Title

Conserved and divergent DNA recognition specificities and functions of R2 retrotransposon N-terminal domains.

Journal

Cell Reports, 43(5)

Authors

Lee, Rosa

Horton, Connor

Van Treeck, Briana

et al.

Publication Date

2024-05-28

DOI

10.1016/j.celrep.2024.114239

Peer reviewed

HHS Public Access

Author manuscript

Cell Rep. Author manuscript; available in PMC 2024 June 26.

Published in final edited form as:

Cell Rep. 2024 May 28; 43(5): 114239. doi:10.1016/j.celrep.2024.114239.

Conserved and divergent DNA recognition specificities and functions of R2 retrotransposon N-terminal domains

Rosa Jooyoung Lee[1,2], **Connor A. Horton**[1], **Briana Van Treeck**[1], **Jeremy J.R. McIntyre**[1], **Kathleen Collins**[1,3,*]

[1]Department of Molecular and Cell Biology, University of California at Berkeley, Berkeley, CA 94720, USA

[2]Present address: Department of Genetics, Stanford University, Stanford, CA 94305, USA

[3]Lead contact

SUMMARY

R2 non-long terminal repeat (non-LTR) retrotransposons are among the most extensively distributed mobile genetic elements in multicellular eukaryotes and show promise for applications in transgene supplementation of the human genome. They insert new gene copies into a conserved site in 28S ribosomal DNA with exquisite specificity. R2 clades are defined by the number of zinc fingers (ZFs) at the N terminus of the retrotransposon-encoded protein, postulated to additively confer DNA site specificity. Here, we illuminate general principles of DNA recognition by R2 N-terminal domains across and between clades, with extensive, specific recognition requiring only one or two compact domains. DNA-binding and protection assays demonstrate broadly shared as well as clade-specific DNA interactions. Gene insertion assays in cells identify the N-terminal domains sufficient for target-site insertion and reveal roles in second-strand cleavage or synthesis for clade-specific ZFs. Our results have implications for understanding evolutionary diversification of non-LTR retrotransposon insertion mechanisms and the design of retrotransposon-based gene therapies.

In brief

Lee et al. elucidate general principles of target-site recognition by the N-terminal DNA-binding domains of site-specific R2 retrotransposon proteins across species and clades. They identify a

*Correspondence: kcollins@berkeley.edu.

AUTHOR CONTRIBUTIONS

R.J.L. designed the study overall and carried out all biochemical and cellular assays and their analyses with the following exceptions. Initial screening for active A-clade R2 proteins was performed by B.V.T. Second-strand nicking assays were developed and performed in part by B.V.T., as well as replicate TPRT assays. ddPCR analyses were designed and performed by J.J.R.M. C.A.H. designed all bioinformatic analyses and performed them in collaboration with R.J.L. K.C. supervised experimental design and analyses. R.J.L. and K.C. wrote the manuscript with input from all other authors.

DECLARATION OF INTERESTS

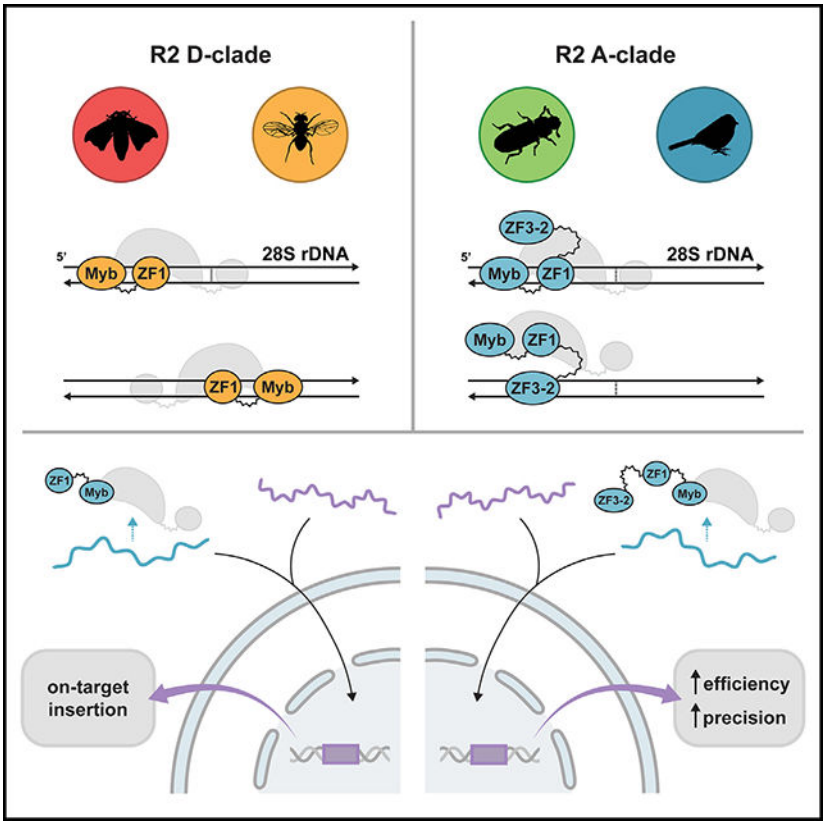
B.V.T. and K.C. are listed inventors on patent applications filed by University of California, Berkeley, related to the transgene insertion technology platform. B.V.T. and K.C. have equity options in Addition Therapeutics, which licensed the University of California, Berkeley technology. K.C. is a consultant and board member of Addition Therapeutics but does not receive personal compensation for these roles.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2024.114239.

minimal set of N-terminal domains sufficient for on-target gene insertion and also find roles for these domains in second-strand cleavage or synthesis.

Graphical Abstract



INTRODUCTION

Retrotransposons are genetic elements that mobilize in the genome via an RNA intermediate. Retrotransposon RNA is reverse transcribed, and the corresponding complementary DNA (cDNA) is inserted at a new genomic locus.[1] Non-long terminal repeat (non-LTR) retrotransposons are early-evolved eukaryotic retroelements[2,3] with only one or two open reading frames (ORFs).[4] They compose significant percentages of many eukaryotic genomes.[4–6] Indeed, the reverse transcriptase (RT) activity of long interspersed element 1 (LINE-1), the only endogenous human non-LTR retrotransposon with observed autonomous mobility, has given rise to over 30% of the human genome, through both self-mobilization and use of the non-coding Alu RNA as template.[7,8] Transposon mobilization and high copy number can both harm and help the host. Mobility can disrupt coding or regulatory sequences and enable genomic rearrangements.[7,9,10] Such genomic instability has been implicated in human genetic disorders, including increased susceptibility for certain cancers.[10,11] However, these very traits also make non-LTR retrotransposons drivers of genomic innovation, generating new genes, gene isoforms, and networks of gene regulation.[12]

The R2 family of non-LTR retrotransposons inserts into a specific, well-conserved site in the gene encoding precursor 28S ribosomal RNA (rRNA),[13,14] which is present as multicopy loci (rDNA) transcribed by RNA polymerase I.[15] R2 is among the most phylogenetically widespread eukaryotic mobile elements but is not present in mammals,[6] which instead have non-site-specific retrotransposons such as human LINE-1. The R2 retrotransposon encodes a single protein, which has nucleic acid-binding, RT, and endonuclease (EN) activities.[14] R2 retrotransposons branch into four subclades, each defined by the number and type of zinc fingers (ZFs) at the protein N terminus (Figure 1A).[6,16,17] Members of the ancestral A clade have three ZFs, termed ZF3, ZF2, and ZF1 (ZF3 is the most N-terminal ZF), whereas members of the extensive D clade have only ZF1. ZF3 and ZF1 are CCHH-type ZFs, while ZF2 is CCHC type. Members of the B and C clades have different sets of two ZFs each. The ZF(s) are followed by a Myb domain, an RT domain, a zinc knuckle (ZK), and a restriction-like EN domain (Figure 1A).[14,18]

The R2 protein, bound to its own RNA transcript, finds its target site and makes a first-strand nick on the rRNA-antisense strand (the rDNA bottom strand in illustrations here). It then does first-strand synthesis by reverse transcribing its bound RNA using the nick-liberated 3′-OH as a primer, a mechanism termed target-primed reverse transcription (TPRT).[13] Next, the second strand (the rDNA top strand in illustrations here) is cleaved, possibly by the R2 protein.[13,19,20] For R2 from *Bombyx mori*, the primary R2 model system, the second strand is cleaved two nucleotides (nt) upstream of the first-strand nick position.[13] In some species, the newly synthesized cDNA 3′ end may base pair with the upstream rDNA. The second strand is then synthesized, and the strand junctions are repaired (Figure 1B). Second-strand synthesis has not been robustly observed in TPRT reactions *in vitro*,[19] and retrotransposon proteins do not encode all necessary enzymatic activities to complete DNA repair, implicating host factor involvement.

Avian R2 proteins from the A clade have promising applications in transgene supplementation of the human genome, providing a mechanism for gene insertion into rDNA as a safe harbor.[21] Safe-harbor transgene delivery would complement CRISPR-Cas approaches for endogenous gene editing. A current limitation is that A-clade R2 protein domains and rDNA sequence elements that support high insertion-site specificity have not been explored. Almost all biochemical characterization of R2 protein has been carried out using the D-clade *B. mori* protein.[14] The recombinant *B. mori* full-length protein binds target rDNA over an extensive ~60-base-pair (bp) region extending from approximately 40 bp upstream of the cleavage site to 20 bp downstream,[20,22,23] proposed to reflect binding of two subunits, with the upstream subunit performing first-strand nicking and cDNA synthesis and the downstream subunit performing second-strand nicking.[20] Previous work using recombinant *B. mori* R2 protein ZF1 and Myb N-terminal domains showed that these domains bind across and downstream of the first-strand nick site.[24] Recent cryoelectron microscopy (cryo-EM) structures of the full-length *B. mori* R2 protein synthesizing cDNA using the 3′ untranslated region (UTR) of its RNA transcript instead place the N-terminal ZF1 and Myb domains bound upstream of the cleavage site.[25,26] Recombinant N-terminal polypeptides from the R2 A-clade retrotransposon from *Limulus polyphemus* and the R2-related R9 retrotransposon from *Adineta vaga* directly recognize sequences upstream of their

cleavage sites,[27,28] but whether these are functional interactions or are broadly representative of A-clade-related R2 retrotransposons remains unknown.

We sought to determine if there are general principles for DNA sequence recognition by the N-terminal domains from R2 proteins within and across clades, as well as how these domains contribute to TPRT *in vitro* and new sequence insertion into rDNA in cells. We discovered shared and distinct DNA interaction specificities of R2 protein domains, comparing across four proteins of two subclades in clade D and two subclades in clade A. Unexpectedly, we found that the A-clade ZF3 and ZF2 domains (together, ZF3-2) make no evident contribution to DNA binding by the full N-terminal region of either A-clade R2 protein tested, and they also are not critical for TPRT activity by the full-length protein *in vitro*. However, using cellular assays, we show that A-clade-specific ZF3-2 do contribute to the efficiency of gene insertion into rDNA. Of particular interest, ZF3-2 removal drastically reduces the precision of gene-insertion 5′ junction formation by second-strand synthesis. Although A-clade R2 proteins do not bind downstream of the target site in the same manner as D-clade R2 proteins, they nonetheless perform second-strand cleavage *in vitro* at the same position as the B. *mori* D-clade R2 protein. Our findings reveal overarching similarities in domain requirements for DNA binding and TPRT across R2 clades and also elaborate distinctions between proteins even from the same clade, suggesting ongoing evolutionary reprogramming of how individual domains support target-site specificity. Additionally, our results indicate unexpected functions for the N-terminal ZFs of A-clade R2 proteins in creating the 5′ rDNA-transgene junction. These insights inform mechanisms of native non-LTR retrotransposon mobility and their manipulation for genome engineering.

RESULTS

D- and A-clade R2 protein N-terminal domains have clade-specific DNA interactions

To investigate the evolutionary dynamics of R2 protein recognition of its 28S rDNA target site, we selected representative R2 retrotransposons that were known or we established to have biochemical and/or biological TPRT activity: the D-clade retrotransposons from *B. mori*[13] (BoMo, clade D2) and *Drosophila simulans*[29] (DroSi, clade D5) and the A-clade retrotransposons from *Tribolium castaneum*[17] (TrCasB, clade A2) and *Zonotrichia albicollis* (ZoAl, clade A3).[6] The target-site regions of rDNA in each species are highly similar, but the species-matched sequence was used for characterization of DNA-binding specificity (Figures 1C and S1A). For each R2 protein, the full N-terminal region comprising the one to three ZFs and the Myb domain (Figure 1D), as well as a series of truncations (Figure 1E), were expressed in *Escherichia coli* as fusion proteins with an N-terminal maltose-binding protein (MBP) tag and a C-terminal 6-histidine (6xHis) tag. Proteins were purified to near homogeneity by two-step chromatography (Figure 1F). Rigorous high-salt wash steps were employed during the purification to remove contaminating nucleic acids. Polypeptides harboring R2 N-terminal DNA-binding domains only are hereafter referred to by their shorthand (Figure 1E); for example, “NBoMo” is used for *B. mori* R2 protein N-terminal region.

For a first readout of how target-site recognition may be mediated by the N-terminal ZF and Myb domains, we screened binding of the full N-terminal-region proteins to upstream

or downstream target rDNA half-sites by electrophoretic mobility shift assay (EMSA). We designed upstream and downstream half-sites of equal length that together extended beyond the footprint of full-length *B. mori* protein expressed in *E. coli*[20,22,23] and were slightly shifted for Dversus A-clade proteins based on prior studies of *B. mori*[24] and *L. polyphemus*[28] N-terminal region DNA binding. The upstream rDNA half-site used for D-clade N-terminal region binding spanned from −50 to −9 and the downstream site from −8 to +34. The half-sites used to test A-clade N-terminal region binding covered a greater length from +50 to −50 in case rDNA binding determinants extended beyond the footprint of *B. mori* R2 protein. The upstream rDNA half-site used for A-clade N-terminal region binding spanned from −50 to −1 and the downstream site from +1 to +50. All target-site oligonucleotide duplexes were 5′-end radiolabeled on the sense strand. To minimize nonspecific protein-DNA interactions, a high concentration of nonspecific competitor was added to EMSA reactions (see STAR Methods). The same binding conditions of DNA concentration (2 nM) and range of protein concentrations (up to 2,000 nM) were used for all EMSAs in Figures 2, 3, and 4.

Reconciling previous studies, we observed that the D-clade R2 N-terminal regions bound to both the upstream and downstream rDNA half-sites (Figures 2A and 2B). NBoMo ZF1-Myb bound with highest affinity to the downstream half-site, consistent with previous studies,[24] but it also bound the upstream half-site, with which it associates in the context of full-length protein[25,26] (Figure 2A). We conclude that the RT domain region in contact with upstream DNA in recent structures[25,26] is not critical for NBoMo association with the target site. In comparison, NDroSi ZF1-Myb also bound both upstream and downstream half-sites but did so with comparable affinity for each (Figure 2B). A single protein-DNA complex was detected for each half-site that, by inference from previous studies, should represent one protein molecule bound to one DNA molecule. It is notable that the compact ZF1-Myb module of D-clade R2 proteins can recognize two distinct, physically separate sequences. The binding of two full-length D-clade proteins at a target site, each via its N-terminal region, could give rise to the entire 60-bp protection demonstrated by the full-length *B. mori* protein.[20,22,23]

Curiously, despite having more ZF domains, the N-terminal regions of both A-clade R2 proteins, NTrCasB and NZoAl, bound to only the upstream half-site (Figures 2C and 2D). As with the D-clade R2 ZF1-Myb polypeptides, A-clade ZF3-Myb polypeptides formed predominantly a single mobility-shifted protein-DNA complex on the upstream half-site, likely corresponding to one protein molecule bound to one DNA molecule. The upstream selectivity of high-affinity binding by A-clade N-terminal domains, in contrast with the ability of D-clade N-terminal domains to bind both upstream and downstream, indicates differences in DNA recognition principles between R2 clades. We cannot exclude the possibility that A-clade R2 proteins also contact DNA downstream of the first-strand nick, but the apparent lack of downstream half-site binding, and the absence of protection in that region (see below), are consistent with results for the *L. polyphemus* A-clade R2 protein N-terminal region.[28]

D-clade ZF1 and Myb domains variably contribute to target-site recognition

We hypothesized that the ability of D-clade R2 protein N-terminal regions to bind both the upstream and downstream sides of the target site could be explained by separate specificities of ZF1 versus the Myb domain. We therefore investigated how the Myb domain with and without ZF1 interacts with DNA by EMSA. Using the full rDNA target-site duplex (−50 to +50), two mobility-shifted protein-DNA complexes were detected with both NBoMo and NDroSi ZF1-Myb, likely reflecting one versus two proteins bound to each DNA molecule (Figures 3A and 3B, left). Proteins lacking ZF1 but retaining the Myb domain showed different binding profiles. NBoMo Myb bound DNA with substantially lowered affinity, with the first mobility shift forming slower-migrating complexes at a protein concentration roughly two orders of magnitude higher than when ZF1 was present (Figure 3A, right). In comparison, NDroSi Myb generated slower-migrating shifted complexes at only very high protein concentrations (Figure 3B, right), showing even greater dependence of the Myb domain on ZF1 for DNA interaction.

To compare binding of the two D-clade N-terminal regions with higher sequence resolution, we employed DNase I footprinting. Protein and target DNA were co-incubated to allow binding, then DNase I was added to reveal regions of DNA not protected by the protein. Partial DNase I digestion products were isolated and resolved by denaturing polyacrylamide gel electrophoresis (PAGE). Regions of protection by the protein appear as gaps in the ladder of cleavage products relative to a control lane without DNA-binding protein. The gaps were mapped by comparison to a reference G + A ladder, the same target DNA chemically cleaved at every G and A in the sequence. The NBoMo ZF1-Myb footprint included protection from −2 to +6, +8 to +15, and +17 to +20 (Figure 3C), generally consistent with previous results.[24] The NDroSi ZF1-Myb footprint in the downstream target site matched the NBoMo ZF1-Myb footprint (Figures 3D and 3E). NDroSi ZF1-Myb also protected the upstream target site from −37 to −25 and −21 to −15 (Figure 3D), about the same length as the downstream footprint and closely matching the A-clade R2 N-terminal region footprints (Figure 3E and see below). Despite comparable binding affinities of NDroSi ZF1-Myb for both half-sites (Figure 2B), the NDroSi ZF1-Myb upstream footprint appeared to give weaker protection than the downstream footprint (Figure 3D). The weaker protection of upstream DNA by NDroSi ZF1-Myb and the lack of an upstream DNA footprint detectable for NBoMo ZF1-Myb (Figures 3C and 3E), consistent with previous work,[24] suggest the possibility that RT domain contact with upstream DNA[25,26] stabilizes the initial ZF1-Myb binding configuration to reduce dissociation (see section “discussion”). Overall, we suggest that ZF1 and Myb from some, if not all, D-clade R2 proteins together constitute a DNA-binding module with two separate sequence specificities that enable protein binding to two separate regions of the target site.

A-clade ZF1 and Myb domains confer high-affinity target-site binding

To investigate the DNA-binding properties of A-clade R2 protein N-terminal domains, we carried out EMSA and DNase I footprinting assays with full and ZF-truncated polypeptides using the upstream rDNA half-site. Surprisingly, for both NTrCasB and NZoAl proteins, removal of ZF3 or ZF3-2 appeared to increase rather than decrease DNA-binding affinity. ZF1-Myb proteins (ΔZF3-2) produced detectable mobility shifts at ~5 nM, whereas the full

ZF3-Myb or ZF2-Myb (ΔZF3) proteins produced a detectable mobility shift at ~25 nM (Figures 4A and 4B). Quantification of EMSA data supported a lower K_d for ZF1-Myb than for full N-terminal-region proteins (Figures 4C and 4D). Additional truncation of NTrCasB and NZoAl ZF1-Myb proteins to the Myb domain alone had strikingly different impact on DNA binding: NTrCasB Myb showed robust DNA binding, equivalent to NTrCasB ZF1-Myb, whereas no specific binding was observed for NZoAl Myb under our assay conditions (Figures 4A and 4B, right; Figures 4C and 4D). Unlike D-clade NBoMo Myb (Figure 3A, right), NTrCasB Myb did not suffer any reduction in binding affinity from loss of ZF1 (Figure 4A, right; Figure 4C), indicating unique features of DNA-binding specificity and affinity in the N-terminal domains of each R2 protein.

DNase I footprinting for NTrCasB and NZoAl gave similar protection of two distinct segments of the target site, one centered in the −30 region and the other centered in the −15 region (Figures 4E and 4F, left; Figure 4G). Extending target site DNA length to include an additional 50 bp of downstream sequence did not change this pattern of protection (Figure S1B). The regions of protection partially overlap results of DNase I footprinting using an N-terminal polypeptide from *L. polyphemus*[28]; however, our assays did not indicate a region of protection around −10. For NTrCasB, protection of the −15 region was less complete than observed for NZoAl. Both A-clade N terminus protein footprints resemble the upstream NDroSi ZF1-Myb footprint (Figures 3D and 3E), which also had a large region of protection centered at −30 and a smaller protection region centered at −15; this pattern of protection could reflect DroSi, ZoAl, and TrCasB DNA-binding domains forming the protein-DNA interactions demonstrated for BoMo ZF1 and Myb domains by cryo-EM.[25,26] Consistent with our finding that the NTrCasB Myb domain retained high DNA-binding affinity on its own (Figure 4A), each truncated NTrCasB protein, including the Myb domain alone, gave the same footprint (Figure 4E). This footprint pattern was also observed for NZoAl proteins lacking ZF3 or ZF3-2 (Figure 4F), with no indication of a previously proposed footprint for ZF3 on the target site.[28] Combined, these findings suggest that the R2 Myb domain is essential for target-site DNA recognition, typically but not universally with a contribution from ZF1.

ZF3-2 domains are not required for TPRT *in vitro*

We next interrogated which A-clade ZFs were functionally critical for the combined DNA binding, first-strand nicking, and primerelongation activities of TPRT. Although we had anticipated that the A-clade ZF3 and ZF2 domains would improve target-site binding, results above suggest otherwise for ZoAl protein and raise the possibility that TrCasB protein might not need even ZF1 for this purpose. On the other hand, because TPRT likely requires changes in DNA configuration after initial sequence recognition, all of the ZF domains could be required. To assay for TPRT, we purified full-length wild-type (WT) and N-terminally truncated TrCasB and ZoAl proteins with an N-terminal FLAG tag, overexpressed in HEK293T cells (Figure 5A). The truncated proteins had approximately equivalent or better (for ΔZF3-1) expression and purification yield than the full-length R2 proteins, monitored by immunoblot for the FLAG tag (Figure 5B). An annealed 64-bp target-site duplex with a 5′-radiolabeled antisense strand (bottom strand in Figure 5C) allowed monitoring of the proportion of target-site strand that was intact, nicked, or nicked

and extended by cDNA synthesis. R2 protein, its cognate R2 3′ UTR, and dNTPs were added to the reaction, allowing first-strand cleavage and TPRT to occur (Figure 5C). Purified input and product DNAs were resolved by denaturing PAGE (Figure 5D).

Product DNAs were produced corresponding to first-strand nicking followed by reverse transcription of one or two consecutive 3′ UTR RNA molecules (Figure 5D, 1X and 2X), resulting from initial TPRT and subsequent elongation of the initial TPRT product by template jumping.[21,30] Unexpectedly, both TrCasB and ZoAl proteins lacking the A-clade-specific ZF3 (ΔZF3) or ZF3 and ZF2 (ΔZF3-2) supported robust target-site nicking and TPRT, but no activity was detected for the ΔZF3-1 proteins lacking all N-terminal ZFs (Figures 5D and 5E). We conclude that A-clade R2 protein ZF1 and Myb domains are required and sufficient to support accurate first-strand nicking, cDNA synthesis initiation, and processive cDNA synthesis *in vitro*. This observation establishes cross-clade conservation of an R2 protein N-terminal module necessary for target-site recognition and TPRT.

ZF3-2 domains facilitate but are not required for transgene insertion into rDNA

In search of possible functions for A-clade ZF3-2, we turned to cellular assays of gene insertion. We recently developed an assay for TPRT-initiated synthesis of autonomous transgenes into rDNA in human cells using ZoAl protein, termed precise RNA-mediated insertion of transgenes (PRINT).[21] For PRINT, an mRNA encoding ZoAl protein and a separate template RNA are co-transfected (Figure 6A, top). In cells, the mRNA is translated into ZoAl protein, which binds the avian R2 3′ UTR in the template RNA (Figure 6A, middle). This ribonucleoprotein (RNP) accesses the nucleus and synthesizes the transgene cassette encoded by the template RNA into the 28S rDNA (Figure 6A, bottom). Here we used a template RNA encoding a transgene expression cassette with a version of the human cytomegalovirus (CMV) immediate-early enhancer and promoter, enhanced green fluorescent protein (GFP) ORF, and minimal polyadenylation signal (PA), together flanked on the 5′ side by a self-cleaving ribozyme (RZ) present at most native R2 retrotransposon 5′ ends and on the 3′ side by an avian R2 3′ UTR and 3′ tail sequence R4A22. The RZ at the template RNA 5′ end has 28 nt of sense-strand rRNA immediately upstream of the first-strand nick position. R4A22 contains 4 nt of rRNA sequence immediately downstream of the first nick position, complementary to the nicked TPRT primer, and a terminal tract of 22 adenosines.

We exploited PRINT to compare the overall efficiency of gene insertion into rDNA by ZoAl WT, ΔZF3, ΔZF3-2, and ΔZF3-1. Template RNA and mRNA encoding one of the ZoAl protein variants were co-transfected into human RPE-1 cells (Figure 6A, top). One day post transfection, cells were harvested and analyzed by flow cytometry to detect GFP-positive cells with functional transgene insertions. A broad distribution of GFP fluorescence intensities was generated (Figure 6B, and see Figure S2 for replicates), in part from variable copy number of insertions into the hundreds of rDNA units per cell.[21] Negative control transfections of mRNA alone or template RNA alone had solely background fluorescence (Figures 6B and S2).

Removal of ZF3 or ZF3-2 reduced the percentage of GFP-positive cells from ~42% to ~10% and decreased the median GFP intensity in the GFP-positive cell pool (Figure 6C), both indicating fewer transgene insertions. Removal of all three ZFs drastically reduced the percentage of cells scored as GFP positive to only marginally above background, and the few cells scored as GFP positive had much lower GFP intensity than in cells with transgenes inserted by ZoAl ΔZF3 or ΔZF3-2, indicating that ZoAl ΔZF3-1 supported minimal if any transgene insertion (Figures 6B, 6C, and S2). PCR amplification was performed to detect the rDNA-inserted transgene 5′ junction, 3′ junction, and GFP ORF from genomic DNA of the transfected cell pools. PCR gave robust signal for GFP ORF and both junctions from cells transfected with template RNA and ZoAl WT, reduced signal from cells transfected with template RNA and ZoAl ΔZF3 or ΔZF3-2, and no signal from cells transfected with template RNA and ZoAl ΔZF3-1 (Figure S3), paralleling transgene insertion efficiency monitored by GFP fluorescence. These comparisons indicate that removal of ZF3 or ZF3-2 reduced but did not eliminate transgene insertion at the rDNA target site. In contrast, removal of ZF3-1 was severely inhibitory. These results indicate that the cross-clade conserved ZF1-Myb N-terminal module is essential for insertion of transgenes into rDNA by A-clade R2 proteins, while the A-clade-specific ZF3 and ZF2 domains are not.

ZF3-2 domains influence the rDNA position of transgene 5′ junctions

To compare transgene insertion by ZoAl WT, ΔZF3, and ΔZF3-2 in more detail, GFP-positive cells from PRINT with ZoAl WT, ΔZF3, or ΔZF3-2 were isolated by cell sorting. Genomic DNA was extracted for droplet digital PCR (ddPCR) and Illumina whole-genome sequencing (WGS). WGS reads were first mapped to a custom scaffold representing a transgene insertion in target-site 28S rDNA (Figure 6A, bottom), with the precise 5′ and 3′ junctions expected from the template RNA 3′ tail annealing to the target-site primer and the cDNA 3′ end annealing to the sense strand of upstream target-site rDNA.[21] Any non-aligned portions of reads containing transgene sequence were aligned to a 45S rDNA reference to detect deletion or duplication of sequence flanking the target site during insertion or repair. Any remaining portions of partially transgene-mapping reads that did not align to rDNA were then mapped to the human genome. Transgene insertions by ZoAl WT, ΔZF3, or ΔZF3-2 were full length or 5′ truncated (Figure S4A), consistent with previous observations[21] and native non-LTR retrotransposon mobility.[31]

Insertion copy number and full-length transgene percentage were quantified by ddPCR in sorted GFP-positive cells. Based on detection of the transgene 3′ end, an average of ~30 insertions per cell occurred with ZoAl WT, reduced to an average of ~15 per cell for ZoAl ΔZF3 and ΔZF3-2 (Figure 6D, left y axis). This trend is consistent with the lower median GFP intensity produced by ZoAl ΔZF3 and ΔZF3-2 (Figure 6C). Considering both the lower percentage of GFP-positive cells and the lower average insertion copy number in sorted GFP-positive cells, ZoAl ΔZF3 and ΔZF3-2 generated ~10-fold fewer insertions (7and 11-fold for ZoAl ΔZF3 and ΔZF3-2, respectively). Furthermore, based on the ratio of ddPCR detection of the transgene 3′ versus 5′ end, ZoAl ΔZF3 and ΔZF3-2 generated a reduced percentage of full-length insertions: 19% and 16%, respectively, compared to 30% for ZoAl WT (Figure 6D, right y axis). We conclude that, in addition to compromised

insertion efficiency, ZoAl ΔZF3 and ΔZF3-2 transgene insertions also had increased 5′ truncations.

Despite these differences, ZoAl WT, ΔZF3, and ΔZF3-2 all retained the previously characterized[21] high target-site fidelity of transgene insertion into rDNA (Figure 6E). Also, all three proteins produced a small minority of transgene 3′ junctions with an extra rDNA nt before the start of cDNA synthesis (Figure S4B), which would result from first-strand nicking 1 nt upstream from the canonical site.[21] ZoAl WT, ΔZF3, and ΔZF3-2 all generated 5′ insertion junctions of the same categories previously described for ZoAl WT.[21] Full-length transgenes can form a 5′ junction by annealing of the cDNA 3′ end to upstream rDNA, which generates a seamless junction (“Anneal” category; Figures 6F and 6G, blue bars). A small fraction of full-length transgene insertions instead occurred by direct joining of the cDNA 3′ end to rDNA, and, as expected, this junction type was common for 5′-truncated transgenes (“Join” category; Figures 6F and 6G, red bars). With all three protein variants, some transgene 5′ ends were followed by a segment of sequence generated by cDNA 3′-end priming of additional synthesis prior to 5′ junction formation (“Snapback” category; Figures 6F and 6G, purple bars). The template for snap-back synthesis was most commonly the cDNA itself, with some snap-back synthesis on nearby rDNA (Figures S4C and S4D). Snap-back synthesis of antisense cDNA is then followed by junction formation with upstream rDNA.[21] Rarely, as previously reported,[21] a U6 small nuclear RNA was copied after transgene synthesis prior to 5′ junction formation, possibly by template jumping (“Extra template” category; Figures 6F and 6G, green bars; Table S1). For both full-length and 5′-truncated insertions, proportions of each category of 5′ junction were not notably different across the three protein variants.

The most striking difference in transgene insertion by ZoAl WT, ΔZF3, and ΔZF3-2 was in the rDNA locations of transgene 5′ junctions. Using the Join category of 5′ junctions, we evaluated where the rDNA was joined to a transgene 5′ end, indicative of target site deletion or duplication. With ZoAl WT, the predominant rDNA positions of 5′ junction formation were at or slightly upstream of the first-strand nick (Figure 6H, top; zoom-in proximal to the target site in Figure S4E). In contrast, with ZoAl ΔZF3 and ΔZF3-2, there was more heterogeneity in the position of rDNA joining to the transgene 5′ end (Figure 6H; bootstrap hypothesis testing for difference in median between samples: WT vs. ΔZF3, $p = 7.9e{-4}$; WT vs. ΔZF3-2, $p = 1.5e{-4}$; ΔZF3 vs. ΔZF3-2, $p = 0.38$). We note that junction formations far upstream or downstream of the target site cannot be definitively identified as rDNA deletions or duplications because rDNA units are present in the genome as tracts of tandem direct repeats. Altogether, we conclude that the N-terminal ZFs present in A-clade R2 proteins but absent in D-clade R2 proteins are not critical for specificity of target-site selection, first-strand nicking, or initiation of cDNA synthesis, but their loss alters the fidelity of 5′ junction formation, resulting in fewer full-length transgene insertions and highly heterogeneous positioning of rDNA fusion to the transgene 5′ end.

ZF3-2 domains bind upstream target-site DNA and may stimulate second-strand nicking

The biological processes underlying second-strand nicking and synthesis are long-standing unresolved questions for non-LTR retrotransposon mobility. *B. mori* R2 protein can nick

the second strand *in vitro*,[13,19] but this activity is inefficient and thought to be dependent on protein interaction with the downstream DNA-binding site,[20] which does not appear to be an interaction shared by A-clade R2 proteins (Figures 2 and 4). If a different DNA interaction specificity is required to position the R2 protein EN domain for second-strand versus first-strand nicking, then the A-clade proteins would need a second mode of DNA binding, beyond ZF1-Myb association with the upstream rDNA target-site. We therefore tested whether the ZF3-2 polypeptide could bind target-site DNA. ZF3-2 are predicted to fold together (Figure S5). EMSAs performed with ZF3-Myb polypeptides required high monovalent ion concentration and a large excess of nonspecific competitor to resolve discrete protein-DNA complexes from a low-mobility smear. Under those conditions, A-clade ZF3-2 alone did not bind target-site DNA in a manner detectable by EMSA. However, under less stringent binding conditions, and with micromolar rather than nanomolar protein concentrations, both NZoAl and NTrCasB ZF3-2 bound target-site duplex (Figures 7A and 7B, lane sets at far right).

We used target-site duplexes with sequential 10- or 20-bp segments of scrambled sequence to test the sequence specificity of A-clade ZF3-2 DNA binding. Surprisingly, for both NZoAl and NTrCasB ZF3-2, EMSAs revealed a binding requirement for the −40 to −21 region (Figures 7A and 7B), within the ZF1-Myb protected footprint (Figure 4G). Consistent with sequence-specific binding to the upstream target site, DNA binding by ZF3-2 was not sensitive to the presence or absence of a first-strand nick (Figure S6A), which we tested based on the characterization of CCHC zinc fingers such as ZF2 in PARP1[32] and DNA ligase III[33] as mediators of nick recognition. Also, ZF3-2 did not detectably bind the downstream region from +2 to +50, even in the presence of single-stranded upstream target-site sequence from −50 to +1 (Figure S6B). It remains possible that ZF2 (or ZF3) recognizes a nicked or gapped structural intermediate of the gene-insertion process in the context of the full-length protein at a specific stage of the process in cells. EMSAs of ZF3-2 and ZF1-Myb polypeptides mixed together, intended to test binding synergy or competition, were uninterpretable because ZF3-2 had no influence under binding conditions that can resolve ZF1-Myb•DNA complexes, and ZF1-Myb shifted DNA to a low-mobility smear under binding conditions that can resolve ZF3-2•DNA complexes.

To investigate whether A-clade R2 proteins can introduce a second-strand nick, we used a target-site duplex 5′ radiolabeled on the sense strand, with the complementary strand either intact or pre-nicked by annealing two shorter oligonucleotides to mimic the product of first-strand cleavage (Figure 7C, top). We note that the physiological second-strand nicking substrate is unknown, so any *in vitro* assay of second-strand nicking remains naive in its design. As an initial foray, for this study we tested the intact and pre-nicked target-site duplexes in reaction conditions with DNA alone or with added template RNA and dNTPs to support TPRT. Starting with just the pre-nicked target-site duplex in the reaction, we assayed for second-strand nicking by the A-clade R2 proteins. As a positive control, we used the *B. mori* R2 protein expressed with the same protein tags (here, BoMo). Both A-clade proteins made a second-strand nick at the same position as D-clade BoMo protein, with the strongest second-strand nicking by TrCasB protein (Figure 7D). ZoAl protein second-strand nicking activity was eliminated by EN active-site mutation (DD1041/1054AA)[21] (Figure 7D, EN dead).

Under the same reaction conditions (pre-nicked target site, no RNA or dNTPs), ZoAl ΔZF3 and ΔZF3-2 failed to nick the second strand (Figure 7E, lanes 1–4). In comparison, using the pre-nicked target site under TPRT conditions (added RNA and dNTPs), second-strand nicking by ZoAl ΔZF3 and ΔZF3-2 was approximately equal to or greater than that of ZoAl WT (Figure 7E, lanes 5–7). Neither assay condition supported second-strand nicking by ZoAl ΔZF3-1 (Figure 7E, lanes 4 and 8). We also tested second-strand nicking of intact target-site duplex. Again, TPRT conditions were required for detectable second-strand nicking by ZoAl ΔZF3 and ΔZF3-2 (Figure 7E, lanes 9–16). In parallel with the ZoAl proteins, we tested the TrCasB proteins. Using pre-nicked target site, the robust TrCasB second-strand nicking activity was largely independent of ZF3-2, with or without TPRT conditions (Figure 7F, lanes 1–8). Even TrCasB ΔZF3-1 generated some nicked top-strand product, albeit at very reduced level and with less specific positioning than TrCasB WT, ΔZF3, and ΔZF3-2. With intact target-site duplex, there was no detectable TrCasB ΔZF3-1 top-strand nicking, but TrCasB WT, ΔZF3, and ΔZF3-2 all retained nicking activity regardless of TPRT conditions (Figure 7, lanes 9–16).

These results indicate that A-clade R2 proteins can perform precise second-strand nicking, despite an apparent lack of high-affinity ZF1-Myb binding downstream of the first-strand nick. If alternate DNA-binding specificities contribute to toggling an R2 protein between first- and second-strand nicking activities, A-clade ZF3-2 binding to upstream target-site sequence could provide the alternate positioning (Figure 7G). Although ZoAl ZF3-2 domains influence second-strand nicking under some conditions, more understanding of the physiological second-strand nicking substrate is required to interpret and extend any *in vitro* assay conclusions. Also, *in vitro* second-strand nicking by any of the R2 proteins does not approach the efficiency of first-strand nicking, so it will be important to explore whether host factors mediate second-strand nicking and synthesis in cells.

DISCUSSION

Over the hundreds of millions of years since the pre-Cambrian origin of the R2 retrotransposon,[2] the 28S rDNA target site has been relatively well conserved, enabling widespread R2 phylogenetic perpetuation without evolutionary pressure for adaptation of new target-site specificity. High fidelity of target-site selection has been maintained in R2 retrotransposons from most characterized species.[6] Experiments above establish that four proteins active for TPRT from the two large R2 clades share interaction specificity for a specific region of rDNA upstream of the first-strand nick, mediated by R2 protein N-terminal domains shared across clades. We also demonstrate clade-specific protein interactions with the downstream target site and within-clade differences in reliance on ZF1 for high DNA-binding affinity. These unexpected findings open the possibility that heterogeneity in DNA-binding properties confers differential efficiency or regulation of endogenous R2 mobility. To test this possibility, we hope to develop an R2 retrotransposition assay. Among other advances, this will require discovery of R2 transcripts that are efficiently translated, since native mobility requirements for R2 transcript processing, nuclear export, and translation are all bypassed using PRINT to assemble an R2 RNP.

DNA binding

A unifying observation for all R2 proteins tested here is the primary role of the ZF1 and Myb domains in providing binding affinity for target-site DNA. Nonetheless, the significance of the ZF1 domain for DNA-binding affinity varied. Recent cryo-EM structures of *B. mori* R2 RNP reveal that motif 6a in the RT domain contacts DNA between the ZF1 and Myb domains.[25,26] We speculate that ZF1-Myb alone could scan the genome for potential target sites, followed by locking in an elongation-productive RNP conformation at the correct target site with DNA contact by motif 6a. Motif 6a participation could stabilize *B. mori* N-terminal region contact with the upstream target site, which is relatively weak compared to other R2 protein N-terminal regions tested here by EMSA.

A second principle to emerge from these studies is the clade-specific versatility of sequence recognition by ZF1-Myb. When assayed by EMSA and nuclease protection, as done in this work, A-clade N-terminal DNA-binding domains interact exclusively with upstream target-site sequence (Figure 7G, right), whereas D-clade N-terminal DNA-binding domains interact with both upstream and downstream target-site sequence in a non-exclusive manner (Figure 7G, left). Our results confirm and reconcile previous demonstrations of *B. mori* N-terminal domains binding to the downstream site[24] and full-length *B. mori* protein binding to the upstream target site.[20,22,25,26] Furthermore, our findings indicate that both upstream-bound and downstream-bound subunits use ZF1-Myb to bind DNA. We speculate that it may be unique to *B. mori* and phylogenetically proximal R2 proteins that binding of a 5′ versus 3′ R2 RNA portion directs full-length protein binding to downstream versus upstream regions, respectively,[23] since, unlike NBoMo, NDroSi readily binds both target-site regions.

Mobility and PRINT

Genome insertions by non-LTR retrotransposons have generated a substantial proportion of many eukaryotic genomes,[4–6] but many of the mechanisms involved in this process, beyond TPRT, are surprisingly opaque. This is a glaring knowledge gap given the biological,[8–10] evolutionary,[12] and disease[7,10,11] significance of non-LTR retrotransposons. D-clade R2 retrotransposons are abundant in arthropods, but it is the A-clade retrotransposons that are common in species phylogenetically closer to mammals.[6] Contrary to prior models, N-terminal truncation of an A-clade R2 protein to the domain structure of a D-clade R2 protein imposed surprisingly minimal impact on nuclease protection of target-site DNA or on TPRT *in vitro*, and even transgene insertion in cells was reduced only on the order of ~10-fold. The most striking functional perturbation of truncations that removed ZoAl ZF3 or ZF3-2 was the loss of a predictable rDNA position of 5′ junction formation. The deficit underlying this loss of precision of 5′ junction formation on the rDNA side could indirectly cause the increase in transgene 5′ truncation observed for ZoAl ΔZF3 and ΔZF3-2, due to a kinetic delay allowing degradation of the cDNA or more reliance on end-joining mechanisms that use internal rather than 3′-terminal cDNA positions for junction formation.

We show that A-clade R2 proteins have site-specific second-strand nicking activity, despite lacking detectable downstream target site recognition, a specificity reported to be necessary for *B. mori* R2 protein second-strand nicking.[20,22,23] For ZoAl protein, under some assay conditions used in this work, *in vitro* second-strand nicking was stimulated by ZF3-2. This

possible activity of ZF3-2 is consistent with the heterogeneity in rDNA position of 5′ junction formation observed for PRINT with ZoAl ΔZF3 or ΔZF3-2. Together these results generate a working hypothesis that the A-clade ZF3-2 domains stimulate second-strand nicking, potentially by binding upstream target-site DNA or through another mechanism such as recruitment of DNA repair factors. However, whether R2 protein is the predominant mediator of second-strand nicking for gene insertion in cells remains to be determined, as do specific role(s) of ZF3-2 in cells.

Overall, this work contributes insights about the protein-DNA interaction specificities and biochemical activities that support site-specific eukaryotic non-LTR retrotransposon mobility. The assays and conclusions developed in this work will have utility in future engineering of R2-derived proteins for gene-delivery applications.[21] For example, to improve or alter target-site specificity, our studies suggest that the combination of ZF1 and Myb domains should be used for optimization or directed-evolution assays. Also, results above indicate that upstream target-site binding is required but not sufficient for TPRT.

Limitations of the study

We assayed the DNA-binding specificity of R2 N-terminal domains separated from the remainder of the full-length protein; therefore, influences of the RT and EN domains on target-site interaction remain to be investigated. Also, given the gymnastics of R2 protein conformation required across the process of new gene synthesis, understanding the principles of target-site selection provides an incomplete inventory of protein-DNA interactions. Although our cellular assays of R2 protein function suggest a possible role for A-clade ZF3-2 domains in second-strand nicking or synthesis, whether R2 protein has second-strand nicking activity in cells remains to be determined. Furthermore, the absence of ZF3-2 could cause transgene 5′ junction heterogeneity by several mechanisms other than inhibition of second-strand nicking or synthesis, such as by failing to protect the upstream target-site region from exonucleolytic degradation.

STAR★METHODS

RESOURCE AVAILABILITY

Lead contact—Further information and requests for resources should be directed to the lead contact, Kathleen Collins (kcollins@berkeley.edu).

Materials availability—Plasmids used in this study will be available from AddGene or by request for constructs of less general utility.

Data and code availability

- Whole genome sequencing data have been deposited at NCBI Sequence Read Archive as SRR24873001, SRR24873002, and SRR24873003 and are publicly available as of the date of publication. Any additional data reported in this paper will be shared by the lead contact upon request.
- All original code has been deposited at Zenodo and is publicly available at DOI https://zenodo.org/doi/10.5281/zenodo.10439695 as of the date of publication.

- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Cell lines—HEK293T cells were grown in DMEM (Gibco) supplemented with 10% fetal bovine serum (FBS) (Avantor) and either 1x Pen/Strep (Gibco) or 10 μg/mL Primocin (InvivoGen) at 37°C, 5% CO_2. RPE-1 hTERT cells (human RPE-1 cells immortalized with an integrated virus expressing human telomerase reverse transcriptase) were grown in DMEM/F12 (Gibco) supplemented with 10% FBS and 10 μg/mL Primocin at 37°C, 5% CO_2. Cell lines were obtained from and authenticated by STR analysis at the UC Berkeley Cell Culture Facility.

METHOD DETAILS

Expression construct generation—Sequences used in this work are given in Table S3. Because the native R2 proteins' position of translation initiation remains unknown, and because different retrotransposon copies have sequence variation, the R2 protein sequences used in this work should not be considered definitive native retroelement proteins. BoMo and ZoAl synthetic ORFs and the ZoAl 3′UTR sequence were reported previously.[21] For DroSi, an identical *D. simulans* R2 retrotransposon DNA sequence was recovered from the Eickbush lab website (University of Rochester, NY; no longer maintained) and Repbase (R2_DSi, www.girinst.org). The translated amino acid sequence, starting after an in-frame methionine N-terminal to ZF1, was expressed using a codon-optimized synthetic ORF (GenScript). For TrCasB, first, an alignment of *T. castaneum* whole genome shotgun DNA sequences was used to derive a composite prediction of protein sequence and 3′UTR. Second, predicted protein sequence alignment with previously characterized active R2 proteins suggested reassignment of an alanine to proline to restore the conserved thumb domain motif PLKP (ALKP changed to PLKP). An amino acid sequence starting N-terminal to ZF1 was given a non-native methionine start codon for expression from a codon-optimized synthetic ORF (GenScript).

Bacterial protein expression: BoMo, DroSi, TrCasB, and ZoAl ORFs were ordered from GenScript. The initial N-terminal domain constructs were built by SLiCE cloning[48] into a pET vector with a C-terminal 6xHis tag (expression vector 2bct from UC Berkeley MacroLab) additionally modified to include an N-terminal MBP tag. N-terminal tagging of the DNA-binding domains does not interfere with full-length R2 protein function.[21] Truncations of the N-terminal constructs were generated by PCR-based mutagenesis. See Table S3 for amino acid sequences.

Mammalian protein expression: TrCasB and ZoAl full-length ORFs with an N-terminal 1xFLAG tag in mammalian expression vector pcDNA3.1(+) were ordered from GenScript. Truncations were generated by PCR-based mutagenesis. See Table S3 for amino acid sequences.

Recombinant protein expression and purification

Bacterial: Plasmids were transformed into chemically competent Rosetta2(DE3)pLysS cells. Cells were grown in 2xYT medium with ampicillin and chloramphenicol to OD_{600} = 0.6. The culture was chilled on ice for at least 20 min and then induced with 0.5 mM isopropylthio-β-galactoside (IPTG) (Gold Bio) at 16°C overnight. Cells were pelleted and lysed by sonication on ice for 3.5 min total in lysis buffer (20 mM Tris•HCl pH 7.5, 1 M NaCl, 20 mM imidazole pH 8, 1 mM $MgCl_2$, 10% glycerol, 0.2% Igepal CA-630 (USB Corporation), 1 mM dithiothreitol (DTT), 0.5 mg/mL lysozyme (Sigma-Aldrich), 0.2 mM phenylmethylsulfonyl fluoride (PMSF), Protease Inhibitor Cocktail (Sigma-Aldrich)). Lysate was cleared by centrifugation at 15,000×g for 20 min at 4°C.

A 2-step purification was employed. First, 500 μL of Ni-NTA agarose resin (Thermo Scientific) per 1 L initial bacterial culture was equilibrated in high salt Ni-NTA wash buffer (20 mM Tris•HCl pH 7.5, 2 M NaCl, 20 mM imidazole pH 8, 10% glycerol, 0.1% Igepal CA-630, 1 mM DTT). Lysate was incubated with resin for 3 h, rotating end-over-end at 4°C. After three washes for 10 min at 4°C in high salt Ni-NTA wash buffer and 1 wash for 5 min at 4°C in low salt Ni-NTA wash buffer (20 mM Tris•HCl pH 7.5, 800 mM NaCl, 20 mM imidazole pH 8, 10% glycerol, 0.1% Igepal CA-630, 1 mM DTT), protein was eluted in 5 mL Ni-NTA elution buffer (20 mM Tris•HCl pH 7.5, 800 mM NaCl, 350 mM imidazole pH 8, 10% glycerol, 0.1% Igepal CA-630, 1 mM DTT). Second, 4 mL amylose resin (NEB) was equilibrated in amylose equilibration buffer (20 mM Tris•HCl pH 7.5, 200 mM NaCl, 10% glycerol, 0.1% Igepal CA-630, 1 mM DTT). Ni-NTA eluate was diluted 4-fold in dilution buffer (20 mM Tris•HCl pH 7.5, 10% glycerol, 1 mM DTT) to a final salt concentration of 200 mM NaCl and incubated with amylose resin for 3 h, rotating end-over-end at 4°C. After 2 washes for 10 min at 4°C in low salt amylose wash buffer (20 mM Tris•HCl pH 7.5, 200 mM NaCl, 10% glycerol, 1 mM DTT) and 1 wash for 5 min at 4°C with high salt amylose wash buffer (20 mM Tris•HCl pH 7.5, 800 mM NaCl, 10% glycerol, 1 mM DTT), protein was eluted in 3 mL amylose elution buffer (20 mM Tris•HCl pH 7.5, 800 mM NaCl, 10% glycerol, 10 mM maltose, 0.1% Igepal CA-630, 1 mM DTT). Purified protein was snap-frozen in liquid nitrogen and stored at −80°C. Protein was purified to near-homogeneity, as validated by SDS-PAGE and Coomassie Blue staining. Concentration was determined using the Pierce BCA Protein Assay Kit (Thermo Scientific). Immediately prior to use in assays, protein was thawed and diluted to a working concentration in protein dilution buffer (20 mM Tris•HCl pH 7.5, 800 mM NaCl, 50% glycerol, 1 mM DTT).

Mammalian: Plasmids were reverse transfected into HEK293T cells at ~80% confluency. Cells were washed with 1x DPBS (Gibco), trypsinized with 0.05% Trypsin-EDTA (Gibco) and replated in DMEM with 10% FBS and either 1x Pen/Strep or 10 μg/mL Primocin. Before cell re-attachment, each 100 mm plate was transfected with 12 μg plasmid DNA using Lipofectamine 3000 (Invitrogen) according to the manufacturer's protocol.

After 24 h, cells were harvested by hypotonic freeze-thaw lysis. Cells were trypsinized, washed once with chilled 1x DPBS with 1 mM PMSF, and resuspended in 4x pellet volume 1x hypotonic lysis buffer (20 mM HEPES pH 8, 2 mM $MgCl_2$, 0.2 mM EGTA, 10% glycerol, 1 mM DTT, 1 mM PMSF, 0.4% Protease Inhibitor Cocktail). After a 5

min incubation on ice, cells were snap-frozen in liquid nitrogen and then thawed at room temperature three times. Samples were brought up to 400 mM NaCl by addition of 5 M NaCl, gently mixed, and incubated on ice for 5 min. Lysate was cleared by centrifugation at 17,000×g for 5 min at 4°C. The supernatant was transferred to a new tube, and an equal volume of 1x hypotonic lysis buffer with 0.2% Igepal CA-630 was added. The lysate was centrifuged again at 17,000×g for 5 min at 4°C. The supernatant was transferred to a new tube.

20 μL of FLAG resin (Sigma-Aldrich) per 100 mm plate was equilibrated in immunoprecipitation (IP) buffer (20 mM HEPES pH 8, 2 mM $MgCl_2$, 0.2 mM EGTA, 10% glycerol, 1 mM DTT, 1 mM PMSF, 0.4% Protease Inhibitor Cocktail, 0.1% Igepal CA-630, 200 mM NaCl). Lysate was incubated with resin for 2 h, rotating end-over-end at room temperature. After 2 quick washes and 2 5-min washes at room temperature in IP buffer, protein was eluted in 40 μL FLAG elution buffer (IP buffer with 50 ng/μL 3xFLAG peptide (Sigma-Aldrich)).

Purified protein was snap-frozen in liquid nitrogen and stored at −80°C. Purified protein was resolved by SDS-PAGE and visualized by immunoblot using 0.45 μM nitrocellulose membranes (Bio-Rad) blocked for 1 h at room temperature in blocking buffer (1x TBST (10 mM Tris•HCl pH 7.5, 150 mM NaCl, 0.2% Tween 20, 0.02% sodium azide) with 5% bovine serum albumin (BSA) (Sigma-Aldrich)). The membrane was probed in blocking buffer with 1:3000 (v/v) anti-FLAG primary antibody (Sigma-Aldrich) then 1:2000 (v/v) Alexa Fluor 680 goat anti-mouse secondary antibody (Invitrogen). The membrane was visualized with an LI-COR Odyssey CLx imager.

Radiolabeling and annealing oligonucleotides—Oligonucleotides were radiolabeled with T4 polynucleotide kinase (NEB) and ATP, [γ−32P] (PerkinElmer) according to the manufacturer's protocol. Radiolabeling reactions were column purified twice with ProbeQuant G-50 Micro Columns (Cytiva). Oligonucleotide annealing conditions were 85°C for 3 min, −1 °C/s/second to 16°C, 10°C hold.

EMSA—EMSA reactions were assembled on ice (0.1 mg/mL BSA, 0 or 500 ng poly[d(I-C)] (Roche), 80 or 160 mM NaCl, 25 mM Tris•HCl pH 7.5, 1 mM DTT, 1 mM $MgCl_2$, 0.5% CHAPS (Sigma-Aldrich), 0.04% bromophenol blue (w/v)). R2 protein was spiked in, and the reaction was incubated for 20 min on ice. Target-site DNA was then spiked in and the reaction was incubated for 20 min on ice. Reactions with 500 ng poly[d(I-C)] had 160 mM NaCl, 2 nM DNA, and 0, 1, 5, 25, 125, 625, 1250, or 2000 nM R2 protein purified from *E. coli*. Reactions with 0 ng poly[d(I-C)] had 80 mM NaCl, 0.5 nM DNA, and 0, 1, 2, or 4 μM R2 protein purified from *E. coli*. The final reaction volume was 10 μL. Reactions were loaded onto a 37.5:1 acrylamide:bisacrylamide 5% native PAGE gel and electrophoresed at 250 V for 2 h at 4°C. Gels were vacuum-dried at 80°C, exposed onto a phosphor imaging plate (Cytiva), and imaged on an Amersham Typhoon Biomolecular Imager.

DNase I footprinting—Oligonucleotides for footprinting were purified by denaturing PAGE size selection prior to use. Reactions were assembled on ice (0.1 mg/mL BSA, 500 ng poly[d(I-C)], 25 mM Tris•HCl pH 7.5, 1 mM DTT, 1 mM $MgCl_2$, 0.5% CHAPS, 0.04%

(w/v) bromophenol blue). R2 protein purified from *E. coli* was spiked in, and reactions were incubated for 20 min on ice. Target-site DNA was then spiked in, and reactions were incubated for 20 min on ice. The reaction volume was 10 μL. Reactions were digested with 1 μL DNase I (NEB) for 1 min at room temperature and stopped by addition of 100 μL of stop buffer (400 mM NaOAc pH 5.2, 0.2% SDS, 10 mM EDTA, 10 μg/mL proteinase K (NEB) (added just before use)). Reactions were incubated for 15 min at 55°C, extracted twice with 1x reaction volume phenol:chloroform:isoamyl alcohol (25:24:1, v/v) (PCI) (Invitrogen), and precipitated for 30 min in a dry ice ethanol bath using 3x reaction volume 100% ethanol with 2 μL 10 mg/mL glycogen as a carrier (Sigma-Aldrich). Samples were pelleted by centrifugation at 17,000×g for 15 min at room temperature. Pellets were resuspended in formamide gel-loading buffer (95% deionized formamide, 0.025% (w/v) bromophenol blue, 0.025% (w/v) xylene cyanol),[49] heated for 2 min at 95°C, cooled on ice, and then resolved by denaturing PAGE. Gels were vacuum-dried at 80°C, exposed onto a phosphor imaging plate, and imaged on an Amersham Typhoon Biomolecular Imager.

RNA production

***Construct generation for* in vitro *transcription (IVT)*:** The ZoAl and TrCasB 3′ UTRs preceded by the T7 RNA polymerase promoter were ordered in the pUC57mini vector from GenScript. Transcription vectors for ZoAl ORF mRNA and GFP template RNA were previously described.[21] See Table S3 for RNA sequences.

IVT: RNAs for TPRT and transgene insertion assays were generated by IVT. IVT templates were created by PCR or by linearizing plasmid with digestion using 5 μL BbsI (NEB) per 5 μg plasmid at 37°C for 4 h. To confirm digestion, products were resolved on an agarose gels containing ethidium bromide and imaged on a Bio-Rad Gel Doc XR+. Digestion reactions were column-purified with the QIAquick PCR purification kit (Qiagen) and eluted in nuclease-free water. RNAs were transcribed with the HiScribe T7 High Yield RNA Synthesis kit (NEB). To generate capped mRNAs with modified nucleotides, R2 ORF mRNAs were transcribed using CleanCap Reagent AG (TriLink) and m1-pseudouridine (TriLink) according to the manufacturer's protocol. RNA synthesis reactions were digested with 2 μL DNase I per reaction at 37°C for 30 min. RNAs were column-purified with a ProbeQuant G-50 Micro Column and extracted with 1x reaction volume PCI. RNAs were precipitated by addition of LiCl to a final concentration of 2.5 M (RNAs for transfection) or NaOAc pH 5.2 to a final concentration of 0.3 M (RNAs for biochemical assays), addition of 3x reaction volume 100% ethanol, incubation in liquid nitrogen for 30 min, then centrifugation at 17,000×g for 30 min at 4°C. RNA pellets were washed 3 times with 70% ethanol, air-dried for 15 min, and resuspended in 1 mM sodium citrate, pH 6.5. Concentrations were quantified on a NanoDrop 1000 Spectrophotometer. RNA quality was assessed by visualizing RNAs by denaturing urea-PAGE, staining with SYBR Gold (Invitrogen), and imaging on an Amersham Typhoon Biomolecular Imager. RNAs were stored at −80°C until use.

TPRT and second-strand nicking assays—TPRT and second-strand nicking assays were performed as described elsewhere.[21] Briefly, reactions were assembled on ice (25 mM Tris•HCl pH 7.5, 75 mM KCl, 5 mM $MgCl_2$, 10 mM DTT, 2% PEG-6K, 0.6 μM

3′ UTR RNA, 0.5 mM dNTPs, 2 μL R2 protein purified from HEK293T cells, 0.025 μM radiolabeled target-site DNA duplex). Standard TPRT and second-strand nicking assay reactions were incubated at 37°C for 15 and 30 min, respectively. Reactions were heat inactivated at 70°C for 5 min. 2 μL 10 mg/mL RNase A (Thermo Scientific) was added and the reaction was incubated at 55°C for 30 more minutes. The RNase A digestion was stopped by addition of 80 μL stop solution (50 mM Tris•HCl pH 7.5, 20 mM EDTA, 0.2% SDS) mixed with a 100 nt 5′-end radiolabeled loading control. Reactions were extracted with 1x reaction volume PCI. Nucleic acids were precipitated for 30 min in a dry ice ethanol bath using NH_4OAc pH 7 added to a final concentration of 0.975 M and 3x reaction volume 100% ethanol with 1 μL 10 mg/mL glycogen as a carrier. Samples were pelleted by centrifugation at 17,000×g for 15 min at room temperature. Pellets were resuspended in formamide gel-loading dye, heated for 2 min at 95°C, cooled on ice, and then resolved by denaturing PAGE. Gels were vacuum-dried at 80°C, exposed onto a phosphor imaging plate, and imaged on an Amersham Typhoon Biomolecular Imager.

Transgene insertion assay—Transgene insertion assays were performed as described elsewhere.[21] In brief, 2 RNAs, an mRNA encoding the R2 protein and a template RNA, were co-reverse transfected into log-phase (30–50% confluency) RPE-1 hTERT cells. Transfections were done in triplicate. Cells were washed with 1x DPBS, trypsinized with 0.05% Trypsin-EDTA, and replated in DMEM/F12 with 10% FBS. 0.75–1 million cells were plated in each well of a 6-well plate. Before cell re-attachment, each well of the 6-well plate was transfected with 1.5 μg of total RNA (template RNA:R2 ORF mRNA molar ratio of 3:1) using MessengerMAX (Invitrogen) according to the manufacturer's protocol. Cells were harvested the next day (approximately 24 h later) for flow cytometry and cell sorting.

Flow cytometry and cell sorting—Cells were washed with 1x PBS and trypsinized with 0.05% Trypsin-EDTA. DMEM/F12 with 5% FBS was added to inactivate Trypsin, to a total volume of 1 mL. Cells were vigorously re-pipetted to break up clumps and transferred to a 5 mL flow cytometry tube.

For flow cytometry analysis, samples were analyzed on the Attune NxT Flow 35 Cytometer (Thermo Fisher) with voltage settings FSC 70V, SSC 280V, and BL1 250V. Flow data was analyzed using FlowJo (10.8.1). Gates were determined using the template RNA alone sample as a negative control. Percentage GFP positive and median GFP intensity values measured for the template RNA alone negative control were subtracted from measures of experimental samples.

For cell sorting, samples were analyzed on the Sony SH800S Cell Sorter with 488 nm and 561 nm lasers. Samples were sorted in normal sorting mode with 130 μm sorting chips (Sony Biotechnology).

Genomic DNA purification and PCR—Cells were washed with 1x DPBS, pelleted by centrifugation at 7000×g for 3 min, snap-frozen in liquid nitrogen, and stored at −80°C. Cell pellets were thawed on ice and resuspended in 200 μL RIPA lysis buffer (150 mM NaCl, 50 mM Tris•HCl pH 7.5, 1 mM EDTA, 1% Triton X-100, 0.5% sodium deoxycholate, 0.1% SDS, 1 mM DTT). 10 μL 10 mg/mL RNase A was added. The samples were vortexed and

incubated at 37°C for at least 30 min. 5 μL 20 mg/mL proteinase K was added. The lysates were incubated at 50°C overnight. Reactions were extracted twice with 1x reaction volume PCI. Using wide bore pipette tips, nucleic acids were precipitated for at least 2h at −20°C by addition of 20 μL 5 M NaCl pH 5 and 3x reaction volume 100% ethanol. Genomic DNA was pelleted by centrifugation at 18,000×g for 30 min at 4°C. Pellets were washed twice with 75% ethanol, air-dried for 20 min, and allowed to resuspend in nuclease-free H_2O overnight at room temperature prior to concentration quantification on a NanoDrop 1000 Spectrophotometer.

Purified genomic DNA was used for PCR detection of insertion junctions. 100 ng genomic DNA was used per 25 μL Q5 PCR (NEB) reaction, assembled according to the manufacturer's protocol. Junction touchdown PCR cycling conditions were 90°C for 3 min; 5 cycles of 98°C for 10 s, 65°C (−1°C/cycle) for 30 s, 72°C for 15 s; 25 cycles of 98°C for 10 s, 60°C for 30 s, 72°C for 15 s; 72°C for 20 s; 4°C forever. GFP ORF touchdown PCR cycling conditions were 90°C for 3 min; 5 cycles of 98°C for 10 s, 65°C (−1°C/cycle) for 30 s, 72°C for 25 s; 25 cycles of 98°C for 10 s, 60°C for 30 s, 72°C for 25 s; 72°C for 20 s; 4°C forever. Products were resolved on agarose gels containing ethidium bromide and imaged on a Bio-Rad Gel Doc XR+.

Transgene copy number analysis by ddPCR was performed exactly as described elsewhere[21] with *RPP30* used as a reference gene. Briefly, genomic DNA was digested overnight with BamHI and XmnI (NEB), and ddPCR reactions were assembled according to the manufacturer's protocol using ddPCR supermix without dUTP (Bio-Rad). The reaction was transferred to a DG8 cartridge (Bio-Rad). Droplet generation oil (Bio-Rad) was added, and droplets were generated with a Bio-Rad QX200 Droplet Generator, transferred to a 96-well plate, and thermal cycled according to the manufacturer's protocol. All PCR primer sequences are listed in Table S2.

Whole genome sequencing and bioinformatic analysis—Genomic DNA samples were submitted to the Vincent J. Coates Genomics Sequencing Lab at UC Berkeley for 30x coverage whole genome shotgun sequencing. Briefly, genomic DNA was sheared to 400–500 bp with Covaris tubes for Illumina library preparation. PE150 sequencing was performed on a NovaSeq 6000 instrument with an S4 flow cell. Bioinformatic analyses were performed on the Berkeley Research Computing Savio cluster with SLURM job scheduling or on an Apple M1 Max processor.

Analyses were performed largely as described previously.[21] Briefly, PCR and optical duplicates were removed with BBMap v38.97[39] and reads were trimmed for quality with Trimmomatic v0.39.[40] Reads shorter than 36 bp or with an overall PHRED quality less than 30 were discarded. All alignments were performed with bwa mem v0.7.17[38] using default parameters. Reads were first aligned to a transgene reference with 840 bp of flanking rDNA. Unmapped mates and clipped portions of reads were mapped to a complete rDNA scaffold (GenBank KY962518.1). Mates and clips remaining unaligned were then aligned to the human genome reference (T2T-CHM13v2.0). Finally, still-unaligned portions were checked for alignment to the context surrounding the 28S insertion site with regex (Python fuzzysearch). Reads without both mates mapped and reads aligning better to the human

genome than to the transgene were discarded. Additionally, reads mapping to a curated list of contaminants, including viral genomes, arising from pooled sequencing were discarded.

Reads spanning the 3′ transgene junction were used to classify site specificity of insertions. If 3′ junction-spanning reads mapped to the anticipated 28S nick site +/− 3 bp, the insertion was classified as “on-target.” If the downstream portion mapped to rDNA outside of the 28S target site, insertions were classified as “rDNA off-target.” All other insertions were classified as “genomic off-target.”

5′ junctions were classified as follows.

1. “Anneal” junctions are characterized by full-length transgene insertions fused to upstream rDNA without deletion or duplication of upstream rDNA, as would be expected from annealing 28 nt of homologous cDNA to upstream rDNA.
2. “Join” junctions contain downstream sequence mapping to the transgene and upstream sequence mapping to rDNA on the same strand. For full-length transgenes, join junctions can result in partial duplication of 28 nt of homologous rDNA sequence.
3. “Snap-back” junctions contain sequence mapping to the opposite than expected strand of the transgene or rDNA scaffold.
4. “Extra template” junctions contain downstream sequence mapping to the transgene and upstream sequence mapping to a cellular RNA sequence consistent with insertion by reverse transcription. Sequences in this category were manually evaluated using NCBI BLAST.[50]
5. “Other” junctions are characterized by upstream sequence mapping somewhere in the genome other than rDNA.

Transgene 5′ junction reads with upstream sequence not mapping anywhere in the genome were not classified. Overall WGS coverage was determined by aligning read pairs to the T2T-CHM13v2.0 human reference genome and calculating mean read depth with samtools depth.

rDNA join positions (i.e., the first base in a 5′ junction-spanning read not mapping to the transgene) were determined for “join” category junctions only. Statistical significance in the difference of median rDNA join positions was determined by bootstrap hypothesis testing: 1 million paired samples were selected with replacement from the joint distribution and the observed difference in medians was compared to the bootstrapped distribution of differences to obtain a *p* value.

Visualizations—Multiple sequence alignments were performed with Clustal Omega[51] and visualized with Jalview.[34] Predicted structures were generated with ColabFold using default settings.[36] Structural alignments were performed and visualized in UCSF ChimeraX.[35]

QUANTIFICATION AND STATISTICAL ANALYSIS

Exact measures of center, precision, or spread, statistical tests used, and numbers of technical replicates (n) are listed in figure legends and results wherever used. If not noted, all results were replicated at least twice. Significance was defined as $p \leq 0.05$. For EMSA quantifications, quantification of free DNA was performed in ImageJ.[37] Binding curves were fitted for 1 site-specific binding in GraphPad Prism. Statistical tests were performed using GraphPad Prism, SciPy,[44] or custom analyses.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank Xiaozhu Zhang for leading development of the PRINT transgene insertion method and Prof. Shawn Christensen (University of Texas at Arlington) for the G + A ladder protocol. We thank Alison Killilea, Willie Hercule, and Mahiya Ellis (UC Berkeley Cell Culture Facility) for cell stocks; the Innovative Genomics Institute for flow cytometer use; and the QB3 Functional Genomics Laboratory and Vincent J. Coates Genomics Sequencing Laboratory for Illumina library production and sequencing. Parts of Figure 6A and the graphical abstract were created with BioRender.com by B.V.T. This work was supported by NIH Pioneer Award DP1 HL156819 (K.C.) with postdoctoral training support from NIH F32 GM139306 (B.V.T.) and predoctoral training support from NIH T32 GM07232 and the Shurl and Kay Curci Foundation (C.A.H.).

REFERENCES

1. Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvák Z, Levin HL, Macfarlan S, et al. (2018). Ten things you should know about transposable elements. Genome Biol. 19, 199. 10.1186/s13059-018-1577-z. [PubMed: 30454069]
2. Malik HS, Burke WD, and Eickbush TH (1999). The age and evolution of non-LTR retrotransposable elements. Mol. Biol. Evol. 16, 793–805. 10.1093/oxfordjournals.molbev.a026164. [PubMed: 10368957]
3. Kapitonov VV, Tempel S, and Jurka J (2009). Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. Gene 448, 207–213. 10.1016/j.gene.2009.07.019. [PubMed: 19651192]
4. Han JS (2010). Non-long terminal repeat (non-LTR) retrotransposons: mechanisms, recent developments, and unanswered questions. Mobile DNA 1, 15. 10.1186/1759-8753-1-15. [PubMed: 20462415]
5. Bao W, Kojima KK, and Kohany O (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA 6, 11. 10.1186/s13100-015-0041-9. [PubMed: 26045719]
6. Kojima KK, Seto Y, and Fujiwara H (2016). The Wide Distribution and Change of Target Specificity of R2 Non-LTR Retrotransposons in Animals. PLoS One 11, e0163496. 10.1371/journal.pone.0163496. [PubMed: 27662593]
7. Kazazian HH, and Moran JV (2017). Mobile DNA in Health and Disease. EnglandN. Engl. J. Med. 377, 361–370. 10.1056/NEJMra1510092.
8. Kojima KK (2018). Human transposable elements in Repbase: genomic footprints from fish to humans. Mobile DNA 9, 2. 10.1186/s13100-017-0107-y. [PubMed: 29308093]
9. Balachandran P, Walawalkar IA, Flores JI, Dayton JN, Audano PA, and Beck CR (2022). Transposable element-mediated rearrangements are prevalent in human genomes. Nat. Commun. 13, 7115. 10.1038/s41467-022-34810-8. [PubMed: 36402840]
10. Konkel MK, and Batzer MA (2010). A mobile threat to genome stability: The impact of non-LTR retrotransposons upon the human genome. Semin. Cancer Biol. 20, 211–221. 10.1016/j.semcancer.2010.03.001. [PubMed: 20307669]

11. Payer LM, and Burns KH (2019). Transposable elements in human genetic disease. Nat. Rev. Genet. 20, 760–772. 10.1038/s41576-019-0165-8. [PubMed: 31515540]
12. Cordaux R, and Batzer MA (2009). The impact of retrotransposons on human genome evolution. Nat. Rev. Genet. 10, 691–703. 10.1038/nrg2640. [PubMed: 19763152]
13. Luan DD, Korman MH, Jakubczak JL, and Eickbush TH (1993). Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: A mechanism for non-LTR retrotransposition. Cell 72, 595–605. 10.1016/0092-8674(93)90078-5. [PubMed: 7679954]
14. Eickbush TH, and Eickbush DG (2015). Integration, regulation, and long-term stability of R2 retrotransposons. Microbiol. Spectr. 3, MDNA3–2014. 10.1128/microbiolspec.MDNA3-0011-2014.
15. Ye J, and Eickbush TH (2006). Chromatin Structure and Transcription of the R1- and R2-Inserted rRNA Genes of Drosophila melanogaster. Mol. Cell Biol. 26, 8781–8790. 10.1128/MCB.01409-06. [PubMed: 17000772]
16. Kojima KK, and Fujiwara H (2005). Long-Term Inheritance of the 28S rDNA-Specific Retrotransposon R2. Mol. Biol. Evol. 22, 2157–2165. 10.1093/molbev/msi210. [PubMed: 16014872]
17. Luchetti A, and Mantovani B (2013). Non-LTR R2 Element Evolutionary Patterns: Phylogenetic Incongruences, Rapid Radiation and the Maintenance of Multiple Lineages. PLoS One 8, e57076. 10.1371/journal.pone.0057076. [PubMed: 23451148]
18. Mahbub MM, Chowdhury SM, and Christensen SM (2017). Globular domain structure and function of restriction-like-endonuclease LINEs: similarities to eukaryotic splicing factor Prp8. Mobile DNA 8, 16. 10.1186/s13100-017-0097-9. [PubMed: 29151899]
19. Khadgi BB, Govindaraju A, and Christensen SM (2019). Completion of LINE integration involves an open '4-way' branched DNA intermediate. Nucleic Acids Res. 47, 8708–8719. 10.1093/nar/gkz673. [PubMed: 31392993]
20. Christensen SM, and Eickbush TH (2005). R2 Target-Primed Reverse Transcription: Ordered Cleavage and Polymerization Steps by Protein Subunits Asymmetrically Bound to the Target DNA. Mol. Cell Biol. 25, 6617–6628. 10.1128/MCB.25.15.6617-6628.2005. [PubMed: 16024797]
21. Zhang X, Van Treeck B, Horton CA, McIntyre JJR, Palm SM, Shumate JL, and Collins K (2024). Harnessing eukaryotic retroelement proteins for transgene insertion into human safe-harbor loci using RNA-only delivery. Nat. Biotechnol. 10.1038/s41587-024-02137-y.
22. Christensen S, and Eickbush TH (2004). Footprint of the Retrotransposon R2Bm Protein on its Target Site Before and After Cleavage. J. Mol. Biol. 336, 1035–1045. 10.1016/j.jmb.2003.12.077. [PubMed: 15037067]
23. Christensen SM, Ye J, and Eickbush TH (2006). RNA from the 5′ end of the R2 retrotransposon controls R2 protein binding to and cleavage of its DNA target site. Proc. Natl. Acad. Sci. USA 103, 17602–17607. 10.1073/pnas.0605476103. [PubMed: 17105809]
24. Christensen SM, Bibillo A, and Eickbush TH (2005). Role of the Bombyx mori R2 element N-terminal domain in the target-primed reverse transcription (TPRT) reaction. Nucleic Acids Res. 33, 6461–6468. 10.1093/nar/gki957. [PubMed: 16284201]
25. Wilkinson ME, Frangieh CJ, Macrae RK, and Zhang F (2023). Structure of the R2 non-LTR retrotransposon initiating target-primed reverse transcription. Science 380, 301–308. 10.1126/science.adg7883. [PubMed: 37023171]
26. Deng P, Tan S-Q, Yang Q-Y, Fu L, Wu Y, Zhu H-Z, Sun L, Bao Z, Lin Y, Zhang QC, et al. (2023). Structural RNA components supervise the sequential DNA cleavage in R2 retrotransposon. Cell 186, 2865–2879.e20. 10.1016/j.cell.2023.05.032. [PubMed: 37301196]
27. Shivram H, Cawley D, and Christensen SM (2011). Targeting novel sites. Mobile Genet. Elem. 1, 169–178. 10.4161/mge.1.3.18453.
28. Thompson BK, and Christensen SM (2011). Independently derived targeting of 28S rDNA by Aand D-clade R2 retrotransposons. Mobile Genet. Elem. 1, 29–37. 10.4161/mge.1.1.16485.
29. Zhang X, and Eickbush TH (2005). Characterization of Active R2 Retrotransposition in the rDNA Locus of Drosophila simulans. Genetics 170, 195–205. 10.1534/genetics.104.038703. [PubMed: 15781697]

30. Bibillo A, and Eickbush TH (2004). End-to-End Template Jumping by the Reverse Transcriptase Encoded by the R2 Retrotransposon. J. Biol. Chem. 279, 14945–14953. 10.1074/jbc.M310450200. [PubMed: 14752111]

31. George JA, Burke WD, and Eickbush TH (1996). Analysis of the 5′ Junctions of R2 Insertions With the 28S Gene: Implications for Non-LTR Retrotransposition. Genetics 142, 853–863. 10.1093/genetics/142.3.853. [PubMed: 8849892]

32. Sefer A, Kallis E, Eilert T, Röcker C, Kolesnikova O, Neuhaus D, Eustermann S, and Michaelis J. (2022). Structural dynamics of DNA strand break sensing by PARP-1 at a single-molecule level. Nat. Commun. 13, 6569. 10.1038/s41467-022-34148-1. [PubMed: 36323657]

33. Mackey ZB, Niedergang C, Murcia JM, Leppard J, Au K, Chen J, de Murcia G, and Tomkinson AE (1999). DNA Ligase III Is Recruited to DNA Strand Breaks by a Zinc Finger Motif Homologous to That of Poly(ADP-ribose) Polymerase: IDENTIFICATION OF TWO FUNCTIONALLY DISTINCT DNA BINDING REGIONS WITHIN DNA LIGASE III. J. Biol. Chem. 274, 21679–21687. 10.1074/jbc.274.31.21679. [PubMed: 10419478]

34. Waterhouse AM, Procter JB, Martin DMA, Clamp M, and Barton GJ (2009). Jalview Version 2−a multiple sequence alignment editor and analysis workbench. Bioinformatics 25, 1189–1191. 10.1093/bioinformatics/btp033. [PubMed: 19151095]

35. Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, Morris JH, and Ferrin TE (2021). UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein Sci. 30, 70–82. 10.1002/pro.3943. [PubMed: 32881101]

36. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, and Steinegger M. (2022). ColabFold: making protein folding accessible to all. Nat. Methods 19, 679–682. 10.1038/s41592-022-01488-1. [PubMed: 35637307]

37. Schneider CA, Rasband WS, and Eliceiri KW (2012). NIH Image to ImageJ: 25 years of image analysis. Nat. Methods 9, 671–675. 10.1038/nmeth.2089. [PubMed: 22930834]

38. Li H, and Durbin R (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 26, 589–595. 10.1093/bioinformatics/btp698. [PubMed: 20080505]

39. Bushnell B (2014). BBMap: A Fast, Accurate, Splice-Aware Aligner. Conference: 9th Annual Genomics of Energy & Environment Meeting. https://www.osti.gov/servlets/purl/1241166.

40. Bolger AM, Lohse M, and Usadel B (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. 10.1093/bioinformatics/btu170. [PubMed: 24695404]

41. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, and Li H (2021). Twelve years of SAMtools and BCFtools. GigaScience 10, giab008. 10.1093/gigascience/giab008. [PubMed: 33590861]

42. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, et al. (2020). Array programming with NumPy. Nature 585, 357–362. 10.1038/s41586-020-2649-2. [PubMed: 32939066]

43. McKinney W (2010). Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference, pp. 56–61. 10.25080/Majora-92bf1922-00a.

44. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods 17, 261–272. 10.1038/s41592-019-0686-2. [PubMed: 32015543]

45. Hunter JD (2007). Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng. 9, 90–95. 10.1109/MCSE.2007.55.

46. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, and de Hoon MJL (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25, 1422–1423. 10.1093/bioinformatics/btp163. [PubMed: 19304878]

47. Waskom M (2021). seaborn: statistical data visualization. J. Open Source Softw. 6, 3021. 10.21105/joss.03021.

48. Zhang Y, Werling U, and Edelmann W (2014). Seamless Ligation Cloning Extract (SLiCE) Cloning Method. Methods Mol. Biol. 1116, 235–244. 10.1007/978-1-62703-764-8_16. [PubMed: 24395368]

49. Formamide Gel-Loading Buffer (2013). Cold Spring Harb Protoc. 2013, pp. Pdb.rec073510. 10.1101/pdb.rec073510.

50. Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ (1990). Basic local alignment search tool. J. Mol. Biol. 215, 403–410. 10.1016/S0022-2836(05)80360-2. [PubMed: 2231712]

51. Madeira F, Pearce M, Tivey ARN, Basutkar P, Lee J, Edbali O, Madhusoodanan N, Kolesnikov A, and Lopez R (2022). Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Res. 50, W276–W279. 10.1093/nar/gkac240. [PubMed: 35412617]

Highlights

- R2 protein N-terminal domains have clade-specific DNA interaction properties
- D-clade R2 protein N-terminal domains bind multiple target DNA sequences
- Only universal DNA-binding domains are essential for new gene insertion
- A-clade R2 protein N-terminal domains increase efficiency and precision of new gene insertion

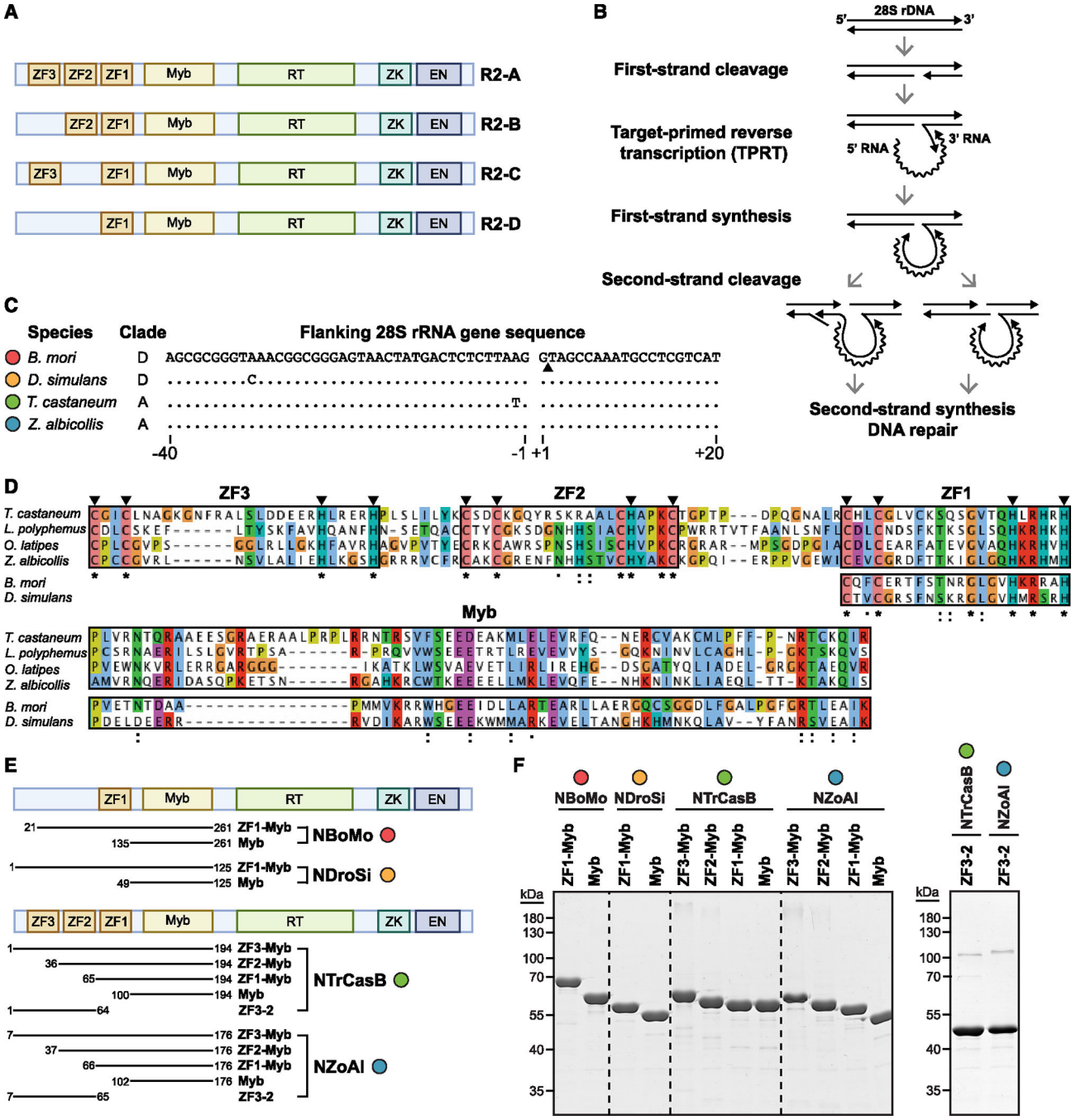


Figure 1. Design and purification of R2 N-terminal-region proteins for DNA-binding assays

(A) Domain schematic of R2 retrotransposon clades A–D. ZF, zinc finger; RT, reverse transcriptase; ZK, zinc knuckle; EN, endonuclease. Domains are not drawn to scale.

(B) Schematic of a new R2 retrotransposon insertion into the 28S rDNA target site. Solid lines denote DNA, and squiggly lines denote RNA. Arrowheads indicate strand 3′ end.

(C) Target site and flanking 28S rDNA sequences from selected species of D- and A-clade R2 retrotransposons relevant for this study. A dot denotes the same nt as in *B. mori*. The conserved position of the first-strand nick is denoted with a black triangle. The numbering

schematic places 0 as the phosphodiester bond at the center of the *B. mori* R2 protein first- and second-strand nick sites and is negative upstream or positive downstream. The color scheme used here (red for BoMo, orange for DroSi, green for TrCasB, and blue for ZoAl) is maintained throughout the figures.

(D) Amino acid sequence alignment of the N-terminal regions of selected D- and A-clade R2 retrotransposons. Black triangles indicate conserved zinc-coordinating residues. Color scheme and characters follow Clustal X convention: an asterisk (*) indicates all residues are identical, a colon (:) indicates conserved substitutions, and a period (.) indicates semi-conserved substitutions.

(E) Schematic for N-terminal-region proteins with amino acid numbering. Schematic is not to scale.

(F) Coomassie blue-stained sodium dodecyl sulfate-PAGE (SDS-PAGE) gel of N-terminal-region proteins purified from *E. coli*. For all gel images shown, an unbroken line bounds samples in the same gel that have the same image contrast settings. A dashed line separates lanes of the same gel, sometimes with removal of empty lanes between samples.

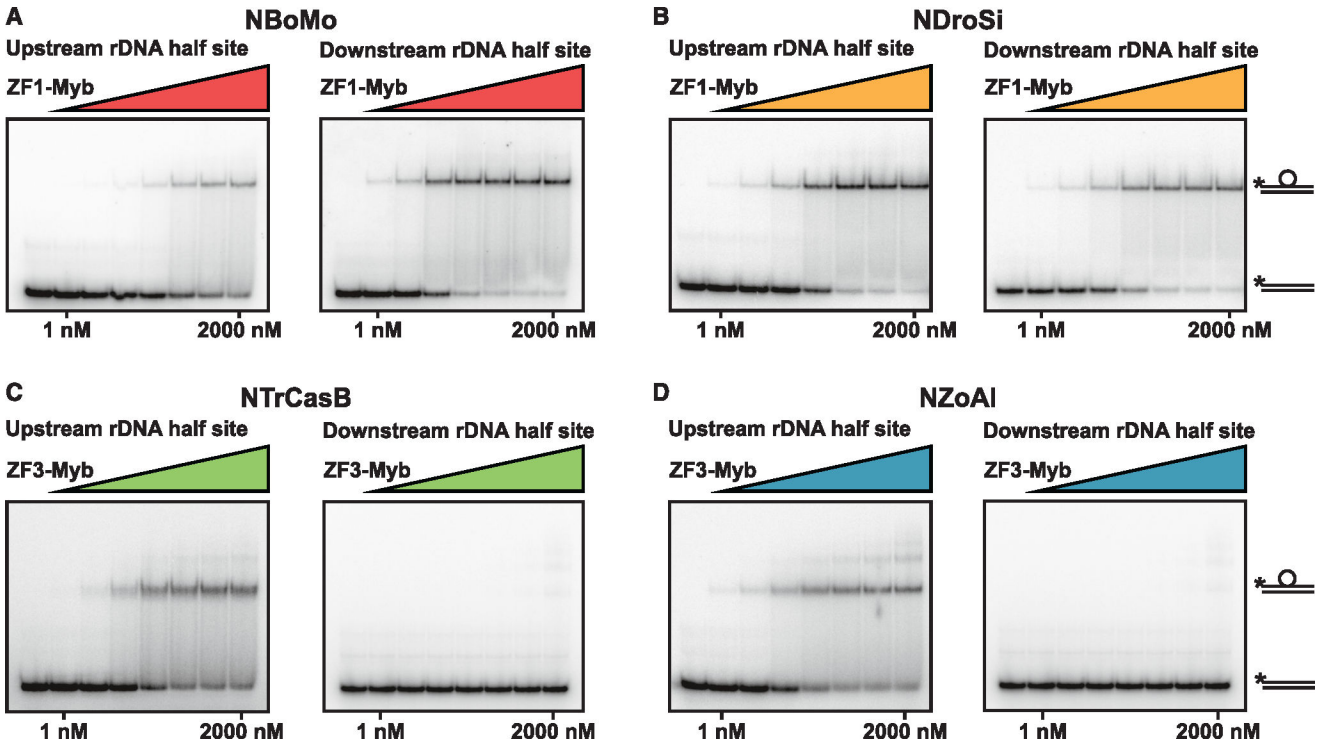


Figure 2. Clade-specific DNA interaction of R2 protein N-terminal domains

Images of EMSA native PAGE gels. Symbols on the far right indicate migration of free versus bound radiolabeled DNA. The circle represents protein, and the straight lines indicate DNA duplex. An asterisk indicates the radiolabeled strand. NBoMo ZF1-Myb (A) and NDroSi ZF1-Myb (B) were tested with upstream target half-site (−50 to −9) on the left or downstream target half-site (−8 to +34) on the right. NTrCasB ZF3-Myb (C) and NZoAl ZF3-Myb (D) were tested with upstream target half-site (−50 to −1) on the left or downstream target half-site (+1 to +50) on the right.

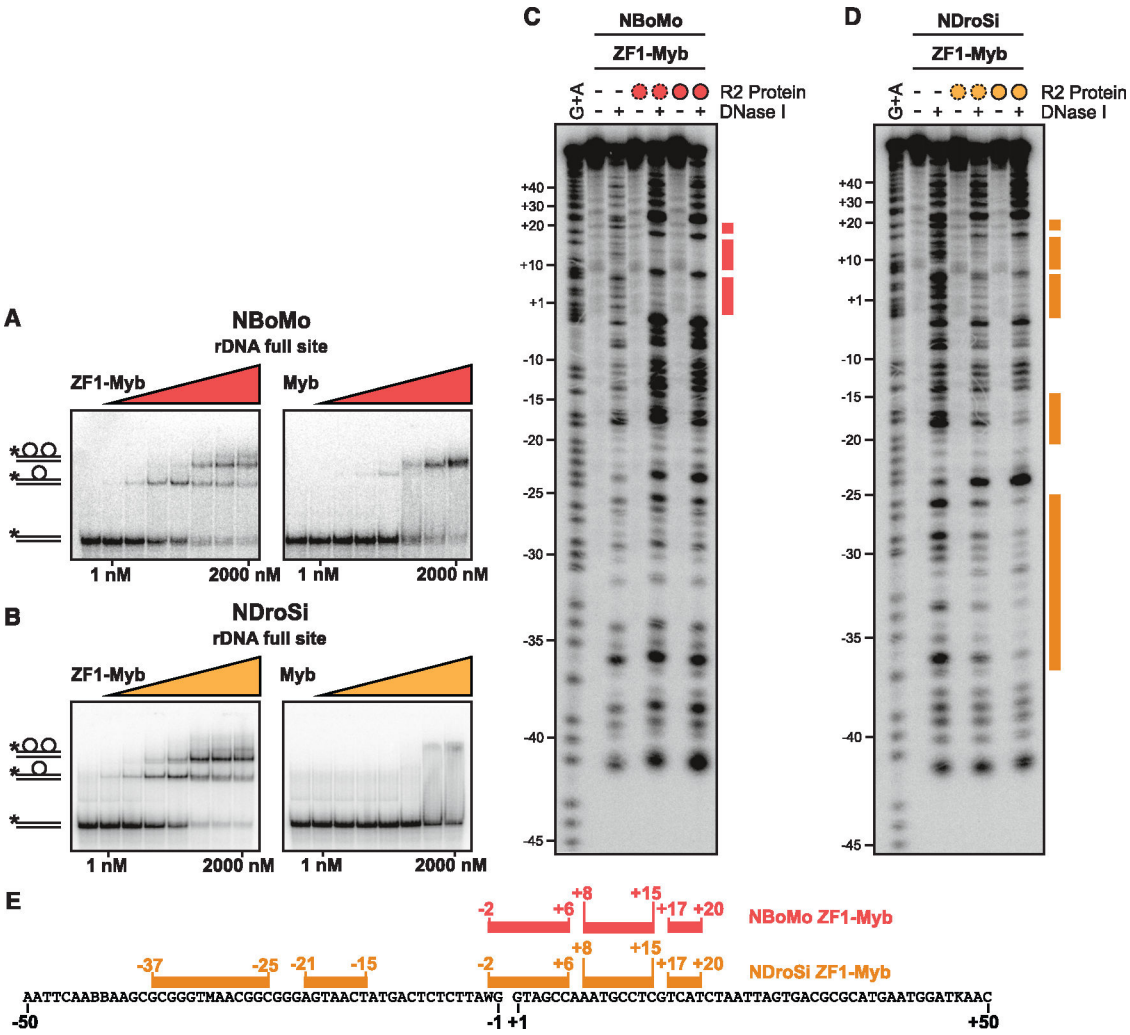


Figure 3. Contribution of D-clade ZF1 and Myb domains to target-site recognition

(A and B) Images of EMSA native PAGE gels. NBoMo proteins (A) and NDroSi proteins (B) were tested with 100-bp duplex target site (−50 to +50).

(C and D) Images of denaturing PAGE gels for DNase I footprinting using 100-bp duplex target site (−50 to +50) with NBoMo ZF1-Myb (C) or NDroSi ZF1-Myb (D). G + A denotes a Maxam-Gilbert sequencing ladder with target-site DNA fragmented at guanosines and adenosines. Numbering on the left indicates target-site DNA position using the numbering scheme of Figure 1C. Circles outlined with dashed or solid lines indicate 125 or 625 nM protein, respectively. Regions of protection are indicated to the right of each gel.

(E) Schematic of DNase I footprints of NBoMo ZF1-Myb and NDroSi ZF1-Myb. The consensus target site for *B. mori* and *D. simulans* is displayed using International Union of Pure and Applied Chemistry (IUPAC) notation.

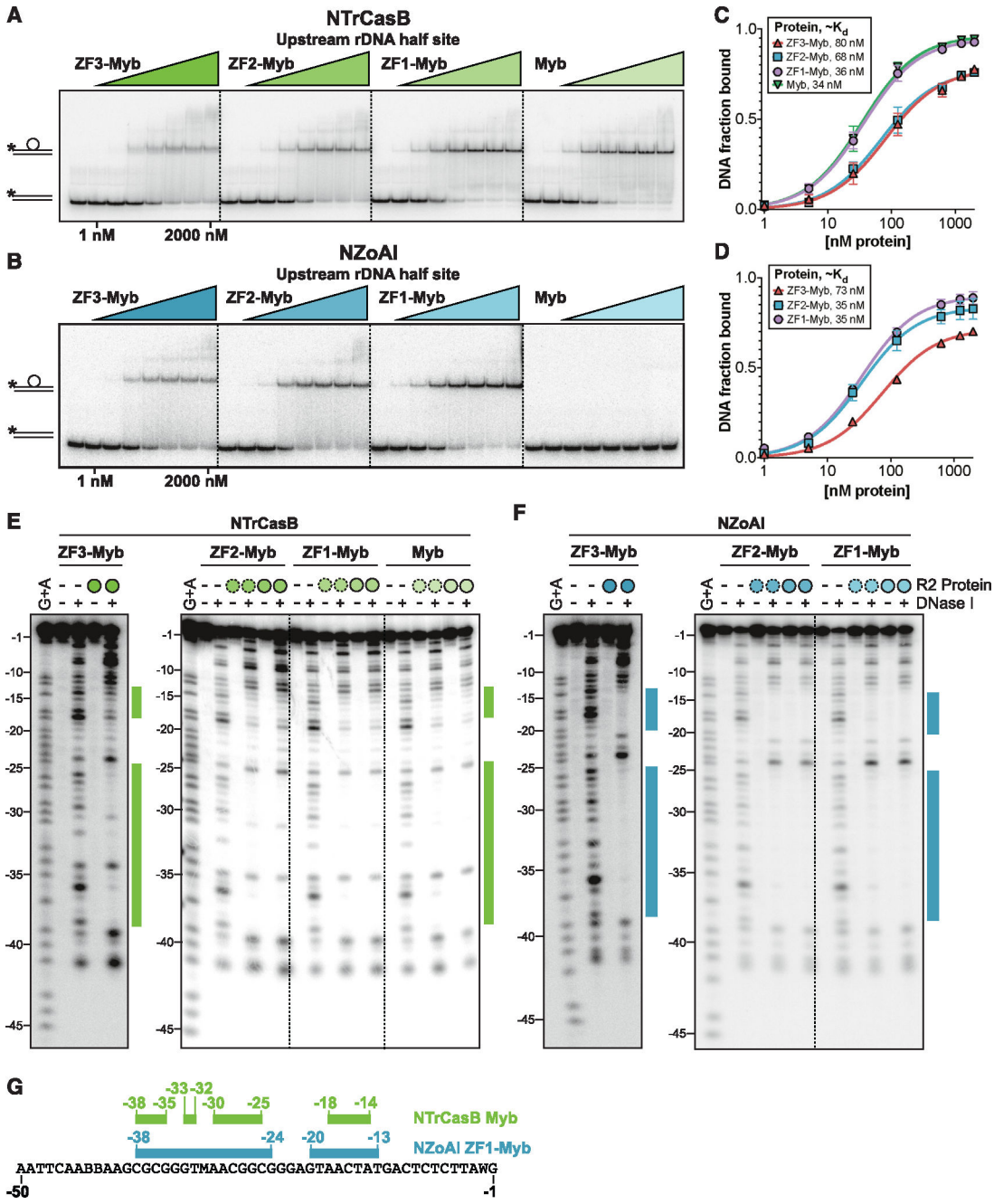


Figure 4. High-affinity target-site binding by A-clade R2 protein ZF1 and Myb domains

(A and B) Images of EMSA native PAGE gels. NTrCasB proteins (A) and NZoAl proteins (B) were tested with upstream target half-site DNA (−50 to −1).

(C and D) Quantifications of (A) and (B) with technical replicates ($n = 3$). The x axis is on a logarithmic scale. Mean ± SEM is plotted.

(E and F) Images of denaturing PAGE gels for DNase I footprinting using upstream target half-site DNA (−50 to −1) with NTrCasB proteins (E) or NZoAl proteins (F). See also Figure S1.

(G) Schematic of DNase I footprints of NTrCasB Myb and NZoAl ZF1-Myb. The consensus target-site sequence for *T. castaneum* and *Z. albicollis* is displayed using IUPAC notation.

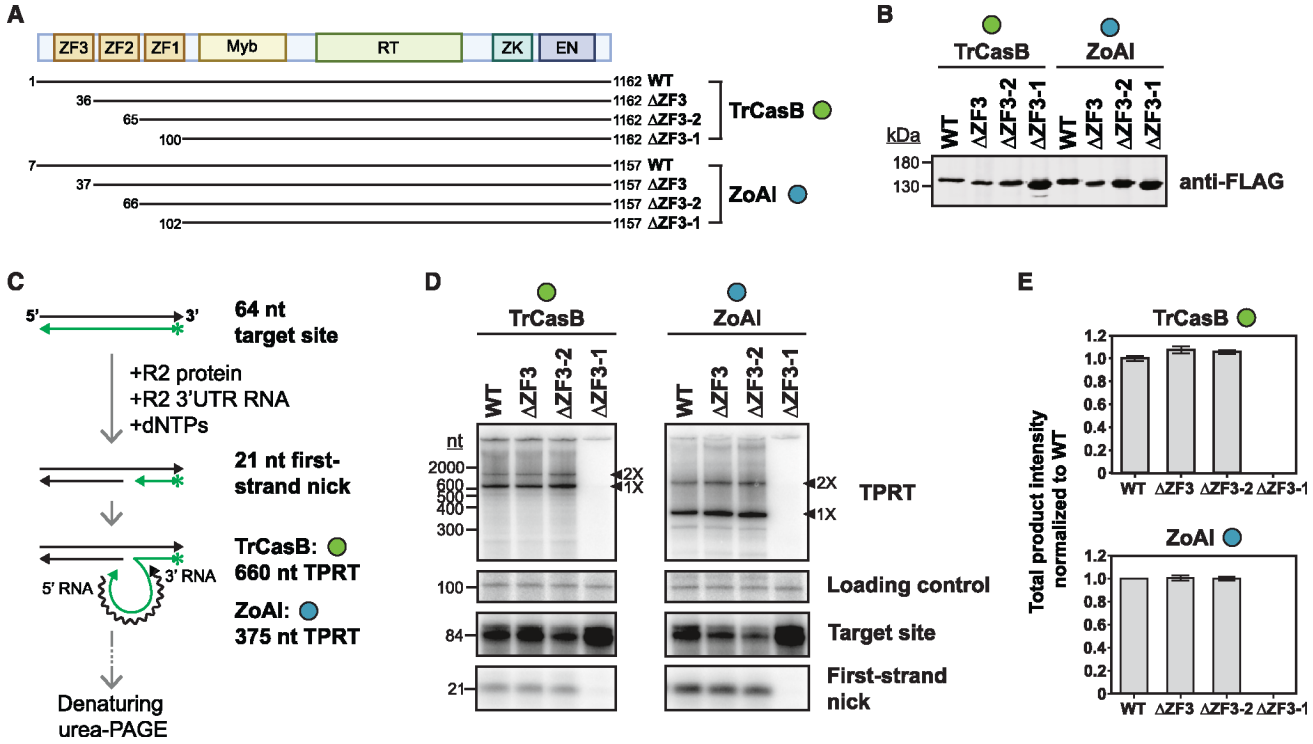


Figure 5. TPRT dependence on N-terminal ZFs *in vitro*

(A) Domain schematic and amino acid numbering for R2 protein versions used in TPRT assays and ZoAI cellular assays.

(B) Purified R2 protein versions used for TPRT assays were resolved by SDS-PAGE and visualized by immunoblot with an anti-FLAG antibody.

(C) Schematic of TPRT assay. Green strands indicate DNA visualizable by 5′ radiolabeling of the antisense strand.

(D) Image from denaturing PAGE of TPRT assay products. TPRT products are indicated with black triangles (1X = cDNA and 2X = cDNA + template jump).

(E) Densitometric quantification of TPRT products (1X and 2X cDNA) and first-strand nick products altogether, from the assays in (D) and technical replicates ($n = 3$). Mean ± SEM is plotted.

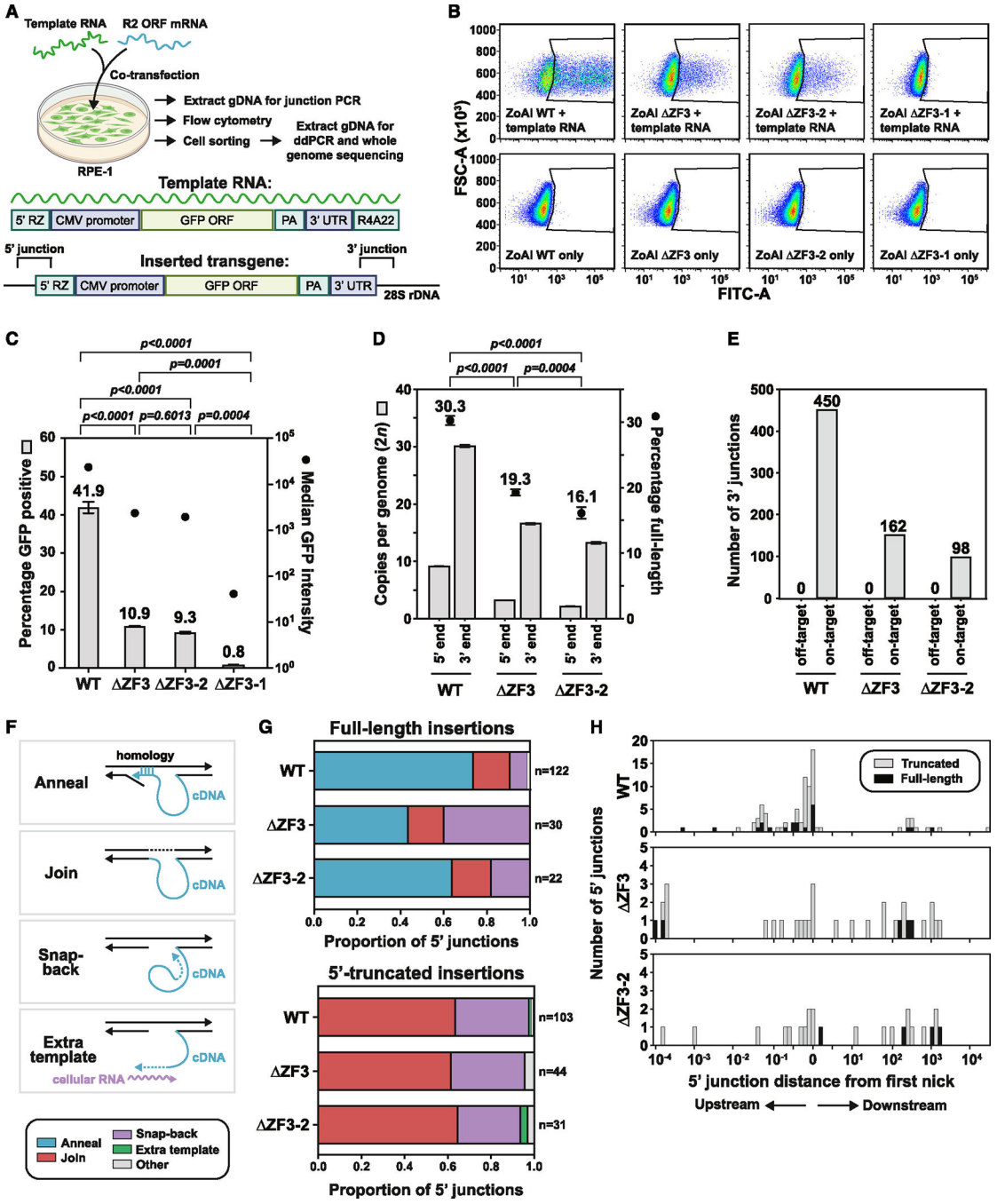


Figure 6. ZF3-2 contributions to new gene insertion in cells

(A) Top, schematic of transgene insertion assay and downstream workflow. RPE-1, retinal pigment epithelium cell line. Middle, schematic of template RNA encoding the GFP expression cassette. Bottom, schematic of transgene inserted into the 28S rDNA target site. 5′ and 3′ junctions are indicated by brackets.

(B) Flow cytometry data for one of three replicates of a parallel set of transgene insertion assays. Cells inside the indicated gating were considered GFP positive. See also Figure S2.

(C) Transgene insertion assays. Flow cytometry results are mean ± SEM of three replicates. Bar plot of percentage GFP-positive cells is on the left y axis. Dot plot of average median GFP intensity is on the right y axis; error bars are not visible because the right y axis is on a logarithmic scale. *p* values for percentage GFP positive comparisons from one-way ANOVA with *post hoc* Tukey's multiple comparisons test are indicated above the plot. See also Figures S2 and S3.

(D) Bar plot of mean ± SEM of insertion copy number from ddPCR is on the left y axis ($n = 4$). Copy number is relative to diploid genome content. Dot plot of average percentage full-length insertions with 95% confidence intervals is on the right y axis ($n = 4$). *p* values for percentage full-length comparisons from one-way ANOVA with *post hoc* Tukey's multiple comparisons test are indicated above the plot. See also Figure S4.

(E) Bar plot of the number of onvs. off-target transgene 3′ junctions.

(F) Schematic of 5′ junction categories. See also Figure S4.

(G) Bar plot of proportion of each type of 5′ junction. Key is to the left under (F). See also Figure S4.

(H) Histogram of rDNA positions of transgene 5′ junctions from the "Join" category. Black and gray bars indicate full-length and truncated transgene insertions, respectively. The x axis is linear between −10 and +10 and otherwise on a logarithmic scale. See also Figure S4.

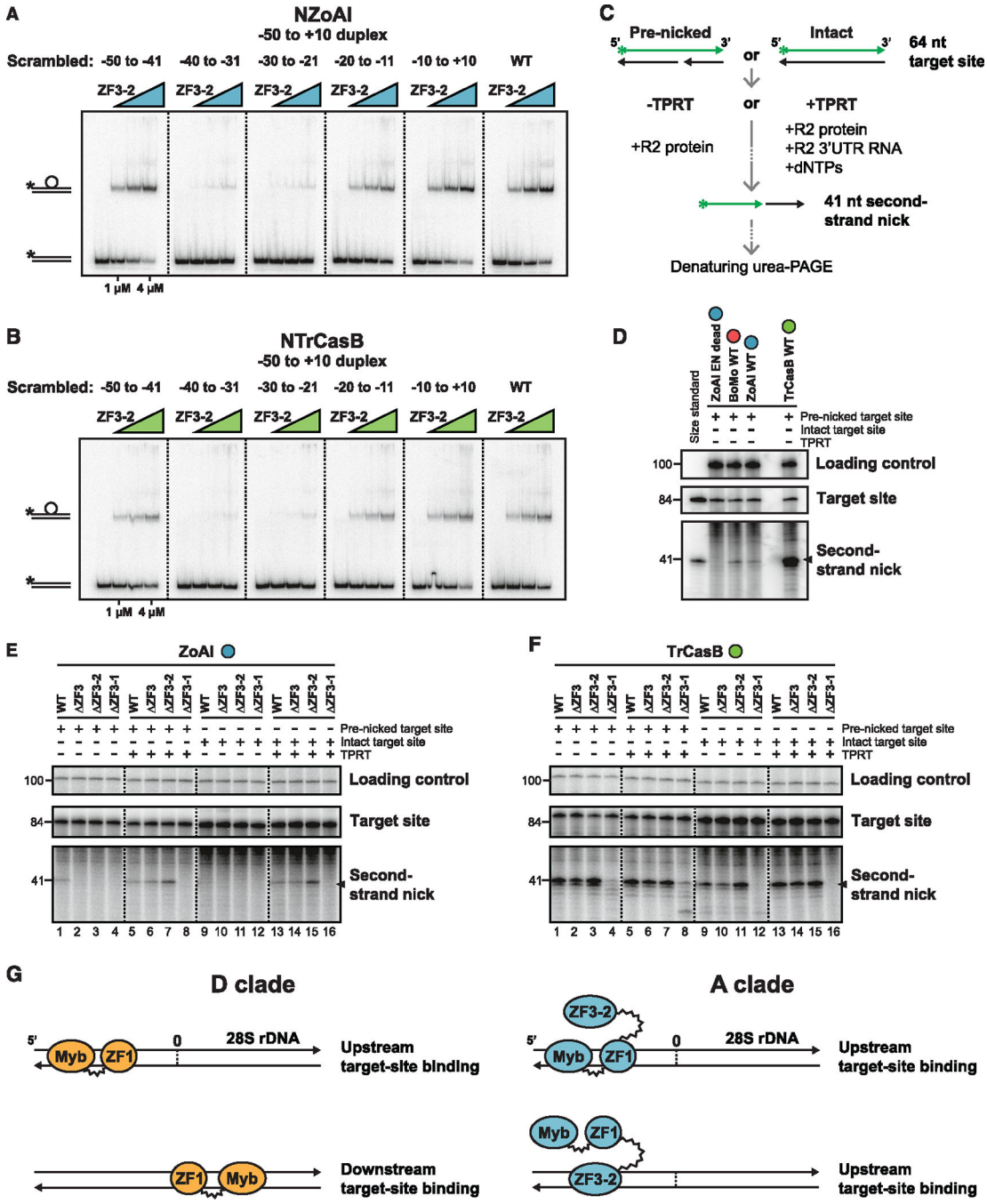


Figure 7. ZF3-2 DNA-binding properties and influence on second-strand nicking

(A and B) Images of EMSA native PAGE gels. NZoAl ZF3-2 polypeptide (A) and NTrCasB ZF3-2 polypeptide (B) were tested with −50 to +10 duplex target-site DNA with a scrambled region indicated above the gel. See also Figure S6.

(C) Schematic of second-strand nicking assay. Green strands indicate DNA visualizable by 5′ radiolabeling.

(D–F) Second-strand nicking position comparison (D) shows similarity across R2 proteins, whereas domain requirements differ for second-strand nicking by ZoAl (E) versus TrCasB (F).

(G) Model for DNA interaction by R2 N-terminal domains of D-clade or A-clade retrotransposons. See also Figure S5.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Monoclonal ANTI-FLAG® M2 antibody produced in mouse	Sigma-Aldrich	Cat#F1804
Goat anti-Mouse IgG (H + L) Cross-Adsorbed Secondary Antibody, Alexa Fluor™ 680	Invitrogen	Cat#A-21057
Bacterial and virus strains		
Rosetta2(DE3)pLysS	Sigma-Aldrich	Cat#71403-M
Chemicals, peptides, and recombinant proteins		
DMEM	Gibco	Cat#10566-016
Seradigm Select Grade USDA Approved Origin Fetal Bovine Serum (FBS)	Avantor	Cat# 89510-186
Penicillin-Streptomycin (10,000 U/mL)	Gibco	Cat#15140122
DMEM/F12	Gibco	Cat#10565-018
Primocin®	Invivogen	Cat#ant-pm-05
Isopropylthio-β-galactoside	Gold Bio	Cat#12481C
Igepal CA-630 ([Octylphenoxy]polyethoxyethanol])	USB Corporation	Cat#19628
Lysozyme from chicken egg white	Sigma-Aldrich	Cat#L6876
Protease Inhibitor Cocktail	Sigma-Aldrich	Cat#P8340
HisPur™ Ni-NTA Resin	Thermo Scientific	Cat#88222
Amylose Resin	NEB	Cat#E8021L
Lipofectamine™ 3000	Invitrogen	Cat#L3000001
DPBS (1X)	Gibco	Cat#14190-144
0.05% Trypsin-EDTA (1X)	Gibco	Cat#25300-054
ANTI-FLAG® M2 Affinity Gel	Sigma-Aldrich	Cat#A2220
3X FLAG® Peptide	Sigma-Aldrich	Cat#F4799
Bovine Serum Albumin	Sigma-Aldrich	Cat#A1470
T4 Polynucleotide Kinase	NEB	Cat#M0201L
ATP, [γ-32P]- 3000 Ci/mmol 10 mCi/ml EasyTide, 250 μCi	Perkin Elmer	Cat#BLU502A250UC
ProbeQuant™ G-50 Micro Columns	Cytiva	Cat# GE28-9034-08
Poly[d(I-C)]	Roche	Cat#10108812001
CHAPS hydrate	Sigma-Aldrich	Cat#C3023
DNase I (RNase-free)	NEB	Cat#M0303L
Proteinase K, Molecular Biology Grade	NEB	Cat#P8107S
UltraPure™ Phenol:Chloroform:Isoamyl Alcohol (25:24:1, v/v)	Invitrogen	Cat#15593031
Glycogen from mussel, Mytilus genus	Sigma-Aldrich	Cat#G1508
BbsI	NEB	Cat#R0539L
CleanCap® Reagent AG	TriLink	Cat#W-7113
N1-Methylpseudouridine-5′-Triphosphate	TriLink	Cat#W-1081

REAGENT or RESOURCE	SOURCE	IDENTIFIER
SYBR™ Gold Nucleic Acid Gel Stain (10,000X Concentrate in DMSO)	Invitrogen	Cat#S11494
RNase A, DNase and protease-free (10 mg/mL)	Thermo Scientific	Cat#EN0531
Lipofectamine™ MessengerMAX™ Transfection Reagent	Invitrogen	Cat#LMRNA001
Q5® High-Fidelity DNA Polymerase	NEB	Cat#M0491L
BamHI	NEB	Cat#R0136
XmnI	NEB	Cat#R0194
ddPCR supermix for Probes (no dUTP)	Bio-Rad	Cat# 1863024
Droplet Generation Oil for Probes	Bio-Rad	Cat #1863005
Critical commercial assays		
Pierce™ BCA Protein Assay Kit	Thermo Scientific	Cat#23227
QIAquick PCR purification kit	Qiagen	Cat#28104
HiScribe® T7 High Yield RNA Synthesis Kit	NEB	Cat#E2040S
Deposited data		
Illumina whole genome shotgun sequencing data	This manuscript	NCBI Sequence Read Archive (SRA): SRR24873001, SRR24873002, SRR24873003
Experimental models: Cell lines		
HEK 293T	UC Berkeley Cell Culture Facility	RRID: SCR_017924
RPE-1 hTERT	UC Berkeley Cell Culture Facility	RRID: SCR_017924
Oligonucleotides		
Target site oligonucleotides, see Table S2	IDT	N/A
Primers for junction PCR, see Table S2	IDT	N/A
Primers for ddPCR, see Table S2	IDT	N/A
Recombinant DNA		
Plasmid: 2bct	UC Berkeley QB3 MacroLab	N/A
Plasmid: 2bct-MBP_NBoMo_ZF1-Myb_6xH	This manuscript	N/A
Plasmid: 2bct-MBP_NBoMo_Myb_6xH	This manuscript	N/A
Plasmid: 2bct-MBP_NDroSi_ZF1-Myb_6xH	This manuscript	N/A
Plasmid: 2bct-MBP_NDroSi_Myb_6xH	This manuscript	N/A
Plasmid: 2bct-MBP_NTrCasB_ZF3-Myb_6xH	This manuscript	N/A
Plasmid: 2bct-MBP_NTrCasB_ZF2-Myb_6xH	This manuscript	N/A
Plasmid: 2bct-MBP_NTrCasB_ZF1-Myb_6xH	This manuscript	N/A
Plasmid: 2bct-MBP_NTrCasB_Myb_6xH	This manuscript	N/A
Plasmid: 2bct-MBP_NTrCasB_ZF3-2_6xH	This manuscript	N/A
Plasmid: 2bct-MBP_NZoAl_ZF3-Myb_6xH	This manuscript	N/A

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Plasmid: 2bct-MBP_NZoAl_ZF2-Myb_6xH	This manuscript	N/A
Plasmid: 2bct-MBP_NZoAl_ZF1-Myb_6xH	This manuscript	N/A
Plasmid: 2bct-MBP_NZoAl_Myb_6xH	This manuscript	N/A
Plasmid: 2bct-MBP_NZoAl_ZF3-2_6xH	This manuscript	N/A
Plasmid: pcDNA3.1(+)_N-FLAG_WT_TrCasB	GenScript	N/A
Plasmid: pcDNA3.1(+)_N-FLAG_ΔZF3_TrCasB	This manuscript	N/A
Plasmid: pcDNA3.1(+)_N-FLAG_ΔZF3-2_TrCasB	This manuscript	N/A
Plasmid: pcDNA3.1(+)_N-FLAG_ΔZF3-1_TrCasB	This manuscript	N/A
Plasmid: pcDNA3.1(+)_N-FLAG_WT_ZoAl	Zhang et al.[21]	N/A
Plasmid: pcDNA3.1(+)_N-FLAG_ΔZF3_ZoAl	This manuscript	N/A
Plasmid: pcDNA3.1(+)_N-FLAG_ΔZF3-2_ZoAl	This manuscript	N/A
Plasmid: pcDNA3.1(+)_N-FLAG_ΔZF3-1_ZoAl	This manuscript	N/A
Plasmid: pcDNA3.1(+)_N-FLAG_ZoAl_DD1041/1054AA	Zhang et al.[21]	N/A
Plasmid: pcDNA3.1(+)_N-FLAG_BoMo	Zhang et al.[21]	N/A
Plasmid: pT7mmRNAF_ZoAl	Zhang et al.[21]	N/A
Plasmid: pT7mmRNAF_ΔZF3_ZoAl	This manuscript	N/A
Plasmid: pT7mmRNAF_ΔZF3-2_ZoAl	This manuscript	N/A
Plasmid: pT7mmRNAF_ΔZF3-1_ZoAl	This manuscript	N/A
Plasmid: L8GGTrCa5RZ_CMVGFP_PAmin_GeFo3	Zhang et al.[21]	N/A
Software and algorithms		
GraphPad Prism 9 for macOS version 9.4.1	GraphPad Software	graphpad.com
Clustal Omega	EMBL-EBI	ebi.ac.uk/Tools/msa/clustalo
Jalview version 2.11.2.6	Waterhouse et al.[34]	jalview.org
UCSF ChimeraX version 1.6rc202304072249	Pettersen et al.[35]	rbvi.ucsf.edu/chimerax
ColabFold version 1.5.2	Mirdita et al.[36]	colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb
ImageJ	Schneider et al.[37]	ImageJ.net/software/ImageJ
Python version 3.6	Python Software Foundation	python.org
Burrows-Wheeler Aligner version 0.7.17	Li and Durbin[38]	bio-bwa.sourceforge.net
BBMap version 38.97	Bushnell[39]	sourceforge.net/projects/bbmap/
Trimmomatic version 0.39	Bolger et al.[40]	usadellab.org/cms/index.php?page=trimmomatic
JDK version 17.0.2	Oracle	oracle.com/java/technologies/downloads
Samtools version 1.8	Danecek et al.[41]	htslib.org
Numpy	Harris et al.[42]	numpy.org
Pandas	McKinney[43]	pandas.pydata.org
SciPy	Virtanen et al.[44]	scipy.org
Matplotlib	Hunter[45]	matplotlib.org
Biopython	Cock et al.[46]	biopython.org

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Pysam	N/A	github.com/pysam-developers/pysam
Seaborn	Waskom[47]	seaborn.pydata.org
Fuzzysearch	N/A	github.com/taleinat/fuzzysearch
Custom Python scripts	This manuscript	zenodo.org/doi/10.5281/zenodo.10439695
Other		
Nitrocellulose Membrane, 0.45 μM	Bio-Rad	Cat#1620115
GE Storage Phosphor Screens	Cytiva	Cat#GE28-9564-74
Sony Sorting Chip-130μm for SH800 and MA900	Sony Biotechnology	Cat#LE-C3213
DG8™ Cartridges and Gaskets	Bio-Rad	Cat #1864007
QX200™ Droplet Generator	Bio-Rad	Cat#1864002
LI-COR Odyssey CLx	LI-COR	Model 9140
Amersham Typhoon Biomolecular Imager	Cytiva	Model 5
Sony Cell Sorter	Sony Biotechnology	Model LE-SH800