

Supplementary Note 1 | Template resetting mechanism of DRT10.

Our experiments substituting each position within the DRT10 ncRNA template loop with its complementary nucleotide support the template resetting mechanism diagrammed in **Fig. 2b**, in which dissociation of the nascent cDNA from A and reannealing to A' provides a DNA primer for a subsequent round of reverse transcription. Notably, mutating the first two bases of A or A' abolished repeat addition, whereas mutating the fourth base was tolerated (**Fig. 2a, Extended Data Fig. 4a**), reflecting a more stringent base-pairing requirement at the 3' end of the DNA primer. Mutation of the third base produced asymmetric effects: the A' mutation (which would cause a T:U mismatch) was tolerated, whereas the A mutation (which would cause an A:A mismatch) was not (**Fig. 2a, Extended Data Fig. 4a**), potentially due to greater steric clash from a purine–purine mismatch. Closer inspection of the sequencing results from mutations to the fourth base of A and A' revealed another intriguing asymmetry: the mutation in A (U110A) yielded a cDNA with a mutated repeat sequence, while the mutation in A' (U119A) preserved the WT repeat sequence (**Extended Data Fig. 4a**). This result demonstrates that templating of cDNA synthesis is dictated by the A–B segment, whereas realignment of the cDNA during repeat addition is dictated by the A' segment.

We further probed the template resetting mechanism by testing additional perturbations to the A–B–A' sequence and larger structural elements of the ncRNA. First, we assessed the contributions of the 3' ends of the A and A' segments to template resetting by mutating the last two bases of A or A', thus leaving only 2-nt of homology (**Supplementary Fig. 2a**). This led to a loss of tandem-repeat cDNA production that could be partially rescued by simultaneous mutation (**Supplementary Fig. 2a**), demonstrating the contributions of these positions to template resetting. We also modulated the GC content of A and A' and found that variations in GC content were generally tolerated, but the highest activity was observed with the WT sequence (**Supplementary Fig. 2b**). Variations in the GC content and length of the B segment were also readily accommodated for tandem-repeat synthesis (**Supplementary Fig. 2c,d**). Finally, we assessed the broader contributions of the ncRNA secondary structure by individually scrambling each stem-loop (SL). All SL perturbations decreased repeat addition, but only mutation of SL5 completely eliminated repeat products (**Supplementary Fig. 2e**), supporting the role of SL5 in setting the 5' template boundary.

We were also interested to better understand variations in the precision of tandem-repeat synthesis — that is, the extent to which the cDNA products contain contiguous repeats of full-length cDNA motifs. In particular, we noticed that *Eco*₁DRT10 exhibits less precise repeat addition than the other homologs (**Extended Data Fig. 2f,h, Supplementary Fig. 2f,g**). Closer inspection of the A–B–A' sequence for *Eco*₁DRT10 revealed a cryptic re-annealing site within the B segment that could potentially interfere with accurate template resetting (**Supplementary Fig. 2h**). In support of this, deletion of the cryptic re-annealing site dramatically improved repeat periodicity (**Supplementary Fig. 2i**).

Collectively, these results underscore the role of A–A' homology in guiding tandem-repeat synthesis and provide insight into natural variations in repeat precision across DRT10 systems.

Supplementary Note 2 | Structural constraints on protein-primed reverse transcription

To assess the importance of the spatial positioning of the DRT10 RT C-terminus for protein-primed reverse transcription, we generated truncated variants of the *Eco*₃ RT C-terminal helix while preserving the terminal serine and tested the variants for phage defense activity. We found that deletion of as few as five residues completely abolished activity (**Supplementary Fig. 5a**), indicating that the precise placement of the priming residue is essential. We next tested whether its chemical identity is similarly constrained, given that distinct DRT10 homologs use either serine or tyrosine. Interestingly, substituting the *Eco*₃ priming serine with tyrosine (S456Y) resulted in loss of phage defense, whereas the reciprocal mutation in *Eco*₁ (Y444S) largely retained defense activity (**Extended Data Fig. 11b**). These results indicate that the RT active site can accommodate different hydroxyl-containing residues for protein priming, but only within compatible structural contexts.

Supplementary Note 3 | DRT10 ncRNA template boundary definition

Our structural and biochemical data indicate that RT interactions with the ncRNA at SL5 and SL1 function to define the 5' and 3' template boundaries, respectively. However, the presence of additional “linker” nucleotides separating the template from each SL led us to investigate the contributions of these linker regions to boundary definition. While a single 5' deletion ($\Delta 105$) diminished repeat synthesis *in vitro* (**Extended Data Fig. 9h**), a 2-nt deletion eliminated detectable WT repeat products but nevertheless appeared to retain high synthetic yields of kilobase-length DNA (**Extended Data Fig. 9h, Supplementary Fig. 6a**). Closer inspection of sequencing data for this mutant revealed the presence of a new repeat motif, CAT, corresponding to an alternative A–B–A' pattern in the template (**Supplementary Fig. 6b**). With this alternative template, synthesis terminates at the same 2-nt distance from SL5 as in the WT ncRNA, supporting the role of SL5 in defining a geometrically constrained window for the 5' template boundary. In agreement with this model, lengthening the 5' linker led to disruption of WT repeat synthesis due to frequent addition of adenosine residues templated by the inserted uridines, indicating read-through beyond the WT boundary (**Extended Data Fig. 9h, Supplementary Fig. 6c**). Furthermore, by shifting the template boundary, these insertions effectively lengthened A without a corresponding modification in A', thus disrupting reannealing of the template and, in turn, processive repeat synthesis.

We also assessed similar perturbations to the 3' linker. As with the 5' linker, truncation of the 3' linker diminished tandem-repeat synthesis (**Extended Data Fig. 9i**), indicating steric constraints on the distance between SL1 and the template. However, insertion mutations were comparatively well tolerated (**Extended Data Fig. 9i**). We attribute the additional flexibility on the 3' side to the role of microhomology between the template and nascent cDNA, which functions to register the position from which each new repeat initiates and thereby at least partially decouples repeat synthesis from strict 3' linker length.

Supplementary Note 4 | Methodological considerations for phylogenetic reconstruction

Reverse transcriptases (RTs) exhibit remarkable sequence diversity. This is unsurprising, given that their frequent mobilization, their presence across all domains of life, and their likely ancient origin^{104,105}.

As most methods for phylogenetic reconstruction are based on multiple sequence alignments (MSAs), such sequence divergence makes accurate phylogenetic inference particularly challenging. Firstly, substitutional saturation can obscure or underestimate true divergence, compressing long branches and promoting a long-branch attraction artifact. Secondly, such divergence also leaves few confidently homologous positions, so stochastic error and alignment uncertainty can introduce spurious synapomorphies that bias branch lengths and topology.

To improve the quality of the MSA underlying our RT phylogeny, we developed a methodology to minimize errors derived from aligning these highly divergent sequences. First, we combined a comprehensive RT domain search with stringent filtering criteria to ensure domain completeness. Second, we leveraged structure-based MSAs that have previously been used to reconstruct the evolutionary history of other ancient enzymes with high sequence divergence, such as B-family DNA polymerases¹⁰⁶ and RNA-dependent RNA polymerases¹⁰⁷. Third, we conservatively trimmed highly gapped positions and removed spurious sequences to maximize reliable homology while retaining informative variability. Below, we provide a detailed explanation of these decisions.

Detailed methodology for RT phylogenetic analysis

Comprehensive RT domain search that avoided truncated domains

To assemble a comprehensive dataset of RTs across all domains of life, we used *hmmsearch* from the HMMER suite to screen a FASTA-formatted version of the AlphaFold Database¹⁰⁸ with Hidden Markov Models (HMMs) of RT domains from the public PFAM⁸⁶, NCBI CDD⁸⁷, and TIGRFAM⁸⁸ databases, as well as custom HMM profiles for specific prokaryotic RT subfamilies reported in recent phylogenies^{2,89,90}.

To maintain a high level of confidence in the retrieved *hmmsearch* hits, we applied several stringent filtering steps. First, we ensured high-quality hits by removing any with a domain length below 100 amino acids, E-value above 1e-10, or HMM coverage below 70%. Second, we conservatively removed proteins for which the best hit was either an RNA-dependent RNA

polymerase (HMM profile PF00978 or any of the NeoRdRP 2.1 HMM profiles⁹³) or a maturase K protein (HMM profile CHL00002), because these enzymes may yield significant hits in RT HMM profiles due to their evolutionary relatedness^{91,92}.

To compile a dataset of RT domains that were as complete as possible, we merged overlapping or adjacent RT hits originating from the same HMM profile and separated by fewer than 30 amino acids for each representative. Such small gaps typically reflect alignment discontinuities due to insertions rather than distinct domains. We then extracted the sequences delineated by the hmmsearch hit coordinates while adding an upstream and downstream buffer of 100 amino acids to ensure complete domain capture, since hmmsearch hit sequences can truncate prematurely. This yielded a large collection of putative complete RT domains, which were clustered permissively at a 40% sequence identity threshold using CD-HIT to create a non-redundant set spanning the diversity of RTs. The resulting 8,533 clusters formed the basis of our unique RT domain dataset.

Structure-based MSA of the RT domain

To align these RT domains and build a high-quality MSA for our phylogenetic analysis, we circumvented the high sequence divergence of RTs by avoiding sequence-only based alignment methods and opted for a structure-based MSA instead, which more accurately aligns proteins with low sequence identity and remote homology^{92,106,109}. We first retrieved the predicted structures of full-length RT sequences corresponding to each RT domain in our dataset from the AlphaFold Database. We then trimmed the coordinates to the same domain boundaries used in the primary sequence-level extraction. This provided a one-to-one mapping between each RT domain sequence and its predicted structure. Finally, we generated structure-based multiple sequence alignments (MSAs) using Reseck⁹⁴ and Muscle-3D with the -super7 algorithm⁹⁵. We used Muscle-3D because it scales well to thousands of proteins, compared to other recently published structure alignment methods.

Processing the MSA to ensure high-quality alignment of complete RT domains

To ensure that the resulting MSA retained only the most reliably aligned positions, it was first trimmed to include only the canonical RT motifs 0–7. Any column with more than 50% gaps was then pruned using trimAl⁹⁶. To further eliminate spurious or incomplete entries, we also

stringently removed any MSA members with gaps at any of the four conserved catalytic residues in RT motif 5, an alignment length shorter than 100 amino acids, or a partial core RT domain that begins after RT motif 1 or ends before RT motif 7.

Altogether, these criteria ensured that only high-quality, full-length RT domains with catalytic motifs were included in the final MSA of 6,734 members, allowing us to reduce artifacts arising from alignment errors and increase the robustness of the downstream maximum-likelihood phylogenetic tree, which we generated using VeryFastTree⁹⁷.

Choice of tree-building algorithm

Pythia¹¹⁰, a tool that quantifies the difficulty of maximum-likelihood-based tree inferences for a given MSA, classified our alignment as “difficult” (difficulty = ~0.7). This scoring predicts that starting trees (or seeds) can produce multiple near-equal-likelihood final topologies, and that convergence can be slow and ambiguous. At this level of difficulty, different maximum-likelihood phylogenetic tools such as FastTree, IQ-TREE, and RAxML may yield differing topologies at similar likelihoods, though IQ-TREE and RAxML are often regarded as more accurate at lower dataset difficulties¹¹¹.

Given this parity in tree-building accuracy, we chose to employ VeryFastTree, a highly optimized implementation of FastTree specifically designed to align massive numbers of sequences⁹⁷, as the large size of our dataset was causing slow convergence of IQ-TREE and RAxML trees. Additionally, the choice of a FastTree-based method (with WAG+ Γ and pseudocounts) was well supported by literature precedent from previous large-scale analyses of prokaryotic RTs^{2,89,90,112–114}.

To assess the robustness of our phylogenetic reconstruction and account for potential uncertainty in the underlying alignment, we calculated bootstrap support using an ensemble approach. Whereas traditional bootstrap methods only evaluate the reconstruction algorithm's variability, the primary source of error in highly divergent protein families, such as RTs, often lies in the alignment process itself¹¹⁵. As such, in employing an ensemble approach, we re-aligned the RT domain sequences 1,000 times while introducing random noise (using the -perturb option in Muscle-3D) and then built phylogenetic trees for each of these perturbed alignments (using the same refinement and tree building process as used for the original alignment). This strategy enabled us to assess the stability of the phylogenetic groupings in the tree and their robustness to

local variations in the alignment algorithm. Finally, to quantify bootstrap support across this ensemble of 1,000 trees, we used the Transfer Bootstrap Expectation (TBE) metric calculated with RAxML-NG. In comparison to traditional bootstrap metrics like Felsenstein's bootstrap proportions, TBE metrics are better suited for large-scale phylogenies with high sequence divergence^{98,116}.

Orthogonal support

To complement our phylogenetic analysis, we further employed a parallel structure-based computational analysis. We built a structure-based tree by running all-vs-all Foldseek comparisons on the full-length AlphaFold models, converting normalized bit scores into distances, and reconstructing the tree using DecentTree (NJ-R). Because full-length comparisons incorporate accessory domains (e.g., RNase H, fingers/thumb extensions, lineage-specific inserts, etc.) in addition to the core RT domain, this analysis assesses overall architectural similarity rather than the core palm domain alone. Despite potential noise from flexible/disordered regions, the structure-based tree recovered broad partitions precisely where the sequence signal is weakest (i.e., resolved canonical group II intron RTs, retron-like RTs, DGR-associated RTs, and LTR/non-LTR lineages). This structure-distance clustering analysis provides an orthogonal line of evidence to our phylogenetic analysis for the broad assignment of RT groupings. Indeed, agreement between the maximum-likelihood phylogenetic tree and this structure-distance clustering analysis at coarse scales increases confidence that proposed groupings are not solely due to alignment artifacts or model misspecification.

Supplementary Figure 1 | Uncropped images of electrophoretic separation assays.

Fig. 1k

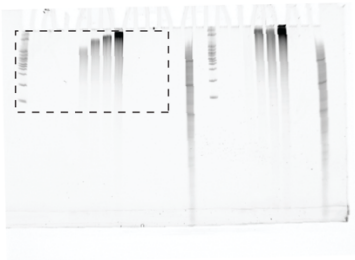


Fig. 1m

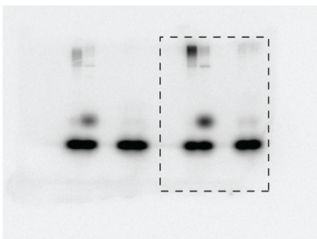


Fig. 3i

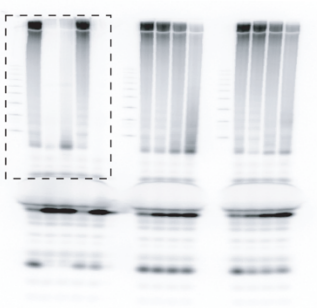


Fig. 4c

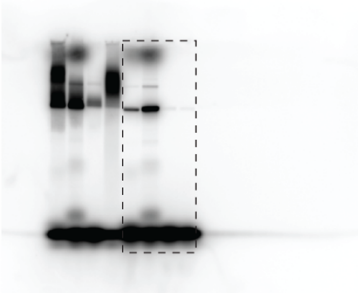


Fig. 4g

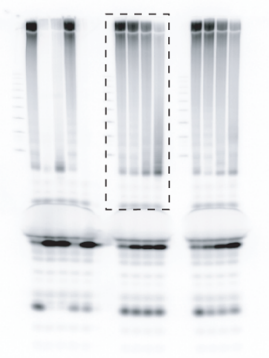
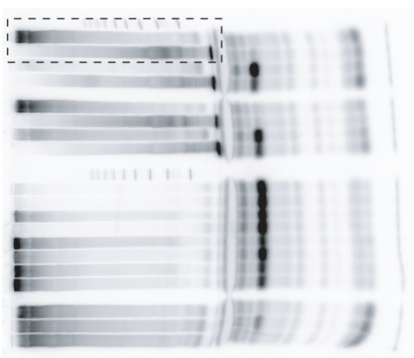
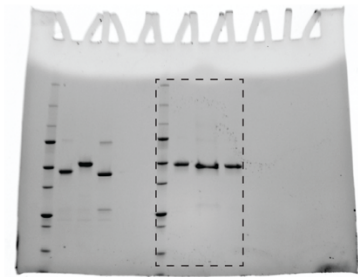


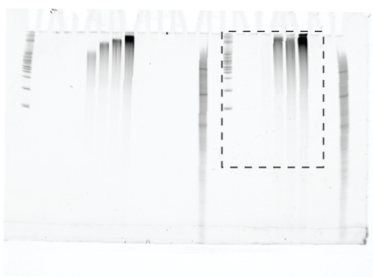
Fig. 4k



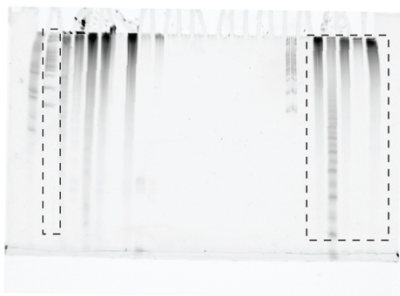
Extended Data Fig. 3a



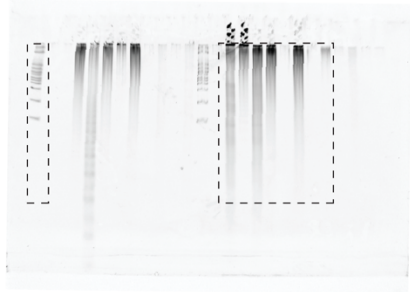
Extended Data Fig. 3b



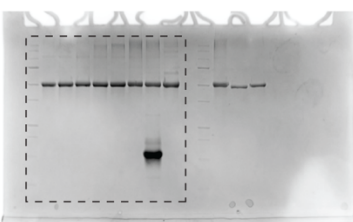
Extended Data Fig. 3c



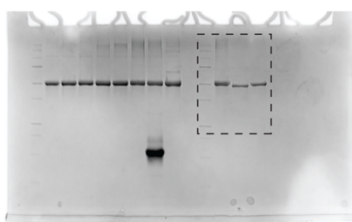
Extended Data Fig. 3d



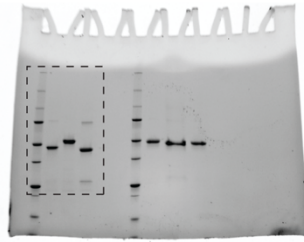
Extended Data Fig. 3e



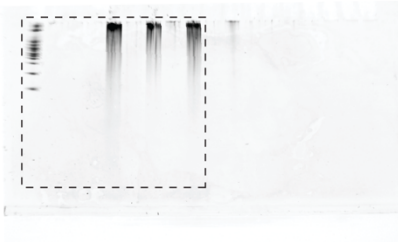
Extended Data Fig. 3f



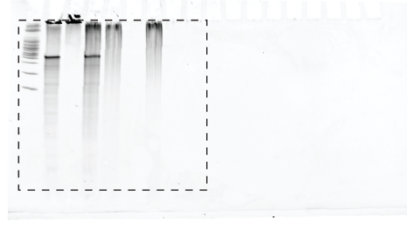
Extended Data Fig. 3i



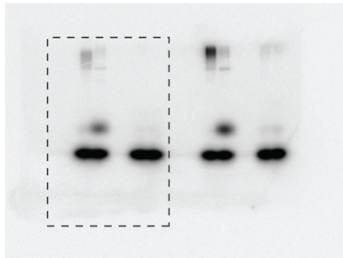
Extended Data Fig. 3j



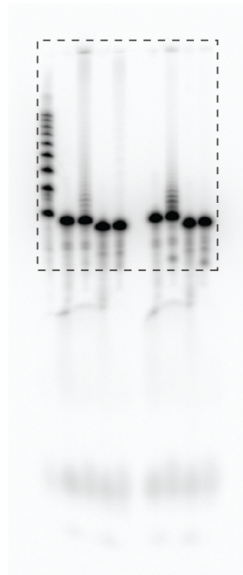
Extended Data Fig. 3k



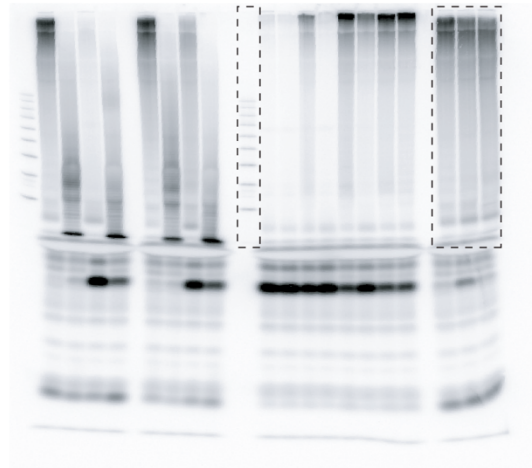
Extended Data Fig. 3l



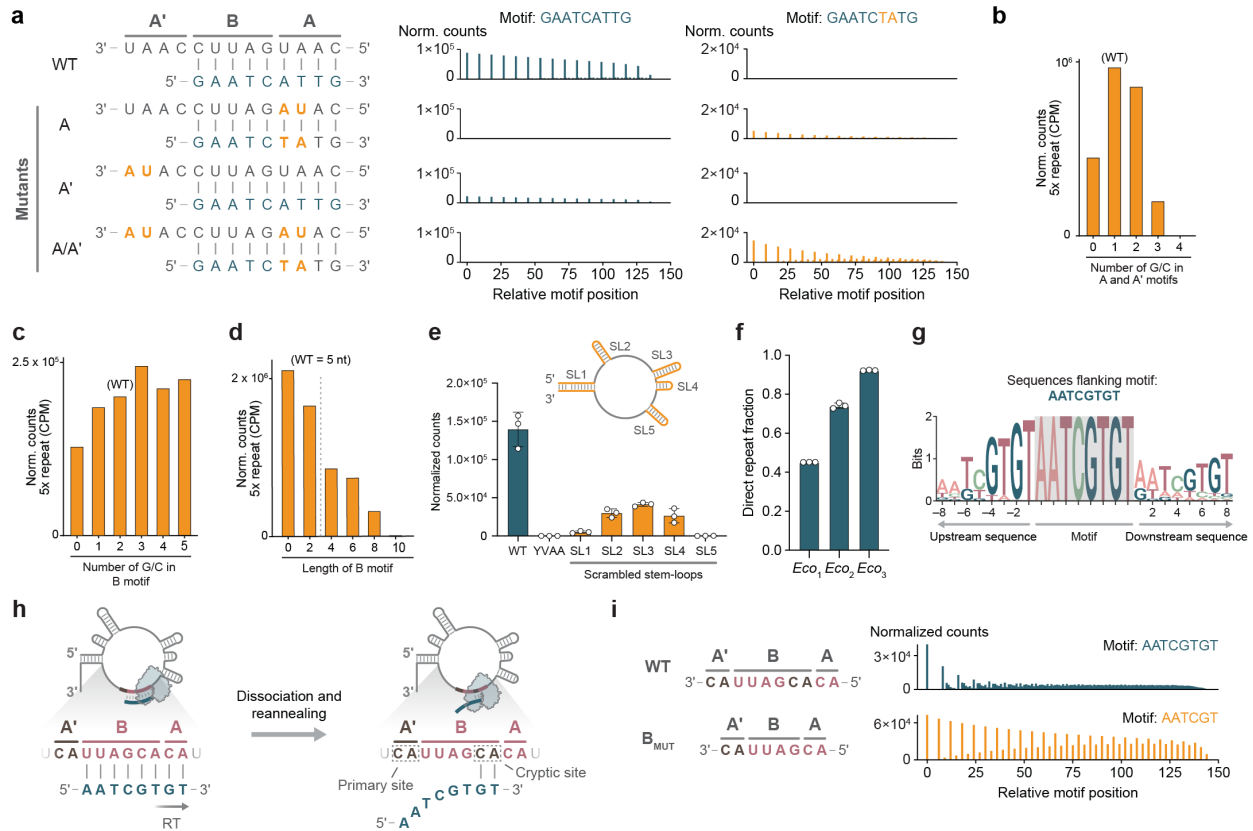
Extended Data Fig. 11e



Supplementary Information Fig. 6a



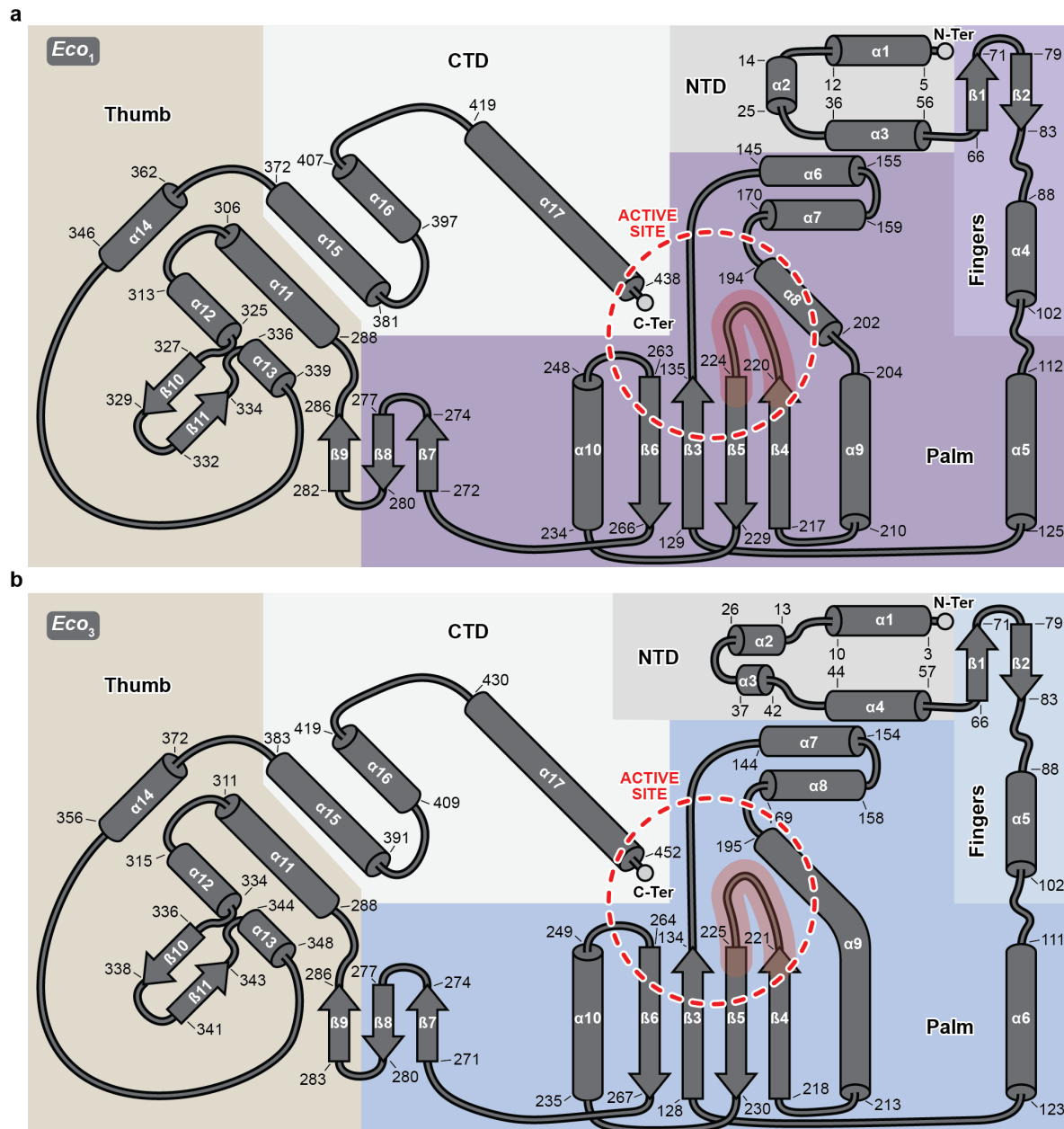
Dashed boxes outline the cropped area used in the indicated figures.

Supplementary Figure 2 | RNA determinants of DRT10 tandem-repeat cDNA synthesis.

a, *Left*, ncRNA template mutants that disrupt and restore homology between the 3'-half of A and A'. *Right*, Miniprep-seq motif position graphs quantifying cDNA repeat motifs templated by WT or mutant (right) ncRNAs. Data are normalized as counts per million plasmid-mapping reads for each sample. **b,c**, Sequencing-based quantification of cDNA products from *in vitro* reactions with ncRNA variants of altered GC content in A/A' (**b**) or B (**c**), with the unaltered region retained as WT. Reads containing ≥ 5 consecutive motif repeats were normalized for sequencing depth as counts per million reads (CPM). **d**, cDNA synthesis by ncRNA variants with altered B-motif length, quantified as in **b,c**. **e**, *Top*, *Eco*₃DRT10 ncRNA secondary structure highlighting stem-loops (SLs) disrupted by mutations. *Bottom*, Bar graph quantifying miniprep-seq reads containing ≥ 3 consecutive cDNA motif repeats for these mutants alongside WT and catalytically inactive (YVAA) controls. Data are normalized as in **a** and represent the mean $\pm$ s.d. for $n = 3$ independent biological replicates. **f**, Bar graph quantifying the “direct repeat fraction,” or the fraction of motif hits in DRT10 miniprep-seq reads precisely flanked by matching motif repeats. *Eco*₁DRT10 has the lowest direct repeat fraction of the tested homologs, indicative of inaccurate repeat addition. Data are mean $\pm$ s.d. for $n = 3$ independent biological replicates. **g**, Sequence logo of motif hits and flanking sequences from *Eco*₁DRT10 miniprep-seq datasets, showing the degeneracy of motif-flanking sequences. **h**, Model of incomplete repeat addition by *Eco*₁DRT10 due to the presence of a cryptic re-annealing site within B, leading to addition of a 2-nt 5'-GT-3' sequence rather than the full 8-nt motif in subsequent extension cycles. **i**, Miniprep-seq motif position graphs for *Eco*₁DRT10 with the template sequences shown at left. Removal of the cryptic re-annealing site

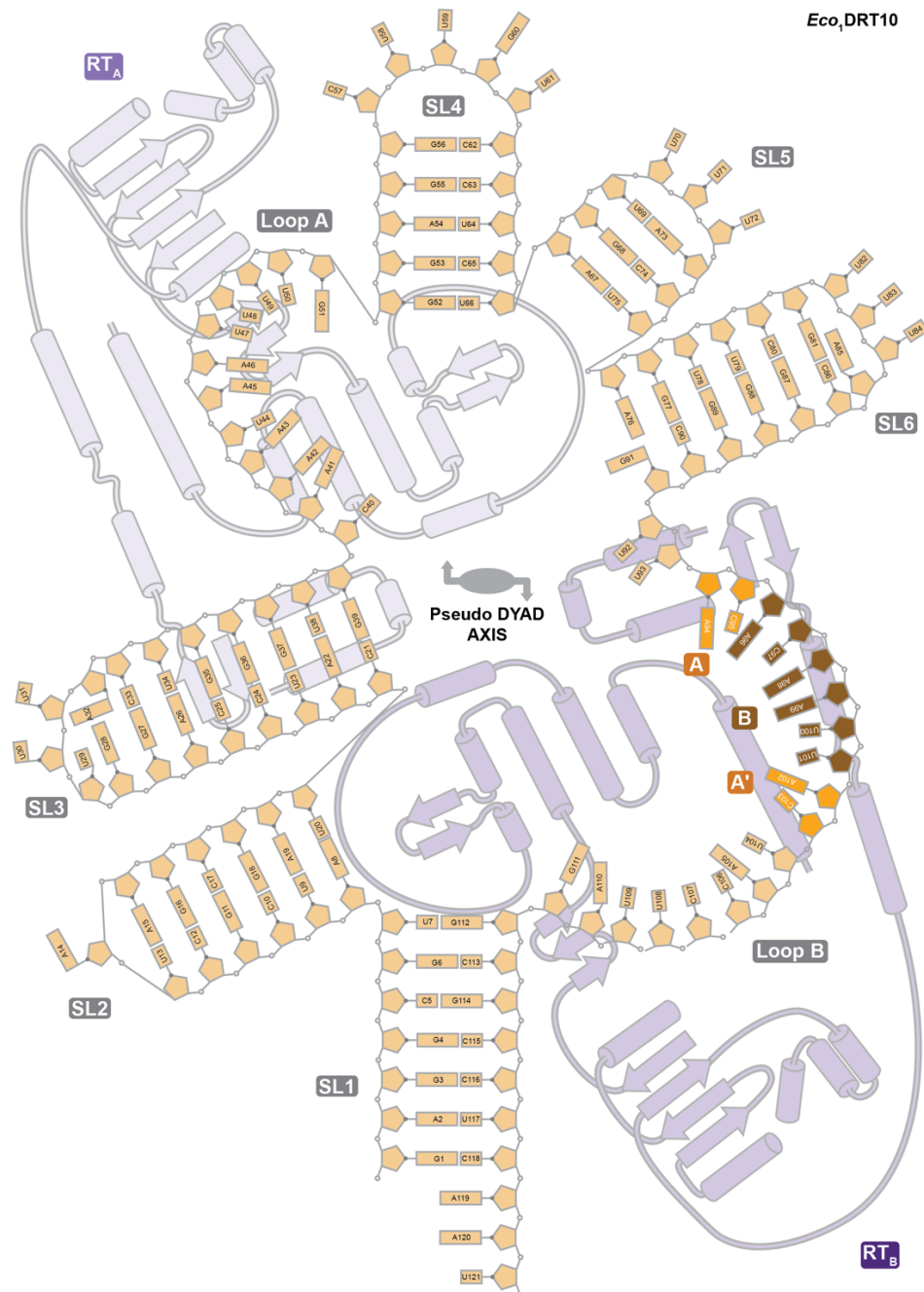
restores efficient synthesis of a 6-nt cDNA repeat with strong periodicity. Data are normalized as in **a**.

Supplementary Figure 3 | Secondary structure topology diagrams of *Eco*₁DRT10 and *Eco*₃DRT10 RTs.



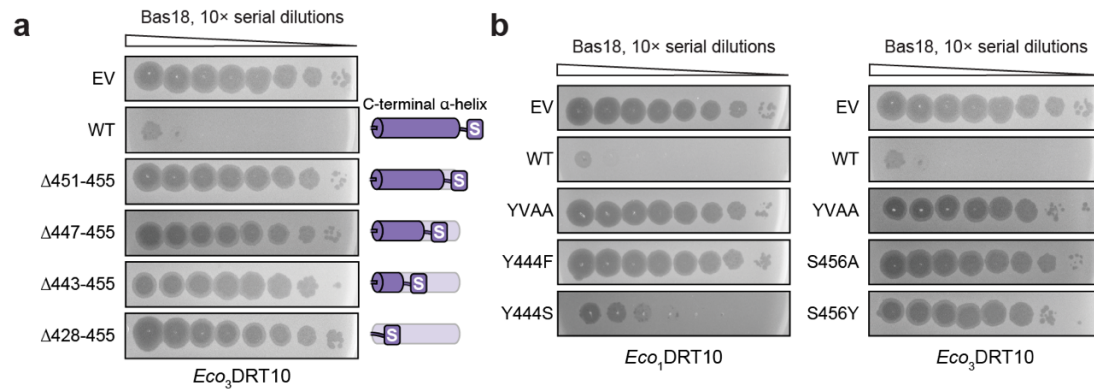
Annotated secondary structure topology diagrams of the RT from *Eco*₁DRT10 (**a**) and *Eco*₃DRT10 (**b**).

Supplementary Figure 4 | Secondary structure and pseudo-symmetric organization of the *Eco*_IDRT10 ncRNA.



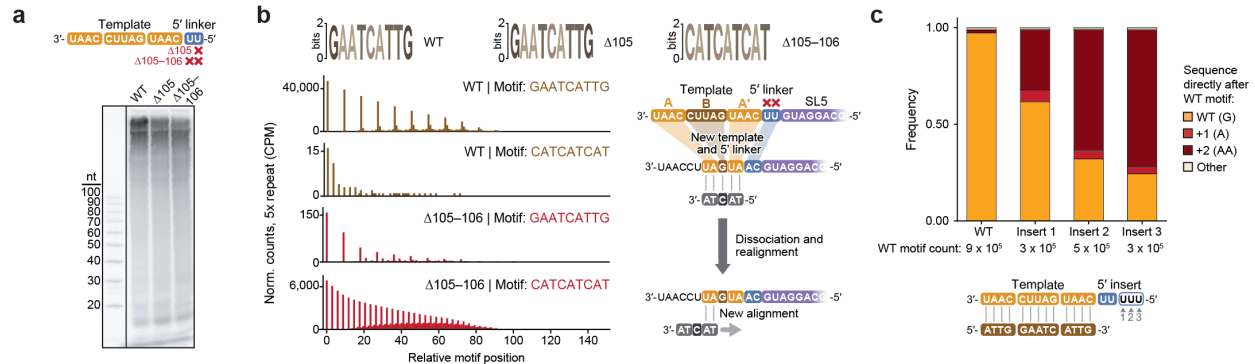
Secondary structure of the *Eco*_IDRT10 ncRNA highlighting stem-loops (SL1–SL6), Loop A, Loop B, and the internal template-containing region. Two structurally related halves form a pseudo-symmetric architecture around a central dyad axis. Regions contacting RT promoters A and B are indicated.

Supplementary Figure 5 | Structural constraints on protein priming within the DRT10 RT catalytic site.

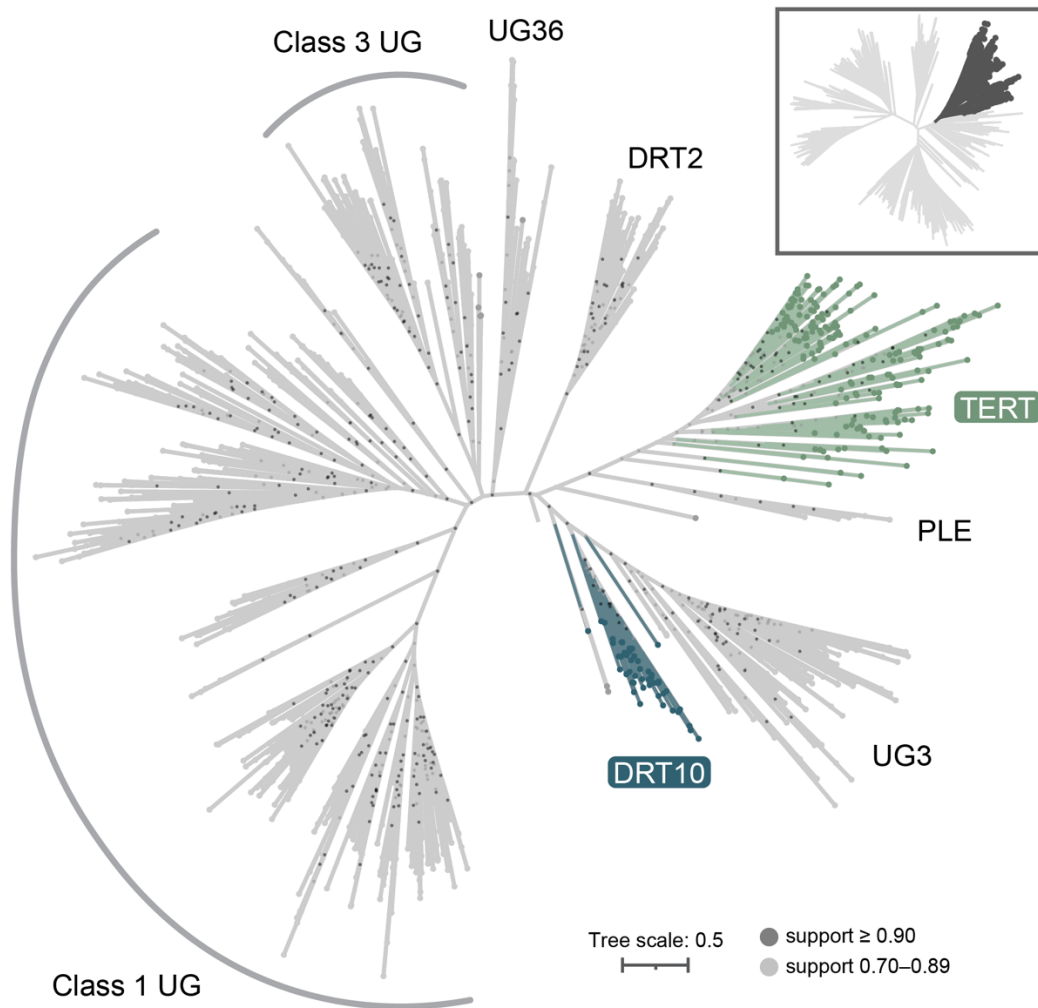


a, C-terminal truncations of *Eco₃DRT10* RT that shorten the α -helix while retaining the priming residue S456, with corresponding plaque assays. **b**, Phage plaque assays of priming residue substitutions in *Eco₁DRT10* and *Eco₃DRT10*, showing partial loss of defense activity in the *Eco₁DRT10* and complete loss of defense activity in the *Eco₃DRT10*. Experiments were repeated twice with similar results.

Supplementary Figure 6 | Analysis of DRT10 ncRNA template 5' linker truncation variants.



a, Denaturing PAGE analysis of $[\alpha\text{-}^{32}\text{P}]\text{dCTP}$ -labeled products from 5' linker truncation variants, showing long products even in variants lacking WT motif repeats by sequencing. **b**, Sequencing analysis of 5' linker truncation variants. *Top*, MEME motif analysis identifies the WT motif in the $\Delta 1$ -nt variant and emergence of a new CAT repeat motif in the $\Delta 2$ -nt variant. *Bottom*, Motif position graphs for WT and CAT repeat motifs. *Bottom right*, Schematic illustrating a putative alternative repeat-addition mechanism in which the RT stops 2 nt before the WT template boundary and uses a new AT re-annealing site to sustain repeat synthesis. **c**, Analysis of template read-through in 5' linker insertions, quantified as nucleotide frequency directly after the WT motif in reads containing that motif.

Supplementary Figure 7 | Phylogenetic relationships between UG and TERT family RTs.

Pruned subtree derived from the larger phylogenetic tree in **Figure 6b**, as shown by the highlighted clade in the inset at top right. RT sequences belonging to the well-supported clade (SH-like local support: 0.987; alignment-ensemble TBE support: 0.844) comprising TERT, PLE, and UG RTs were retained, enabling closer inspection of their evolutionary relationships.

REFERENCES

104. Cojocaru, R. & Unrau, P. J. Transitioning to DNA genomes in an RNA world. *Elife* **6**, e32330 (2017).
105. Mustafin, R. N. & Khusnutdinova, E. K. The Role of Reverse Transcriptase in the Origin of Life. *Biochemistry (Mosc)* **84**, 870–883 (2019).
106. Kazlauskas, D., Krupovic, M., Guglielmini, J., Forterre, P. & Venclovas, Č. Diversity and evolution of B-family DNA polymerases. *Nucleic Acids Res* **48**, 10142–10156 (2020).
107. Venkataraman, S., Prasad, B. V. L. S. & Selvarajan, R. RNA Dependent RNA Polymerases: Insights from Structure, Function and Evolution. *Viruses* **10**, 76 (2018).
108. Varadi, M. *et al.* AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res* **52**, D368–D375 (2024).
109. Černý, J., Černá Bolfíková, B., de A Zanotto, P. M., Grubhoffer, L. & Růžek, D. A deep phylogeny of viral and cellular right-hand polymerases. *Infect Genet Evol* **36**, 275–286 (2015).
110. Haag, J. & Stamatakis, A. Pythia 2.0: New Data, New Prediction Model, New Features. Preprint at <https://doi.org/10.1101/2025.03.25.645182> (2025).
111. Höhler, D., Haag, J., Kozlov, A. M. & Stamatakis, A. A representative Performance Assessment of Maximum Likelihood based Phylogenetic Inference Tools. Preprint at <https://doi.org/10.1101/2022.10.31.514545> (2022).
112. Toro, N., Martínez-Abarca, F. & González-Delgado, A. The Reverse Transcriptases Associated with CRISPR-Cas Systems. *Sci Rep* **7**, 7089 (2017).
113. Toro, N. & Nisa-Martínez, R. Comprehensive phylogenetic analysis of bacterial reverse transcriptases. *PLoS One* **9**, e114083 (2014).
114. Silas, S. *et al.* On the Origin of Reverse Transcriptase-Using CRISPR-Cas Systems and Their Hyperdiverse, Enigmatic Spacer Repertoires. *mBio* **8**, e00897-17 (2017).
115. Edgar, R. C. Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nat Commun* **13**, 6968 (2022).
116. Lemoine, F. *et al.* Renewing Felsenstein’s phylogenetic bootstrap in the era of big data. *Nature* **556**, 452–456 (2018).