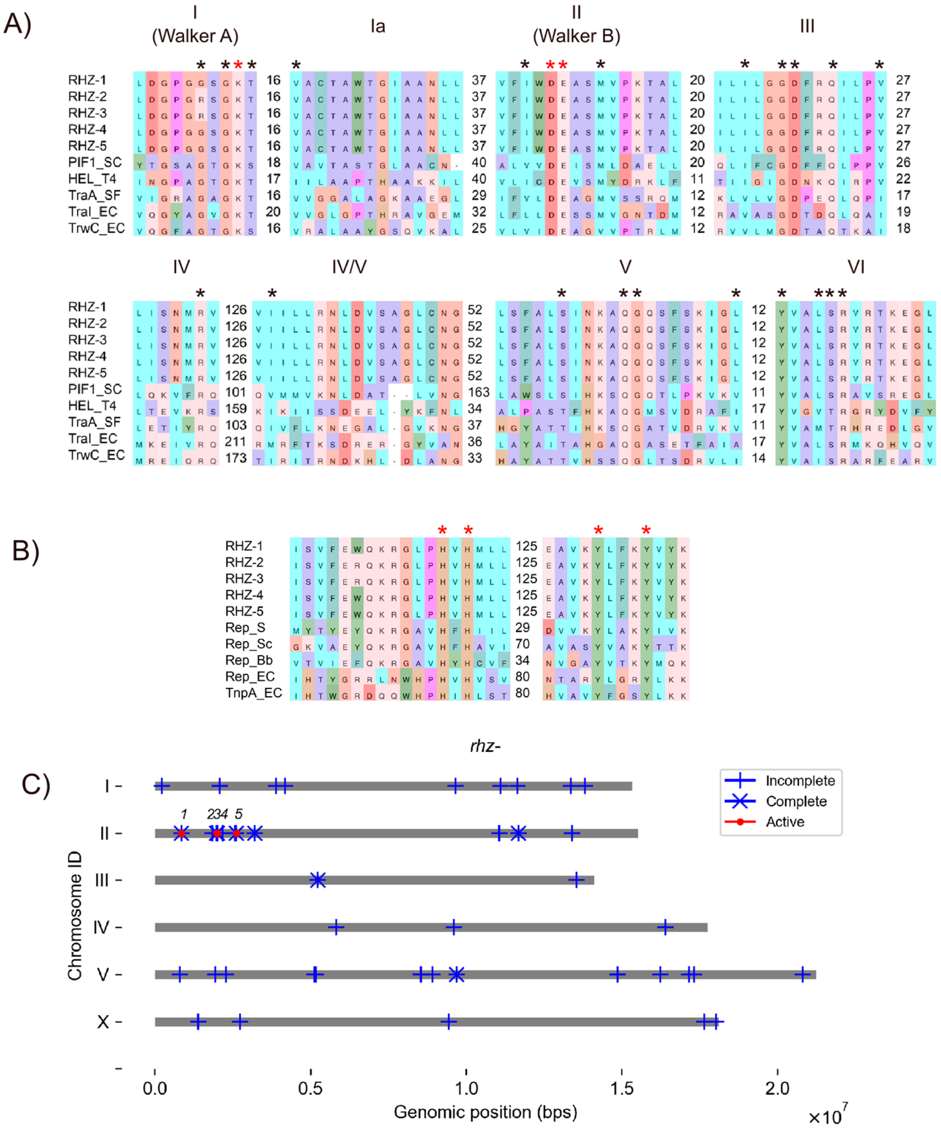


Figure 4. Potentially functional *rhz*s encoded in Helitrons. (A) Alignment of amino acid sequences of helicase (HEL) domains with reference HELs. PIF1 in *Saccharomyces cerevisiae* [P07271], HEL_T4 in Enterobacteria phage T4 [P32270], TraA *Sinorhizobium fredii* [P55418], TraI_EC *Escherichia coli* [P14565], TRWC *E. coli* [Q47673]. (B) Alignment of amino acid sequences of REP domains with reference REPs. Rep_Bb in *Brevibacillus borstelensis* [BAA07788.1], Rep plasmid pVT736-1 [AAC37125.1], Rep_Pf3P. phage Pf3 [AAA88392.1], Rep_EC IS91 [BCN22733.1], TnpA_EC IS91 TnpA *E. coli* [QIC00531.1]. Asterisks indicate conserved residues. Red asterisks indicate residues used for identifying potentially functional TDPGs. (C) Genomic positions of 5 potentially functional *rhz*s in the *Caenorhabditis elegans* genome. Gray lines represent chromosomes. Red circles indicate positions of potentially functional genes of *rhz-n*. Numbers indicate numbers of *rhz* genes. Vertical ticks and × marks indicate incomplete and complete ORFs, respectively, of *rhz*s.

DIAMOND to the 5 MP copies. Two DNA POLB-related genes (*polB_MP*) encoding 378 and 93 amino acids were located in MP Ia and MP X, respectively (Supplementary Table 7). Two IN genes (*inz_MP*) were located in MP Ib and MP Ic (Supplementary Table 8). Interestingly, a gene encoding the helix-turn-helix 48 (HTH48) domain-containing protein, which is conserved in some Transposases (TPs), is located in MP Xa (Supplementary Tables 9 and 10). Because TPs are functionally and evolutionarily related to INs,[84,85] we considered these 3 IN-related genes to be *inz_MP*s.

DNA POLB has 5 conserved motifs from Motifs I to V.[86] By aligning PolB_MP-1 (at MP Ib) and PolB_MP-2 (at MP Xa) with the reference DNA POLBs, we found that PolB_MP-1 lacked most of the N-terminal motifs from I to III and only had YnDTD conserved in Motif IV. In addition, PolB_MP-2 only had conservation of a short N-terminus fragment

Table 3. List of potentially functional *rhz*s.

GENE ID	CLASS	CHR	POSITION	HEL	REP
rhz-1	Helitron1_CERC/Helitron	II	(858688, 863243)	GKT V IDEM IGDQV R I SQGL YLSR	H H Y Y
rhz-2	Helitron1_CERC/Helitron	II	(1985427, 1990301)	GKT V IDEM IGDQV R I SQGL YLSR	H H Y Y
rhz-3	Helitron1_CERC/Helitron	II	(1994314, 1999188)	GKT V IDEM IGDQV R I SQGL YLSR	H H Y Y
rhz-4	Helitron1_CERC/Helitron	II	(2018435, 2023309)	GKT V IDEM IGDQV R I SQGL YLSR	H H Y Y
rhz-5	Helitron1_CERC/Helitron	II	(2619731, 2624240)	GKT V IDEM IGDQV R I SQGL YLSR	H H Y Y

Evolutionarily conserved amino acids within the Helicase ("Hel") and Endonuclease ("REP") domains within potentially functional RHZs. The consecutive amino acid single letters without spaces indicate conserved amino acids within the same domain. See amino acids indicated by asterisks in Figure 4A and B. A dash (-) indicates the absence of a conserved amino acid. The "Class" denotes the TE identity as determined by RepeatMasker analysis. Chromosome ID ("Chr") and genomic positions ("Position") are provided for each gene.

and did not exhibit conservation of Motifs I to V. Therefore, we considered that these PolB_MPs likely do not encode functional proteins. Next, by aligning INZ_MP-1 (at MP Ib), INZ_MP-2 (at MP Ic), and INZ_MP-3 (MP Xa) with reference INs, we found that INZ_MP-1 and INZ_MP-2 encoded glutamic acid triads that align with the reference DDE triad (black asterisks in Figure 5C). Based on this potential functional conservation, we do not reject the possibility that *inz_MP*-1 and *inz_MP*-2 encode functional genes. INZ_MP-3 did not align with the reference DDE triads, whereas INZ_MP-3 aligned with the DDD/E triad in the Tc1 family TPs (red asterisks in Figure 5C). We thus also considered *inz_MP-3* to be a potentially functional gene.

The IN encoded in a MP copy has been identified as a cellular IN (c-Integrase) that is homologous to the retrotransposon IN.[22] In addition, the IN in MP is highly homologous to the TP encoded in ginger DNA transposon.[87] Maverick/Polintons has been proposed to be involved in gene transfer between eukaryotic DNA mobile elements (dsDNA viruses, adenoviruses, small ssDNA viruses, Mavirus-like virophages, icosahedral viruses).[88-90] Considering that *inz_MP-3* at MP Xa encoded an HTH48 domain, and that the DDD motif was homologous to the Tc family of DNA TPs, we note the possibility of evolutionary interactions between the Tc family of DNA transposons and MP in *C. elegans*.

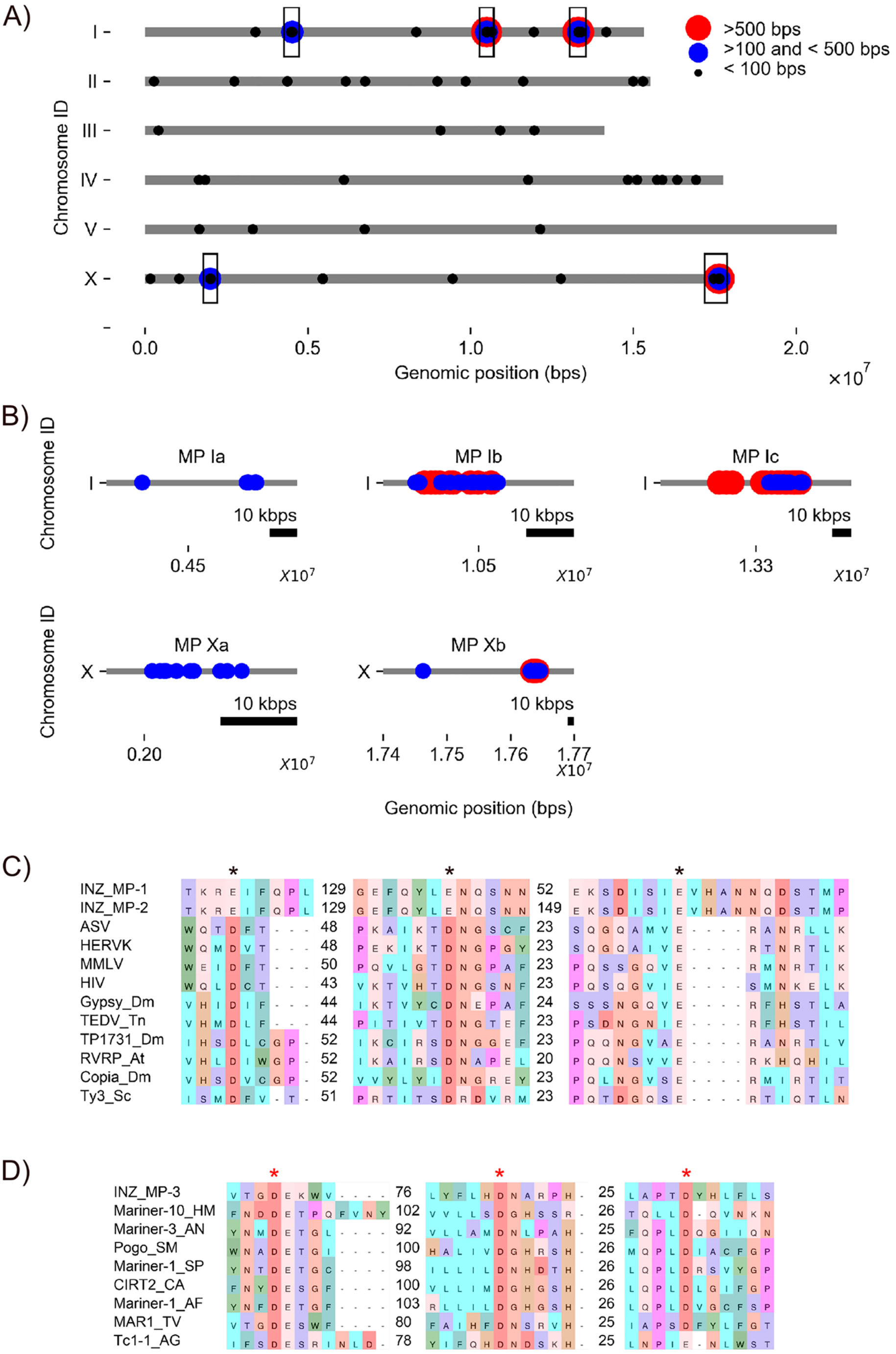
TDPGs in DIRS, PLE, and Crypton elements in C. elegans

Our RepeatMasker analysis did not identify copies of DIRS or Crypton. A previous homology-based search of 34 nematode species to identify Tyr-REC genes that mediate transposition of DIRS and Crypton identified an incomplete cDNA for a Tyr-REC gene in chromosome II of *C. elegans*.[91] We applied SNAP to a genomic region of this incomplete cDNA with 20 kbp 5′ flanking and 20 kbp 3′ flanking genomic regions in VC2010, but did not find any complete ORF. From this result and previous studies, we conclude that *C. elegans* may not have a functional Tyr-REC gene. Finally, consistent with a previous report,[92] our analysis on VC2010 did not identify any PLE copies.

Discussion

Extensive bioinformatic studies on TEs have led to the identification of novel TEs, the inferred mechanisms of transposition from enzymes coded on TEs, and through phylogenetics analysis, the clarification of the evolutionary processes of TEs and the genome.[6,7,26] Despite such studies, the physiological functions of TEs, such as their roles in development, and aging, as well as evolution, are largely unverified experimentally. In this report, we searched for TDPGs in all the representative TE orders in the latest genome assembly of *C. elegans* (VC2010). According to our RepeatMasker analysis, more than 50 000 TE loci exist in the latest *C. elegans* genome assembly (Figure 6A and Supplementary Table 1). These loci encode 428 complete ORFs, 66.8% of which (285 genes) are TDPGs. Among them, we identified 142 potentially functional genes, including 8 *rtz_LTR*s, 7 *inz_LTR*s, 28 *rtz_LINE*s, 94 *tpz*s, and 5 *rhz*s (Figure 6B). In addition, our manual analysis identified 3 potentially functional *inz_MP*s (Figure 6B). In total, we identified 145 potentially functional TDPGs.

The preponderance of total TE loci relative to those retaining potential for mobility is well noted but has rarely been quantified. The best studied example is the retrotransposon L1, a member of the well-studied LINE order. The human genome has more than 500 000 loci corresponding to L1, accounting for 17% of the human genome.[93] However, bioinformatic analysis has identified only 146 copies of full-length L1 in the human genome, and only 107 of these L1 copies have conserved intact RT genes.[94] *In vitro* experimental tests of the mobility of L1 copies showed that fewer than 100 copies of L1 were active, with 6 L1 copies accounting for most of the activity of L1 in the human genome.[95] Similarly, cancer cell genome analysis and genome comparison in human population showed that limited number of L1 loci comprise the bulk of mobile activity of L1.[96-99] These results indicate that a limited number



```
INZ_MP-1      TKREIFQPL 129  GEFQYLENQSNN 52   EKSDISIEVHANNQDSTMP
INZ_MP-2      TKREIFQPL 129  GEFQYLENQSNN 149  EKSDISIEVHANNQDSTMP
ASV           WQTDFT---  48  PKAIKTDNGSCF 23   SQGQAMVE----RANRLLK
HERVK         WQMDVT---  48  PEKIKTDNGPGY 23   SQGQAIVE----RTNRTLK
MMLV          WEIDFT---  50  PQVLGTDNGPAF 23   PQSSGQVE----RMNRTIK
HIV           WQLDCT---  43  VKTVHTDNGSNF 23   PQSQGVIE----SMNKELK
Gypsy_Dm      VHIDIF---  44  IKTVYCDNEPAF 24   SSSNGQVE----RFHSTLA
TEDV_Tn       VHMDLF---  44  PITIVTDNGTEF 23   PSDNGNIE----RFHSTIL
TP1731_Dm     IHSDLCGP-  52  IKCIRSDNGGEF 23   PQQNGVAE----RANRTLV
RVRP_At       VHLDIWGP-  52  IKAIRSDNAPEL 20   PQQNSVVE----RKHQHIL
Copia_Dm      VHSDVCGP-  52  VVYLYIDNGREY 23   PQLNGVSE----RMIRTIT
Ty3_Sc        ISMDFV-T-  51  PRTITSDRDVRM 23   PQTDGQSE----RTIQTLN
```

```
INZ_MP-3      VTGDEKWV----  76  LYFLHDNARPH-  25  LAPTDYHLFLS
Mariner-10_HM FNDDETPQFVNY 102  VVLLSDGHSSR-  26  TQLLD-QVNKN
Mariner-3_AN  YNMDETGL----  92  VLLAMDNLPAH-  25  FQPLDQGIIQN
Pogo_SM       WNADETGI---- 100  HALIVDGHRSH-  26  MQPLDIACFGP
Mariner-1_SP  YNTDETGC----  98  ILLILDNHDTH-  26  LQPLDRSVYGP
CIRT2_CA      FNYDESGF---- 100  VLLIMDGHGSH-  26  LQPLDLGIFGP
Mariner-1_AF  YNFDETGF---- 103  RLLILDGHGSH-  26  LQPLDVGCFSP
MAR1_TV       VTGDESWF----  80  FAIHFDNSRVH-  25  IAPSDFYLFGT
Tc1-1_AG      IFSDESRINLD-  78  YIFQHDNDSKH-  25  LNPIE-NLWST
```

Figure 5. Maverick/Polintons (MPs) in the *Caenorhabditis elegans* genome. (A) Genomic positions of homologous regions with AL110478.1. Five rectangular regions are as follows: in chromosome I: MP Ia, 3543 bp region from 0.4521550 × 10^7 to 0.4525093 × 10^7 bps; MP Ib, 17 276 bp region from 1.0486778 × 10^7 bps to 1.0504054 × 10^7 bps, and MP Ic, 43 817 bp region from 1.3280931 × 10^7 bps to 1.3324748 × 10^7 bps; and in chromosome X: MP Xa, 11 758 bp region from 0.2000999 × 10^7 bps to 0.2012757 × 10^7 bps, and MP Xb, 183 668 bp region from 1.7462913 × 10^7 bps to 1.7646581 × 10^7 bps. (B) Scattered distribution of homologous DNA regions with AL110478.1 in each of 5 MP copies. Red dots indicate homologous regions > 500 bps. Blue dots indicate homologous regions of < 500 bps and > 100 bps. (C) Alignment of catalytic core domains of INZ_MP-1 and INZ_MP-2 with reference Integrases (INs). Abbreviations: ASV: IN in avian sarcoma virus [1ASU_A], HERVK: Pol protein in human endogenous retrovirus K [CAA76885], MMLV: p46 IN Moloney murine leukemia virus (MoMLV) [NP_955592.1], HIV: IN in HIV-1 [1BIZ_A], Gypsy_Dm: IN, Gypsy endogenous retrovirus in *Drosophila melanogaster* [CAB69645], TEDV_Tm: ORFB in TED virus in *Trichoplusia ni* [YP_009507248], TP1731: Pol polyprotein in transposon_1731 in *D. melanogaster* [S00954], RVRP_At: retrovirus-related like Polyprotein in *Arabidopsis thaliana* [CAB78488_1], Copia_Dm: Gag-Int-Pol protein in COPIA in *D. melanogaster* [P04146], Ty-3_Sc: Gag-Pol polyprotein in Ty3-G in *Saccharomyces cerevisiae* [GFP69998.1]. (D) Alignment of catalytic core domains of INZ_MP-3 with reference Transposases (TPs). Amino acid sequences of Mariner-10_HM, Mariner-3_AN, Pogo_SM, Mariner-1_SP, CIRT2_CA, Mariner-1_AF, MAR1_TV, and Tc1-1_AG as examples of Tc1/mariner class TPs were arbitrarily obtained from Yuan and Wessler.[65] Asterisks indicate conserved amino acids in catalytic core.

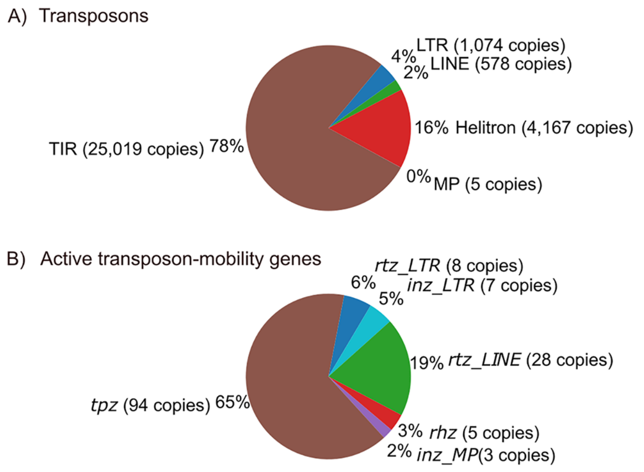


Figure 6. Summary of potentially functional TDPGs in the *Caenorhabditis elegans* genome. (A) The percentage of copy numbers of LTR element (blue), LINE (green), Helitron (red), MP (purple), and TIR element (brown) in the total copy number of TEs in the *C. elegans* genome. (B) The percentage of potentially functional TDPGs; 8 *rtz_LTR* (blue) and 7 *inz_LTR* (light blue), 28 *rtz_LINE* (green), 5 *rhz* (red), 3 *inz_MP* (purple), and 94 *tpz* (brown) in total 145 potentially functional TDPGs identified in this study.

of L1 copies encode functional TDPGs, which is consistent with our finding of the stark contrast between the number of potentially functional TDPGs and the total number of TEs.

Critical amino acids at the catalytic cores of TDPGs have been studied in LTR elements,[27-29] LINEs,[30-32] and TIR elements[33] in *C. elegans*. We compared the 145 potentially functional TDPGs identified in this study to those identified previously. In previous studies of LTR elements, researchers identified 24[27] or 10[28] full-length copies among 124 or 62 LTR element copies, respectively, in *C. elegans*. Another study reported 17 *rtz_LTR*s with conserved Asp residues in Motifs A and C, and 15 *inz_LTR*s with a conserved DDE triad.[29] We identified 8 potentially functional *rtz_LTR* and 7 potentially functional *inz_LTR* genes. Thus, we found fewer potentially functional *rtz_LTR*s and *inz_LTR*s in this study than were identified by previous studies.[27-29] Differences in the genome assembly used in Ganko et al[27] and Bowen and McDonald,[29] as discussed in Yoshimura et al,[39] might contribute to this discrepancy. On the other hand, using the same VC2010 assembly as us, Kanzaki et al[28] identified 10 LTR elements; we identified nearly the same number but slightly fewer TDPGs. A detailed analysis of the coding sequence of the catalytic core of enzymes in the 10 full-length LTR elements[28] may provide consilience with the number of potentially functional LTR element genes identified here. Notably, TDPGs in the most-studied *C. elegans* LTR element, Cer1 on chromosome III,[27,100] were found on our list of potentially functional genes (i.e. *rtz_LTR-5* and *inz_LTR-5*; Figure 1C). Cer1 is both biologically active and mobile[101-103] in recent evolutionary history, based on a comparison of natural isolates of *C. elegans*.[104] For LINE retrotransposons, among more than 1000 copies of LINE,[30] 6 copies of the RTE LINE suborder[31] and 17 copies of the T1/CR1 LINE suborder[32] encode the *rtz_LINE* that conserves the Asp residues in Motifs A and C. We identified a similar number of LINE copies (618 copies; Supplementary Table 1). Among them, 28 *rtz_LINE*s preserved potentially functional RT domains, which is more than total number of potentially functional *rtz_LINE*s reported previously. For TIR elements, a previous report showed that 61 *tpzs* had the conserved DDD/E motifs in 127 copies of Tc/mariner family TIR element.[33] We identified 94 potentially functional *tpzs* in 189 copies of the Tc/mariner family TIR element, which is a larger number than was reported previously. For Helitron DNA transposon, we found that 1.65% of the *C. elegans* genome was occupied with Helitron copies (Supplementary Table 1), similar to previous reports (~2%).[18] There was no further amino acid sequence analysis about *rhzs*.

In addition, we investigated the latest version of Wormbase (WS292). For Tc1/mariner, among its 636 copies, 129 *tpzs* were found, with 112 being potentially functional *tpzs*. Regarding LINE, among its 62 copies, 9 out of 10 potentially functional *rtz_LINE*s were found. For LTR element, among its 15 copies with 4 ORFs registered, none showed potentially functional *rtz_LTR* and *inz_LTR*. In the case of MP, among its 12 copies, 3 *polB_MP*s (C33E10.6, Y106G6G.5, Y26D4A.9) were registered, but *inz_MP* was not found. Among these PolB_MP, only Y26D4A.9 preserved motif IV out of 5 conserved motifs, and thereby these *pol_MP*s were not considered potentially functional genes. Likely due to discrepancies in the genome sequences, corresponding ORFs for these 3 *polB*s were not found in the VC2010 genome assembly. Helitron, DIRS, Crypton, and PLE were not registered. Taken together, our catalog of potentially functional TDPGs includes an equivalent or, for some TEs, greater number of genes compared with the number identified in previous reports, and reflects various lines of available biological and bioinformatic evidence, supporting the validity of our annotation.

In human and mouse genomes, L1 is the most abundant among TE classes, with 145 copies in human and 2811 copies in *Mus musculus* being conserved at full length.[94,105,106] In *D. melanogaster*, *Saccharomyces cerevisiae*, and *Arabidopsis thaliana*,[107] LTR element is the most abundant class, with 325 copies in *D. melanogaster*[108] and 51 copies in *S. cerevisiae*[109] being conserved at full length. Long terminal repeat elements are only the TE encoded in *Schizosaccharomyces pombe*, with 13 copies being conserved at full length.[110] DNA transposons are the most abundant in *Danio rerio* (zebrafish), with 2.3 million copies, but it is unknown how many of these copies are active.[111] Similarly, it is unknown how many of the 286 LTR elements in *A. thaliana* are active.[107] In our study, TIR elements were the most abundant transposon class (20 852 copies; Supplementary Table 1) in *C. elegans*, encoding 94 copies of potentially functional *tpz*. Our identification of 145 potentially functional TDPGs indicates that *C. elegans* encodes the fewest number of potentially functional TDPGs among conventional metazoan model organisms. We propose that *C. elegans* can be a useful model metazoan to study the conserved physiological, pathological, and evolutionary roles of transposon mobility.

Conclusions

Our TDPG catalog, which contains 145 potentially functional TDPGs encoded in LTR elements, LINEs, TIR elements, Helitrons, and Mavericks/Polintons, serves as a source of information for conducting RNA interference (RNAi) experiments to systematically manipulate the mobility of both autonomous and non-autonomous TEs. By designing primer sets to specifically amplify certain genome regions encoding a TDPG, RNAi experiments can simultaneously manipulate the mobility of TEs that rely on the target TDPGs, other TEs that rely on homologous copies of the TDPG, and non-autonomous TEs that depend on these TDPGs. RNAi targeting a TDPG in a homologous gene cluster of *rtz_LINE* and *tpz* (Supplementary Figures 1 and 2) can more efficiently and simultaneously affect the mobility of multiple TEs. Our TDPG catalog could potentially promote the study of the physiological functions of TE mobility.

Limitation

Our gene catalog provides substantial advantages over its predecessors, such as Wormbase for conducting experimental studies but is imperfect. The TDPGs in our catalog were curated based on amino acid sequences within the catalytic cores; some of these genes might be truncated in other regions, such as the N- or C-terminus or within inter-functional domains. The TDPGs in our catalog might lack DNA elements necessary for proper gene transcription. Due to mutations in regions responsible for translational regulation,[112] it is possible that genes may not produce functional transcripts or proteins. In addition, our current TDPG search scheme does not detect genes located in the genome regions that RepeatMasker did not recognize as TEs. Thus, the possibility remains that there are still unidentified TDPGs. To address these shortcomings, it will be necessary to combine multiple algorithms for TE and gene identification across different versions of wild-type genome assemblies, based on continuously updated reference libraries for TEs, and to take the union of the TDPGs identified through these bioinformatics searches. In addition, integrating *in vivo* transcript information obtained from Expressed Sequence tags and RNA-seq data could further refine the gene catalog. Experimental evolution and re-sequencing organisms after targeting the current set of TDPGs would provide an orthogonal approach to evaluate whether this catalog is exhaustive. These expanded analyses could provide a more rigorous basis for experimental manipulation of TE mobility *in vivo*.

Acknowledgements

We thank Dr Damien Farrell at University College Dublin for the sequence alignment viewer. We also would like to thank Thomas J. LaRocca at Colorado State University for his helpful comments and suggestions on manuscript. Finally, we extend our gratitude to WormBase (www.wormbase.org) for a valuable help for our analysis.

Author Contributions

Y.A. and N.P. conceptualized the study. Y.A. and P.J. performed the data analysis. Y.S. provided supervision, contributed to the interpretation of the results, and critically reviewed the manuscript. Y.A. drafted the manuscript.

ORCID iD

Yukinobu Arata https://orcid.org/0000-0002-8687-4678

SUPPLEMENTAL MATERIAL

Supplemental materials for this article are available online.

REFERENCES

1. Sessegolo C, Burlet N, Haudry A. Strong phylogenetic inertia on genome size and transposable element content among 26 species of flies. *Biol Lett*. 2016;12:20160407. doi:10.1098/rsbl.2016.0407
2. Huang CR, Burns KH, Boeke JD. Active transposition in genomes. *Annu Rev Genet*. 2012;46:651-675. doi:10.1146/annurev-genet-110711-155616
3. Deniz Frost ÖJM, Branco MR. Regulation of transposable elements by DNA modifications. *Nat Rev Genet*. 2019;20:417-431. doi:10.1038/s41576-019-0106-6
4. Gorbunova V, Seluanov A, Mita P, et al. The role of retrotransposable elements in ageing and age-associated diseases. *Nature*. 2021;596:43-53. doi:10.1038/s41586-021-03542-y
5. Kapitonov VV, Jurka J. A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat Rev Genet*. 2008;9:411-412. doi:10.1038/nrg2165-c1
6. Wells JN, Feschotte C. A field guide to eukaryotic transposable elements. *Annu Rev Genet*. 2020;54:539-561. doi:10.1146/annurev-genet-040620-022145
7. Wicker T, Sabot F, Hua-Van A, et al. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet*. 2007;8:973-982. doi:10.1038/nrg2165
8. Cappello J, Handelsman K, Lodish HF. Sequence of Dictyostelium DIRS-1: an apparent retrotransposon with inverted terminal repeats and an internal circle junction sequence. *Cell*. 1985;43:105-115. doi:10.1016/0092-8674(85)90016-9
9. Goodwin TJ, Poulter RT. The DIRS1 group of retrotransposons. *Mol Biol Evol*. 2001;18:2067-2082. doi:10.1093/oxfordjournals.molbev.a003748
10. Goodwin TJ, Poulter RT. A new group of tyrosine recombinase-encoding retrotransposons. *Mol Biol Evol*. 2004;21:746-759. doi:10.1093/molbev/msh072
11. Poulter RTM, Butler MI. Tyrosine recombinase retrotransposons and transposons. *Microbiol Spectr*. 2015;3:MDNA3-0036-2014. doi:10.1128/microbiolspec.mdna3-0036-2014
12. Pyatkov KI, Arkhipova IR, Malkova NV, Finnegan DJ, Evgen'ev MB. Reverse tanscriptase and endonuclease activities encoded by Penelope-like retroelements. *Proc Natl Acad Sci U S A*. 2004;101:14719-14724. doi:10.1073/pnas.0406281101
13. Hickman AB, Dyda F. DNA transposition at work. *Chem Rev*. 2016;116:12758-12784. doi:10.1021/acs.chemrev.6b00003
14. Hickman AB, Perez ZN, Zhou L, et al. Molecular architecture of a eukaryotic DNA transposase. *Nat Struct Mol Biol*. 2005;12:715-721. doi:10.1038/nsmb970
15. Nesmelova IV, Hackett PB. DDE transposases: structural similarity and diversity. *Adv Drug Deliv Rev*. 2010;62:1187-1195. doi:10.1016/j.addr.2010.06.006
16. Chandler M, de la Cruz F, Dyda F, Hickman AB, Moncalian G, Ton-Hoang B. Breaking and joining single-stranded DNA: the HUH endonuclease superfamily. *Nat Rev Microbiol*. 2013;11:525-538. doi:10.1038/nrmicro3067
17. Grabundzija I, Messing SA, Thomas J, et al. A Helitron transposon reconstructed from bats reveals a novel mechanism of genome shuffling in eukaryotes. *Nat Commun*. 2016;7:10716. doi:10.1038/ncomms10716

18. Kapitonov VV, Jurka J. Rolling-circle transposons in eukaryotes. *Proc Natl Acad Sci U S A*. 2001;98:8715-8719. doi:10.1073/pnas.151269298
19. Kapitonov VV, Jurka J. Helitrons on a roll: eukaryotic rolling-circle transposons. *Trends Genet*. 2007;23:521-529. doi:10.1016/j.tig.2007.08.004
20. Thomas J, Pritham EJ. Helitrons, the eukaryotic rolling-circle transposable elements. *Microbiol Spectr*. 2015;3. doi:10.1128/microbiolspec.mdna3-0049-2014
21. Feschotte C, Pritham EJ. Non-mammalian c-integrases are encoded by giant transposable elements. *Trends Genet*. 2005;21:551-552. doi:10.1016/j.tig.2005.07.007
22. Gao X, Voytas DF. A eukaryotic gene family related to retroelement integrases. *Trends Genet*. 2005;21:133-137. doi:10.1016/j.tig.2005.01.006
23. Kapitonov VV, Jurka J. Self-synthesizing DNA transposons in eukaryotes. *Proc Natl Acad Sci U S A*. 2006;103:4540-4545. doi:10.1073/pnas.0600833103
24. Feschotte C, Pritham EJ. DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet*. 2007;41:331-368. doi:10.1146/annurev.genet.40.110405.090448
25. Goodwin TJD, Butler MI, Poulter RTM. Cryptons: a group of tyrosine-recombinase-encoding DNA transposons from pathogenic fungi. *Microbiology*. 2003;149:3099-3109. doi:10.1099/mic.0.26529-0
26. Ivancevic AM, Kortschak RD, Bertozzi T, Adelson DL. LINEs between species: evolutionary dynamics of LINE-1 retrotransposons across the eukaryotic tree of life. *Genome Biol Evol*. 2016;8:3301-3322. doi:10.1093/gbe/evw243
27. Ganko EW, Fielman KT, McDonald JF. Evolutionary history of Cer elements and their impact on the *C. elegans* genome. *Genome Res*. 2001;11:2066-2074. doi:10.1101/gr.196201
28. Kanzaki N, Tsai IJ, Tanaka R, et al. Biology and genome of a newly discovered sibling species of *Caenorhabditis elegans*. *Nat Commun*. 2018;9:3216. doi:10.1038/s41467-018-05712-5
29. Bowen NJ, McDonald JF. Genomic analysis of *Caenorhabditis elegans* reveals ancient families of retroviral-like elements. *Genome Res*. 1999;9:924-935. doi:10.1101/gr.9.10.924
30. Bessereau JL. Transposons in *C. elegans*. *WormBook*. 2006. doi:10.1895/wormbook.1.70.1
31. Youngman S, Van Luenen HG, Plasterk RH. Rte-1, a retrotransposon-like element in *Caenorhabditis elegans*. *FEBS Lett*. 1996;380:1-7. doi:10.1016/0014-5793(95)01525-6
32. Marín I, Plata-Rengifo P, Labrador M, Fontdevila A. Evolutionary relationships among the members of an ancient class of non-LTR retrotransposons found in the nematode *Caenorhabditis elegans*. *Mol Biol Evol*. 1998;15:1390-1402. doi:10.1093/oxfordjournals.molbev.a025867
33. Fischer SE, Wienholds E, Plasterk RH. Continuous exchange of sequence information between dispersed Tc1 transposons in the *Caenorhabditis elegans* genome. *Genetics*. 2003;164:127-134. doi:10.1093/genetics/164.1.127
34. Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob DNA*. 2021;12:1-14. doi:10.1186/s13100-020-00230-y
35. Korf I. Gene finding in novel genomes. *BMC Bioinformatics*. 2004;5:59. doi:10.1186/1471-2105-5-59
36. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*. 2014;12:59-60. doi:10.1038/nmeth.3176
37. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013;30:772-780. doi:10.1093/molbev/mst010
38. Katoh K, Misawa K, Kuma KI, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. 2002;30:3059-3066. doi:10.1093/nar/gkf436
39. Yoshimura J, Ichikawa K, Shoura MJ, et al. Recompleting the *Caenorhabditis elegans* genome. *Genome Res*. 2019;29:1009-1022. doi:10.1101/gr.244830.118
40. Laricchia KM, Zdraljevic S, Cook DE, Andersen EC. Natural variation in the distribution and abundance of transposable elements across the *Caenorhabditis elegans* species. *Mol Biol Evol*. 2017;34:2187-2202. doi:10.1093/molbev/msx155
41. Stein LD, Bao Z, Blasiar D, et al. The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLoS Biol*. 2003;1:E45. doi:10.1371/journal.pbio.0000045
42. Poch O, Sauvaget I, Delarue M, Tordo N. Identification of four conserved motifs among the RNA-dependent polymerase encoding elements. *EMBO J*. 1989;8:3867-3874. doi:10.1002/j.1460-2075.1989.tb08565.x
43. Xiong Y, Eickbush TH. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J*. 1990;9:3353-3362. doi:10.1002/j.1460-2075.1990.tb07536.x
44. Delarue M, Poch O, Tordo N, Moras D, Argos P. An attempt to unify the structure of polymerases. *Protein Eng*. 1990;3:461-467. doi:10.1093/protein/3.6.461
45. Steitz TA. A mechanism for all polymerases. *Nature*. 1998;391:231-232. doi:10.1038/34542
46. Le Grice SFJ. Human immunodeficiency virus reverse transcriptase: 25 years of research, drug discovery, and promise. *J Biol Chem*. 2012;287:40850-40857. doi:10.1074/jbc.R112.389056
47. Jacobo-Molina A, Ding J, Nanni RG, et al. Crystal structure of human immunodeficiency virus type 1 reverse transcriptase complexed with double-stranded DNA at 3.0 Å resolution shows bent DNA. *Proc Natl Acad Sci U S A*. 1993;90:6320-6324. doi:10.1073/pnas.90.13.6320
48. Larder BA, Purifoy DJM, Powell KL, Darby G. Site-specific mutagenesis of AIDS virus reverse transcriptase. *Nature*. 1987;327:716-717. doi:10.1038/327716a0
49. Le Grice SF, Naas T, Wohlgensinger B, Schatz O. Subunit-selective mutagenesis indicates minimal polymerase activity in heterodimer-associated p51 HIV-1 reverse transcriptase. *EMBO J*. 1991;10:3905-3911. doi:10.1002/j.1460-2075.1991.tb04960.x
50. Boyer PL, Ferris AL, Hughes SH. Cassette mutagenesis of the reverse transcriptase of human immunodeficiency virus type 1. *J Virol*. 1992;66:1031-1039. Accessed November 25, 2024. https://journals.asm.org/journal/jvi
51. Kaushik N, Rege N, Yadav NS, Sarafianos SG, Modak MJ, Pandey VN. Biochemical analysis of catalytically crucial aspartate mutants of human immunodeficiency virus type 1 reverse transcriptase. *Biochemistry*. 1996;35:11536-11546. Accessed November 25, 2024. https://pubs.acs.org/sharingguidelines
52. Smith RA, Anderson DJ, Preston BD. Hypersusceptibility to substrate analogs conferred by mutations in human immunodeficiency virus type 1 reverse transcriptase. *J Virol*. 2006;80:7169-7178. doi:10.1128/jvi.00322-06
53. Mitchell M, Gillis A, Futahashi M, Fujiwara H, Skordalakes E. Structural basis for telomerase catalytic subunit TERT binding to RNA template and telomeric DNA. *Nat Struct Mol Biol*. 2010;17:513-518. doi:10.1038/nsmb.1777
54. Barik S, Rud EW, Luk D, Banerjee AK, Kang CY. Nucleotide sequence analysis of the L gene of vesicular stomatitis virus (New Jersey serotype): identification of conserved domains in L proteins of nonsegmented negative-strand RNA viruses. *Virology*. 1990;175:332-337.
55. Castro C, Smidansky ED, Arnold JJ, et al. Nucleic acid polymerases use a general acid for nucleotidyl transfer. *Nat Struct Mol Biol*. 2009;16:212-218. doi:10.1038/nsmb.1540
56. Canard B, Chowdhury K, Sarfati R, Doublié S, Richardson CC. The motif D loop of human immunodeficiency virus type 1 reverse transcriptase is critical for nucleoside 5-triphosphate selectivity. *J Biol Chem*. 1999;274:35768-35776. Accessed November 25, 2024. http://www.jbc.org
57. Engelman A, Craigie R. Identification of conserved amino acid residues critical for human immunodeficiency virus type 1 integrase function *in vitro*. *J Virol*. 1992;66:6361-6369. doi:10.1128/jvi.66.11.6361-6369.1992
58. Van Gent DC, Groeneger AAMO, Plasterk RHA. Mutational analysis of the integrase protein of human immunodeficiency virus type 2. *Proc Natl Acad Sci U S A*. 1992;89:9598-9602. doi:10.1073/pnas.89.20.9598
59. Kulkosky J, Jones KS, Katz RA, Mack JP, Skalka AM. Residues critical for retroviral integrative recombination in a region that is highly conserved among retroviral/retrotransposon integrases and bacterial insertion sequence transposases. *Mol Cell Biol*. 1992;12:2331-2338. doi:10.1128/mcb.12.5.2331-2338.1992
60. Hare S, Maertens GN, Cherepanov P. 3′-Processing and strand transfer catalysed by retroviral integrase in crystallo. *EMBO J*. 2012;31:3020-3028. doi:10.1038/emboj.2012.118
61. Maertens GN, Engelman AN, Cherepanov P. Structure and function of retroviral integrase. *Nat Rev Microbiol*. 2021;20:20-34. doi:10.1038/s41579-021-00586-9
62. Drelich M, Wilhelm R, Mous J. Identification of amino acid residues critical for endonuclease and integration activities of HIV-1 IN protein *in vitro*. *Virology*. 1992;188:459-468.
63. Kines KJ, Sokolowski M, deHaro DL, et al. The endonuclease domain of the LINE-1 ORF2 protein can tolerate multiple mutations. *Mob DNA*. 2016;7:8. doi:10.1186/s13100-016-0064-x
64. Morrish TA, Gilbert N, Myers JS, et al. DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nat Genet*. 2002;31:159-165. doi:10.1038/ng898
65. Yuan YW, Wessler SR. The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. *Proc Natl Acad Sci U S A*. 2011;108:7884-7889. doi:10.1073/pnas.1104208108
66. Davies DR, Goryshin IY, Reznikoff WS, Rayment I. Three-dimensional structure of the tn5 synaptic complex transposition intermediate. *Science*. 2000;289:77-85. doi:10.1126/science.289.5476.77
67. Lovell S, Goryshin IY, Reznikoff WR, Rayment I. Two-metal active site binding of a Tn5 transposase synaptic complex. *Nat Struct Biol*. 2002;9:278-281. doi:10.1038/nsb778
68. Richardson JM, Colloms SD, Finnegan DJ, Walkinshaw MD. Molecular architecture of the mos1 paired-end complex: the structural basis of DNA transposition in a eukaryote. *Cell*. 2009;138:1096-1108. doi:10.1016/j.cell.2009.07.012
69. Bolland S, Kleckner N. The three chemical steps of Tn10/IS10 transposition involve repeated utilization of a single active site. *Cell*. 1996;84:223-233. doi:10.1016/S0092-8674(00)80977-0

70. Naumann TA, Reznikoff WS. Tn5 transposase active site mutants. *J Biol Chem*. 2002;277:17623-17629. doi:10.1074/jbc.M200742200
71. Hall MC, Matson SW. Helicase motifs: the engine that powers DNA unwinding. *Mol Microbiol*. 1999;34:867-877. doi:10.1046/j.1365-2958.1999.01659.x
72. Walker JE, Saraste M, Runswick MJ, Gay NJ. Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J*. 1982;1:945-951. doi:10.1002/j.1460-2075.1982.tb01276.x
73. Ambudkar SV, Kim IW, Xia D, Sauna ZE. The A-loop, a novel conserved aromatic acid subdomain upstream of the Walker A motif in ABC transporters, is critical for ATP binding. *FEBS Lett*. 2006;580:1049-1055. doi:10.1016/j.febslet.2005.12.051
74. Gorbalenya AE, Koonin EV. Helicases: amino acid sequence comparisons and structure-function relationships. *Curr Opin Struct Biol*. 1993;3:419-429. doi:10.1016/S0959-440X(05)80116-2
75. Koonin EV. A common set of conserved motifs in a vast variety of putative nucleic acid-dependent ATPases including MCM proteins involved in the initiation of eukaryotic DNA replication. *Nucleic Acids Res*. 1993;21:2541-2547. doi:10.1093/nar/21.11.2541
76. Raney KD, Byrd AK, Aarattuthodiyil S. Structure and mechanisms of SF1 DNA helicases. *Adv Exp Med Biol*. 2013;767:17-76. doi:10.1007/978-1-4614-5037-5_2
77. Velankar SS, Soultanas P, Dillingham MS, Subramanya HS, Wigley DB. Crystal structures of complexes of PcrA DNA helicase with a DNA substrate indicate an inchworm mechanism. *Cell*. 1999;97:75-84. doi:10.1016/S0092-8674(00)80716-3
78. Brosh RM Jr, Matson SW. Mutations in motif II of *Escherichia coli* DNA helicase II render the enzyme nonfunctional in both mismatch repair and excision repair with differential effects on the unwinding reaction. *J Bacteriol*. 1995;177:5612-5621. doi:10.1128/jb.177.19.5612-5621.1995
79. Graves-Woodward KL, Gottlieb J, Challberg MD, Weller SK. Biochemical analyses of mutations in the HSV-1 helicase-primase that alter ATP hydrolysis, DNA unwinding, and coupling between hydrolysis and unwinding. *J Biol Chem*. 1997;272:4623-4630. doi:10.1074/jbc.272.7.4623
80. Walker SL, Wonderling RS, Owens RA. Mutational analysis of the adeno-associated virus type 2 Rep68 protein helicase motifs. *J Virol*. 1997;71:6996-7004. doi:10.1128/jvi.71.9.6996-7004.1997
81. Weng Y, Czaplinski K, Peltz SW. Genetic and biochemical characterization of mutations in the ATPase and helicase regions of the Upf1 protein. *Mol Cell Biol*. 1996;16:5477-5490. doi:10.1128/mcb.16.10.5477
82. Pritham EJ, Putliwala T, Feschotte C. Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. *Gene*. 2007;390:3-17. doi:10.1016/j.gene.2006.08.008
83. Bush ZD, Alice FS, Dinwiddie D, Albers C, Hillers KJ, Libuda DE. Comprehensive detection of structural variation and transposable element differences between wild type laboratory lineages of *C. elegans*. *bioRxiv*. 2023. doi:10.1101/2023.01.13.523974
84. Capy P, Langin T, Higuet D, Maurer P, Bazin C. *Do the Integrases of LTR-Retrotransposons and Class II Element Transposases Have a Common Ancestor?* Vol. 100. Kluwer Academic Publishers; 1997.
85. Capy P, Vitalis R, Langin T, Higuet D, Bazin C. Relationships between transposable elements based upon the integrase-transposase domains: is there a common ancestor? *J Mol Evol*. 1996;42:359-368. doi:10.1007/BF02337546
86. Iwai T, Kurosawa N, Itoh YH, Kimura N, Horiuchi T. *Sequence Analysis of Three Family B DNA Polymerases from the Thermoacidophilic Crenarchaeon*. Vol. 7.Sulfurisphaera Ohwakuensis; 2000. Accessed November 25, 2024. https://academic.oup.com/dnaresearch/article/7/4/243/496032
87. Bao W, Kapitonov VV, Jurka J. Ginger DNA transposons in eukaryotes and their evolutionary relationships with long terminal repeat retrotransposons. *Mob DNA*. 2010;1:3. doi:10.1186/1759-8753-1-3
88. Koonin EV, Dolja VV, Krupovic M. Origins and evolution of viruses of eukaryotes: the ultimate modularity. *Virology*. 2015;479-480:2-25. doi:10.1016/j.virol.2015.02.039
89. Krupovic M, Koonin EV. Polintons: a hotbed of eukaryotic virus, transposon and plasmid evolution. *Nat Rev Microbiol*. 2015;13:105-115. doi:10.1038/nrmicro3389
90. Yutin N, Shevchenko S, Kapitonov V, Krupovic M, Koonin EV. A novel group of diverse Polinton-like viruses discovered by metagenome analysis. *BMC Biol*. 2015;13:95. doi:10.1186/s12915-015-0207-4
91. Szitenberg A, Koutsovoulos G, Blaxter ML, Lunt DH. The evolution of tyrosine-recombinase elements in Nematoda. *PLoS ONE*. 2014;9:e106630. doi:10.1371/journal.pone.0106630
92. Arkhipova IR. Distribution and phylogeny of penelope-like elements in eukaryotes. *Syst Biol*. 2006;55:875-885. doi:10.1080/10635150601077683
93. Cordaux R, Batzer MA. The impact of retrotransposons on human genome evolution. *Nat Rev Genet*. 2009;10:691-703. doi:10.1038/nrg2640
94. Penzkofer T, Jäger M, Figlerowicz M, et al. L1Base 2: more retrotransposition-active LINE-1s, more mammalian genomes. *Nucleic Acids Res*. 2017;45:D68-D73. doi:10.1093/nar/gkw925
95. Brouha B, Schustak J, Badge RM, et al. Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci U S A*. 2003;100:5280-5285. doi:10.1073/pnas.0831042100
96. Tubio JMC, Li Y, Ju YS, et al. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science*. 2014;345:1251343. doi:10.1126/science.1251343
97. Rodriguez-Martin B, Alvarez EG, Baez-Ortega A, et al. Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nat Genet*. 2020;52:306-319. doi:10.1038/s41588-019-0562-0
98. Gardner EJ, Lam VK, Harris DN, et al. The mobile element locator tool (MELT): population-scale mobile element discovery and biology. *Genome Res*. 2017;27:1916-1929. doi:10.1101/gr.218032.116
99. Ebert P, Audano PA, Zhu Q, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*. 2021;372:eabf7117. doi:10.1126/science.abf7117
100. Britten RJ. Active gypsy/Ty3 retrotransposons or retroviruses in *Caenorhabditis elegans*. *Proc Natl Acad Sci U S A*. 1995;92:599-601. doi:10.1073/pnas.92.2.599
101. Dennis S, Sheth U, Feldman JL, English KA, Priess JR. *C. elegans* germ cells show temperature and age-dependent expression of Cer1, a Gypsy/Ty3-related retrotransposon. *PLoS Pathog*. 2012;8:e1002591.
102. Moore RS, Kaletsky R, Lesnik C, et al. The role of the Cer1 transposon in horizontal transfer of transgenerational memory. *Cell*. 2021;184:4697-4712.e18. doi:10.1016/j.cell.2021.07.022
103. Sun B, Kim H, Mello CC, Priess JR. The CERV protein of Cer1, a *C. elegans* LTR retrotransposon, is required for nuclear export of viral genomic RNA and can form giant nuclear rods. *PLoS Genet*. 2023;19:e1010804. doi:10.1371/journal.pgen.1010804
104. Palopoli MF, Rockman MV, TinMaung A, et al. Molecular basis of the copulatory plug polymorphism in *Caenorhabditis elegans*. *Nature*. 2008;454:1019-1022. doi:10.1038/nature07171
105. Waterston RH, Lindblad-Toh K, Birney E, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*. 2002;420:520-62. doi:10.1038/nature01262
106. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001409:860-921. doi:10.1038/35057062
107. Zhang X, Wessler SR. Genome-wide comparative analysis of the transposable elements in the related species *Arabidopsis thaliana* and *Brassica oleracea*. *Proc Natl Acad Sci U S A*. 2004;13:5589-5594. doi:10.1073/pnas.0401243101
108. Bergman CM, Quesneville H, Anxolabéhère D, Ashburner M. Recurrent insertion and duplication generate networks of transposable element sequences in the *Drosophila melanogaster* genome. *Genome Biol*. 2006;7:R112. doi:10.1186/gb-2006-7-11-r112
109. Kim JM, Vanguri S, Boeke JD, Gabriel A, Voytas DF. Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. *Genome Res*. 1998;8:464-478. doi:10.1101/gr.8.5.464
110. Bowen NJ, Jordan IK, Epstein JA, Wood V, Levin HL. Retrotransposons and their recognition of pol II promoters: a comprehensive survey of the transposable elements from the complete genome sequence of *Schizosaccharomyces pombe*. *Genome Res*. 2003;13:1984-1997. doi:10.1101/gr.1191603
111. Howe K, Clark MD, Torroja CF, et al. The zebrafish reference genome sequence and its relationship to the human genome. *Nature*. 2013;496:498-503. doi:10.1038/nature12111
112. Havecker ER, Gao X, Voytas DF. The Diversity of LTR Retrotransposons, 2004. Accessed November 25, 2024. http://genomebiology.com/2004/5/6/225