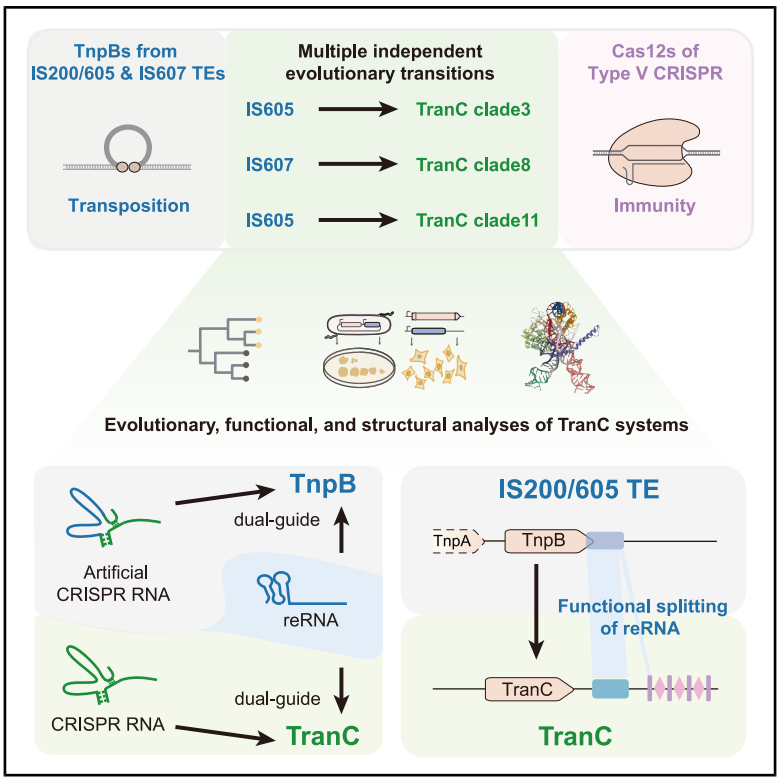
Cell

Functional RNA splitting drove the evolutionary emergence of type V CRISPR-Cas systems from transposons

Graphical abstract



Highlights

- TranC systems are evolutionary intermediates between transposons and Cas12 effectors
- TranCs mediate DNA cleavage guided by both reRNAs and CRISPR RNAs
- Cryo-EM structures reveal a split guide RNA architecture composed of tracrRNA and crRNA
- Functional RNA splitting represents a key step in the evolution of type V CRISPR immunity

Authors

Shuai Jin, Zixu Zhu, Yunjia Li, ..., Yong E. Zhang, Jun-Jie Gogo Liu, Caixia Gao

Correspondence

zhangyong@ioz.ac.cn (Y.E.Z.), junjiegogoliu@tsinghua.edu.cn (J.-J.G.L.), cxgao@genetics.ac.cn (C.G.)

In brief

Identification of TranC systems reveals evolutionary intermediates bridging from transposon-encoded TnpBs and CRISPR-Cas12 effectors. Evolutionary, structural, and functional analyses show that RNA splitting into tracrRNA and crRNA was a pivotal step in the emergence of type V CRISPR systems.

Jin et al., 2025, Cell *188*, 1–18
October 30, 2025
https://doi.org/10.1016/j.cell.2025.09.004

Cell

CellPress

Article

Functional RNA splitting drove the evolutionary emergence of type V CRISPR-Cas systems from transposons

Shuai Jin,[1,7] Zixu Zhu,[1,2,7] Yunjia Li,[1,2,7] Shouyue Zhang,[3,7] Yijing Liu,[1,2] Danyuan Li,[3] Yuanqing Li,[4] Yingfeng Luo,[5] Zhiheng Cheng,[1,2] Kevin Tianmeng Zhao,[6] Qiang Gao,[6] Guanglei Yang,[1,2] Hongchao Li,[1] Ronghong Liang,[1] Rui Zhang,[1] Jin-Long Qiu,[5] Yong E. Zhang,[4,*] Jun-Jie Gogo Liu,[3,*] and Caixia Gao[1,2,8,*]

[1]New Cornerstone Science Laboratory, Center for Genome Editing, Laboratory of Advanced Breeding Technologies, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China

[2]College of Advanced Agricultural Sciences, University of Chinese Academy of Sciences, Beijing, China

[3]Beijing Frontier Research Center for Biological Structure, State Key Laboratory of Membrane Biology, Tsinghua-Peking Center for Life Sciences, School of Life Sciences, Tsinghua University, Beijing, China

[4]State Key Laboratory of Animal Biodiversity Conservation and Integrated Pest Management, Institute of Zoology, Chinese Academy of Sciences, Beijing, China

[5]State Key Laboratory of Microbial Diversity and Innovative Utilization, Institute of Microbiology, Chinese Academy of Sciences, Beijing, China

[6]Qi Biodesign, Beijing, China

[7]These authors contributed equally

[8]Lead contact

*Correspondence: zhangyong@ioz.ac.cn (Y.E.Z.), junjiegogoliu@tsinghua.edu.cn (J.-J.G.L.), cxgao@genetics.ac.cn (C.G.)

https://doi.org/10.1016/j.cell.2025.09.004

SUMMARY

Transposon-encoded TnpB nucleases gave rise to type V CRISPR-Cas12 effectors through multiple independent domestication events. These systems use different RNA molecules as guides for DNA targeting: transposon-derived right-end RNAs (reRNAs or omega RNAs) for TnpB and CRISPR RNAs for type V CRISPR-Cas systems. However, the molecular mechanisms bridging transposon activity and CRISPR immunity remain unclear. We identify TranCs (transposon-CRISPR intermediates) derived from distinct IS605- or IS607-TnpB lineages. TranCs utilize both CRISPR RNAs and reRNAs to direct DNA cleavage. The cryoelectron microscopy (cryo-EM) structure of LaTranC from *Lawsonibacter sp.* closely resembles that of the ISDra2 TnpB complex; however, unlike a single-molecule reRNA, the LaTranC guide RNA is functionally split into a tracrRNA and crRNA. An engineered RNA split of ISDra2 TnpB enabled activity with a CRISPR array. These findings indicate that functional RNA splitting was the primary molecular event driving the emergence of diverse type V CRISPR-Cas systems from transposons.

INTRODUCTION

Mobile genetic elements (MGEs), such as transposable elements (TEs) and viruses, are major drivers of phenotypic evolution across all domains of life.[1–3] A notable example of MGE-driven innovation is the CRISPR-Cas system, which originated from domesticated TEs to confer adaptive immunity in prokaryotes.[4–6] The key effector proteins, Cas9 (type II) and Cas12 (type V), evolved from insertion sequences Cas9-like OrfB (IscB) and transposon-encoded nuclease TnpB (TnpB), respectively, both of which are encoded by IS200/IS605 TEs (and in some cases also by IS607).[7–9] These RNA-guided nucleases utilize CRISPR RNAs (crRNAs), often in conjunction with trans-activating crRNAs (tracrRNAs), to recognize and cleave invading nucleic acids flanked by protospacer-adjacent motifs (PAMs).[4,10] The programmable nature of Cas9 and Cas12 has revolutionized genome editing, enabling both fundamental discoveries and diverse biotechnological applications.[11–13]

Although both nucleases originated from TEs, their evolutionary trajectories diverged markedly. Type II CRISPR-Cas9 systems likely originated from a single IscB domestication event,[8,14–17] whereas type V CRISPR-Cas12 systems exhibit remarkable diversity, suggesting that TnpB nucleases underwent multiple independent domestication events.[7–9,18–21] This recurrent transition from TnpB to Cas12 is thought to be driven by the extraordinary abundance and sequence diversity of TnpB nucleases, which are found at more than one million loci across prokaryotic genomes.[8,20] TnpBs are encoded by several insertion sequence (IS) families, including IS605, IS607, and IS1341.[7,22,23] The IS605 and IS607 families are autonomous, encoding both TnpB and the transposase TnpA (Y1 transposase

and serine recombinase, respectively). By contrast, IS1341 encodes only a TnpB. These TnpBs are guided by transposon-derived right-end RNAs (reRNAs, also known as ωRNAs) to cleave DNA sequences flanked by target-adjacent motifs (TAMs),[8,9] thereby promoting homologous recombination events that restore and retain the mobile elements.[24,25]

It has been proposed that the transition to CRISPR-based immunity occurred through numerous independent insertions of transposons near preexisting CRISPR arrays, followed by the emergence of functional associations between TnpB proteins and these repeats.[5,18,19,26] Phylogenetic analyses support this scenario by revealing multiple independent recruitment events of distinct TnpB ancestors into CRISPR loci.[18,21] This is further consistent with the extensive structural and functional diversity observed among type V systems, including variations in nucleic acid targeting mechanisms and guide RNA processing.[21,27–34] However, the molecular mechanisms that enabled this evolutionary emergence remain poorly understood. The type V-U supergroup is thought to encompass multiple nascent CRISPR subtypes.[19,21] However, experimentally and structurally characterized V-U clades (e.g., Cas12f and Cas12m) differ substantially from TnpBs in protein structure and domain composition and appear to be separated from the initial transformation by a considerable evolutionary period.[33,35–37] Therefore, identifying evolutionary intermediates that capture transitional states between transposition and CRISPR immunity is critical to understanding this remarkable functional transition (Figure 1A).[38,39]

To address this gap, we identified a class of TranCs (transposon-CRISPR intermediates) that exhibit hallmark features of the transposon-to-CRISPR transition. Through integrative evolutionary, functional, and structural analyses, we demonstrate that the functional splitting of reRNA into tracrRNA and crRNA represents a conserved and essential mechanism driving the initial transformation from transposon-encoded TnpBs to type V CRISPR-Cas systems.

RESULTS

Identification of evolutionary intermediates between transposons and CRISPR nucleases

To investigate the evolutionary transition from transposon-encoded TnpBs to type V CRISPR-Cas systems, we sought to identify putative intermediate systems—hereafter referred to as TranCs. To balance sensitivity, specificity, and computational feasibility, we designed an *in-silico* pipeline to mine TranC candidates from high-quality genomic and metagenomic datasets (Figure 1B; Data S1.1; STAR Methods). Given the substantial sequence diversity among TnpB and Cas12 proteins (Figure S1A; domain architectures in Data S1.2),[7,21,26,32] we implemented three mining methods to identify proteins sharing conserved sequence features of both TnpB and Cas12: (1) sequence-based homology searches using BLASTp against reference TnpB and Cas12 sequences, (2) profile-based domain screening using three RuvC-associated hidden Markov models (HMMs) shared by both protein families, and (3) motif-based scanning targeting three conserved motifs located around the catalytic residues (Figure 1B; Data S1.1; STAR Methods). After integrating the results from all approaches and applying multiple rounds of filtering, we identified 146 high-confidence TranC candidates (Figures 1B and S1B–S1D; Data S1.1). These proteins are compact, TnpB-like nucleases located adjacent to CRISPR arrays and may function as nascent CRISPR systems (Table S1).

To classify the TranC candidates within the type V CRISPR framework, we reconstructed a maximum-likelihood phylogeny using 7,566 protein sequences, including previously reported canonical TnpBs, established Cas12 subtypes, uncharacterized V-U members,[19,21] and all 146 newly identified TranC candidates (Figure 1C; Table S1). The resulting phylogeny recapitulated previously described evolutionary relationships among TnpB groups (e.g., typical vs. derived) and known Cas12 families (Figure 1C; Data S1.1).[21,40] After filtering out orphan sequences, we obtained 25 distinct TranC candidate clades nested within diverse TnpB branches (Figure 1C).

Given the extensive diversity of TnpB nucleases and the multiple independent origins of TranCs, we aimed to comprehensively identify additional related TnpB or TranC homologs within each candidate TranC clade to better elucidate their evolutionary trajectories (Figure 2A). To enhance sensitivity and coverage, we conducted both sequence-based and structure-informed homology searches for all members of the 25 candidate clades, utilizing the MGnify metagenomic database and publicly available protein structure repositories, respectively (Figure 2A; STAR Methods). We retrieved 3,538 non-redundant homologs. We then reconstructed 25 clade-specific phylogenies and identified six bona fide TranC clades (clades 3, 8, 11, 12, 13, and 14), in which CRISPR-associated TranC groups consistently formed high-confidence sister-group relationships with specific TnpB lineages (branch support $\geq$ 80%; Figures 2B–2E; full trees in Figures S2A–S2C; Data S1.3–S1.5). In addition to the TranC candidates, these six clades also contained newly identified closely related TnpB proteins and uncharacterized V-U members (Table S1). Clade8, which includes Cas12n homologs,[40] was termed clade8-Cas12n. Consequently, we identified these six as bona fide TranC clades, each comprising a TranC group and a TnpB sister group sharing a common ancestor (Figures 2A and 2B).

We next examined the TranCs' association with CRISPR arrays and the presence of the DED catalytic residues (Figures 2B–2E and S2A–S2C; Data S1.3–S1.5). Clade3 comprises five subclades, each containing at least three members (Figures 2B and 2C). Among them, clade3-subclade3 lacks an associated CRISPR array, while clade3-subclade4 exhibits substitutions or loss of the DED catalytic residues. By contrast, the remaining three subclades contain both a CRISPR array and the conserved DED motif. Similarly, we observed that one of four subclades in clade8-Cas12n and one of five subclades in clade11 lacked a stable association with a CRISPR array (Figures 2D and 2E). Clade12, clade13, and clade14 were distinct in that none of their subclades contained the conserved DED catalytic motif, suggesting a potential loss or alteration of nuclease activity (Data S1.3–S1.5). These findings suggest a transposon domestication scenario for TnpBs characterized by an unstable functionality with a CRISPR array and the DED catalytic residues, supporting TranC's nascent status. Further spacer analysis of the putatively functional clade3, clade8-Cas12n, and clade11 identified 103 spacers predicted to target 55 viral genomes, 40 viral

Please cite this article in press as: Jin et al., Functional RNA splitting drove the evolutionary emergence of type V CRISPR-Cas systems from transposons, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.004

CellPress

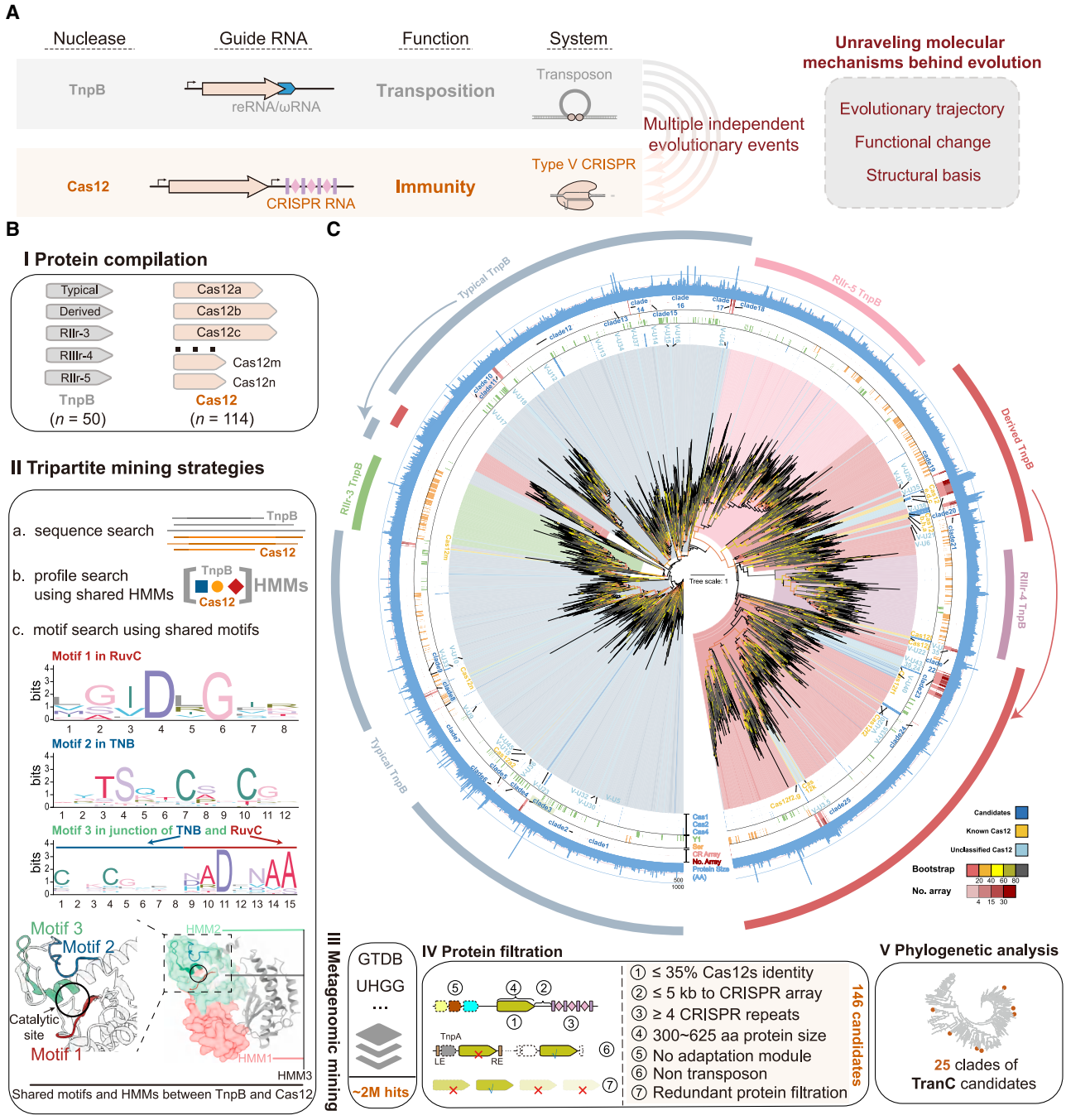


Figure 1. Metagenomic mining and phylogenetic analyses of TranC candidates

(A) Conceptual framework illustrating the evolutionary transition from TnpB-encoded transposons to CRISPR-Cas12 systems.

(B) Overview of the *in silico* mining pipeline used to identify TranC candidates. Three complementary mining methods were applied: (a) sequence-based search using reference proteins, (b) profile-based search using shared hidden Markov models (HMMs), and (c) motif-based search targeting conserved catalytic residues.

(C) Maximum-likelihood phylogenetic tree comprising 7,566 protein sequences, including known TnpBs, TranC candidates, unclassified Cas12s (V-Us), and previously classified Cas12 subtypes. Colored rings indicate the presence or enrichment of elements (Cas1, Cas2, Cas4, Y1-type TnpA, serine-type TnpA, and CRISPR arrays) as well as the number of CRISPR arrays ("no. array") and the amino acid (aa) length. The outermost ring indicates TnpB classification. The three representative candidate TranC clades were separately highlighted using distinct background colors. Branch support is visualized by the color intensity of each node.

See also Figure S1 and Data S1.

Please cite this article in press as: Jin et al., Functional RNA splitting drove the evolutionary emergence of type V CRISPR-Cas systems from transposons, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.004

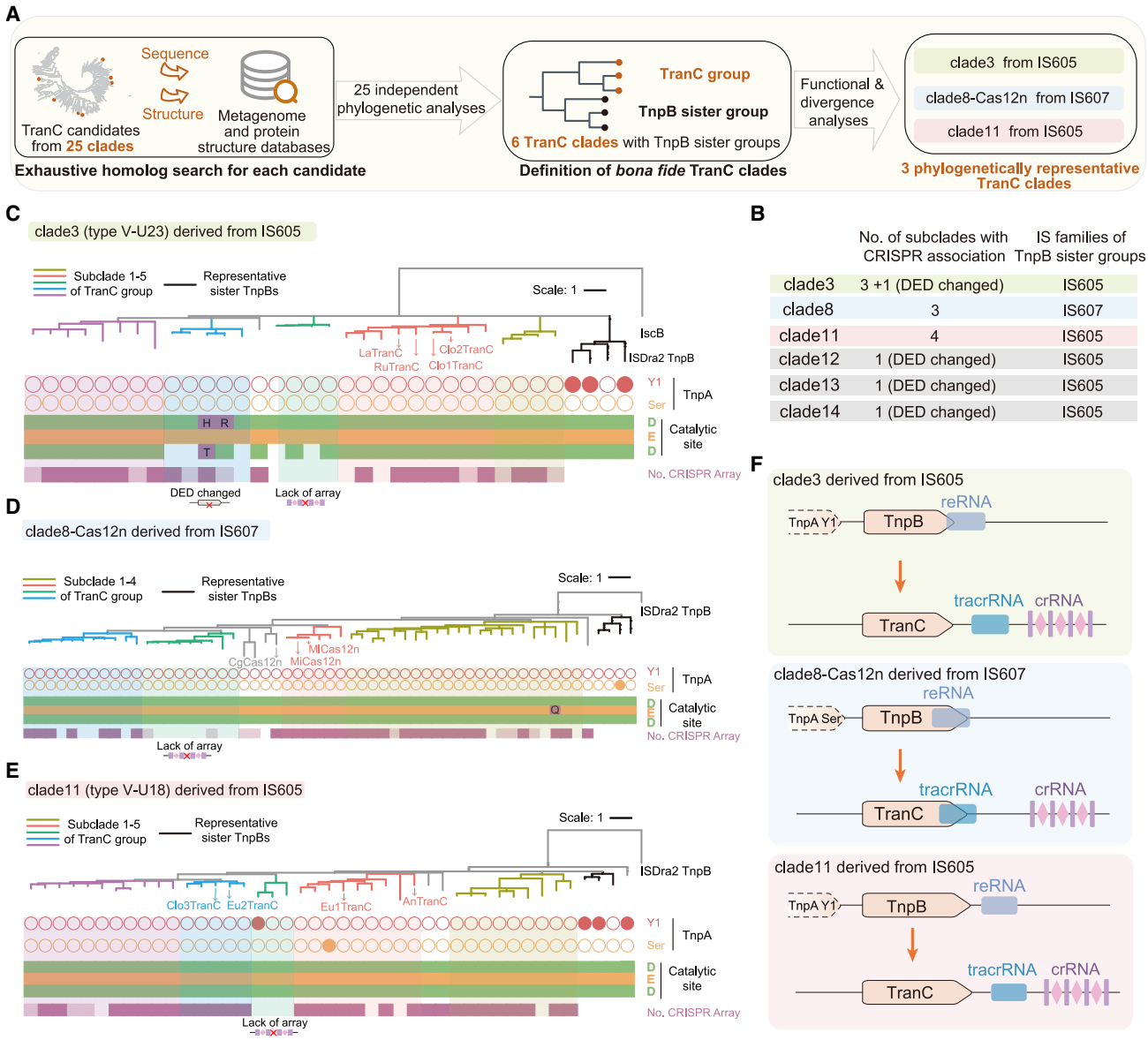


Figure 2. Identification and classification of TranC clades

(A) Schematic illustrating the identification of bona fide TranC clades based on refined phylogenetic inference.

(B) Overview of six bona fide TranC clades. The conserved DED catalytic triad is characteristic of TnpB-like proteins. Some subclades show substitutions or loss of DED catalytic residues, suggesting impaired or absent nuclease activity. Clade12 through clade14, characterized by altered DED motifs, are highlighted with uniform gray backgrounds.

(C–E) Zoomed-in views of phylogenetic trees for clade3 (C), clade8-Cas12n (D), and clade11 (E), corresponding to Figures S2A–S2C. Distinct subclades, orphan lineages, and sister TnpBs are color-coded. Colored circles and bars beneath the trees indicate the presence of associated elements, including Y1-type TnpA transposase (Y1), serine recombinase-type TnpA transposase (Ser), DED catalytic motifs, and the number of CRISPR arrays ("no. CRISPR array"). Subclade4 of clade3 exhibits DED substitutions or deletions and is labeled "DED changed." Subclades lacking CRISPR arrays are labeled "lack of array." Lighter shading indicates repeat counts < 4.

(F) Gene organization of three representative TranC clades. Dashed boxes indicate the optional presence of TnpA genes.

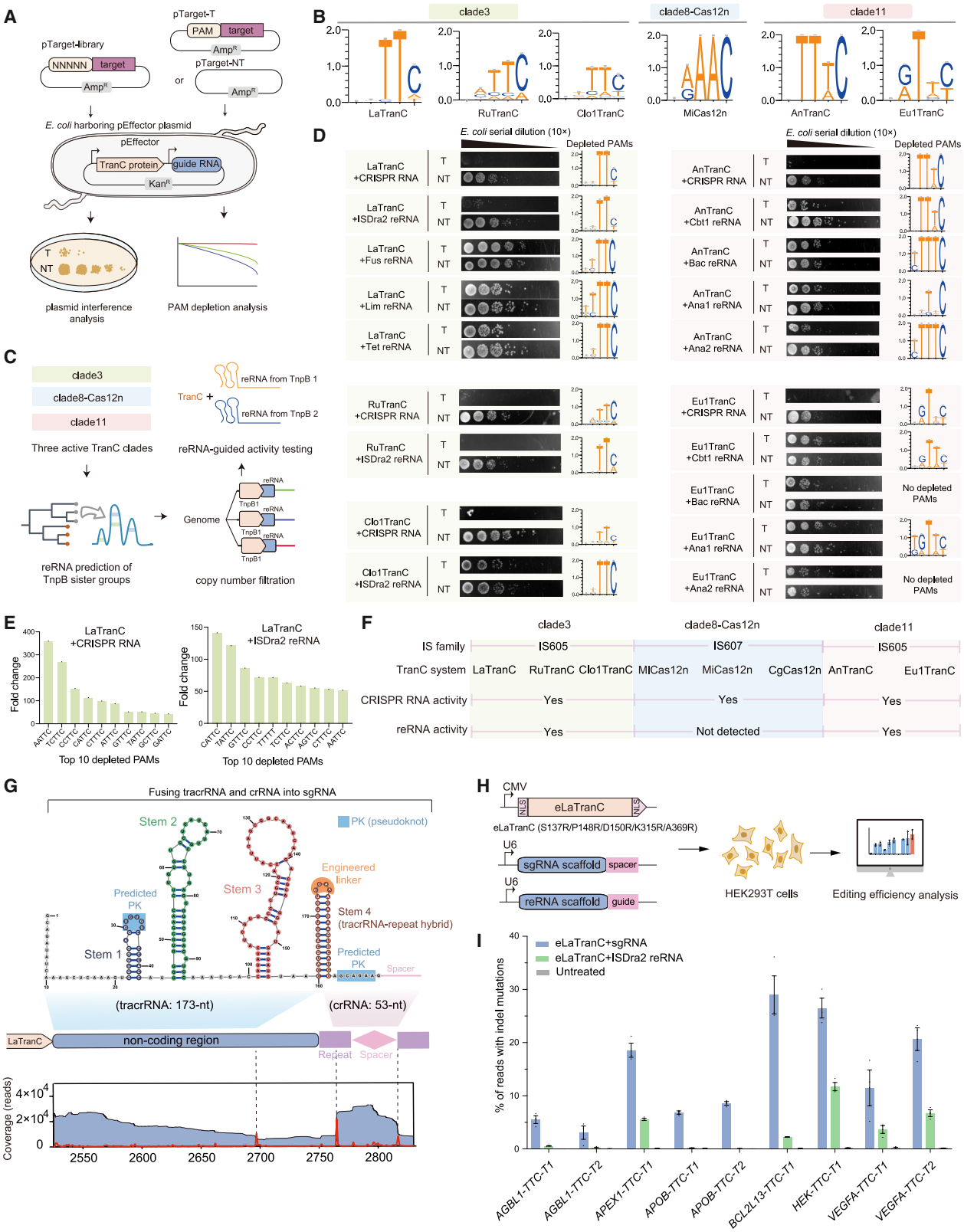
See also Figure S2 and Data S1.

metagenomic sequences, and 8 plasmids, supporting a potential immune function (Table S1 and STAR Methods).

We next focused on the distinctions among these three clades. TnpA association analysis on each clade's respective TnpB sister groups revealed that clade3 and clade11 originated from IS605 TnpBs (sister TnpBs being associated with TnpA Y1 transposases), while clade8-Cas12n was derived from IS607 TnpBs (associated with Ser transposases; Figures 2C–2E and S2A–S2C). To analyze gene organization, we identified reRNAs and tracrRNA-crRNA pairs located downstream of TnpB or TranC

Please cite this article in press as: Jin et al., Functional RNA splitting drove the evolutionary emergence of type V CRISPR-Cas systems from transposons, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.004

CellPress



(legend on next page)

Please cite this article in press as: Jin et al., Functional RNA splitting drove the evolutionary emergence of type V CRISPR-Cas systems from transposons, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.004

open reading frames (ORFs) using clade-specific RNA covariance models (Figure 2F; STAR Methods). Clade8-Cas12n exhibited the highest degree of genetic overlap between the effector protein and guide RNAs (Figures 2F and S2D). On average, the TnpB ORFs overlapped with 88.4% of the reRNA sequence, while Cas12n ORFs overlapped with 65.0% of their respective tracrRNA sequences. By contrast, clade3 and clade11 generally showed little or no overlap, except for the TnpBs in clade3, which showed a modest 22.6% overlap. In addition, these three clades are scattered across evolutionarily distant TnpB branches, representing high sequence diversity (Figure 1C). Collectively, we identified three diverse phylogenetically representative TranC clades, each associated with a specific TnpB sister group and differing in their transposon family origins, gene organizations, and phylogenetic positions. These clades represent the initial transformation stages in multiple independent evolutionary trajectories from TnpB to type V CRISPR effectors.

TranCs can be guided by both crRNA and reRNA

Given their intermediate evolutionary position between transposon-encoded TnpBs and Cas12 effectors, TranC systems may retain the ancestral ability to utilize both reRNAs and crRNAs (with tracrRNA) for DNA targeting. To test this hypothesis, we evaluated the activity of TranC nucleases in *E. coli* using both a PAM depletion assay and a plasmid interference assay (Figures 3A–3D). The PAM depletion assay profiles reRNA-guided nuclease requirements via deep sequencing, and the plasmid interference assay assesses nuclease activity against invading DNA (Figure 3A).[8,20,40,41] We synthesized DNA constructs encoding TranC effectors together with a minimal CRISPR locus (harboring a putative tracrRNA region and a reprogrammed CRISPR array targeting the *GFP-T1* site in pTarget) and cloned them into a pEffector vector (Figures 3A and S3A; STAR Methods).

We first evaluated crRNA-guided activity of 11 TranC members from three representative TranC clades, including clade3, clade8-Cas12n, and clade11 (Figures 2C–2E; Table S1). These TranCs were selected in subclades predicted to be functional based on stable CRISPR array association and conservation of the DED catalytic triad (Figures 2C–2E). Eight out of eleven TranC nucleases showed specific PAM recognition (Figures 3B and S3B–S3D; Table S1). LaTranC (from *Lachnospiraceae sp.*), RuTranC (from *Ruminiclostridium sp.*), and Clo1TranC (from *Clostridia sp.*) from clade3 recognized 5′-TTC PAMs; AnTranC (from *Anaeromassilibacillus senegalensis*) and Eu1TranC (from *Eubacteriales sp.*) from clade11 recognized a 5′-TTTC PAM and a 5′-RTTY PAM, respectively; and MiCas12n (from *Micrococcus sp.*) from clade8-Cas12n recognized a 5′-RAAC PAM. Previously reported MlCas12n and CgCas12n exhibited preferences for 5′-AAN and 5′-AAC PAMs, respectively, consistent with previous studies (Figures S3B–S3D).[40] We subsequently performed the plasmid interference assay to test whether these eight TranCs could cleave invading DNA in *E. coli* and found that indeed all eight systems exhibited robust interference activity (Figures 3D and S3B–S3D).

Next, we tested whether these eight TranCs could also function with reRNAs derived from closely related TnpBs identified in the clade-specific phylogenetic analyses (Figure 3C; Table S1; STAR Methods). In clade3 derived from IS605, all three tested TranCs (LaTranC, RuTranC, and Clo1TranC) were active when paired with reRNA from well-characterized insertion sequence from *Deinococcus radiodurans* (ISDra2) TnpB, recognizing a conserved 5′-TTC PAM, as evidenced by PAM depletion assays (Figures 3D and 3E). Plasmid interference assays also confirmed robust cleavage activity for the ISDra2 reRNA-guided LaTranC and RuTranC systems (Figure 3D). LaTranC, which is most similar to ISDra2 TnpB, was further tested with three metagenome-derived TnpB-reRNAs (Fus TnpB, Lim TnpB, and Tet TnpB, Tables S1 and S2). All three were active but showed lower fold changes than the ISDra2 reRNA in PAM depletion assays and failed to support notable plasmid interference, indicating weaker activity (Figures 3D and S3G).

In clade11, which is derived from another IS605 lineage, AnTranC and Eu1TranC were tested with four reRNAs (from Cbt1 TnpB, Bac TnpB, Ana1 TnpB, and Ana2 TnpB). AnTranC was active with all four reRNAs in PAM depletion assays (5′-TTTC PAM), and the Cbt1 reRNA supported detectable plasmid interference (Figure 3D). Eu1TranC exhibited nuclease activity with Ana1 and Cbt1 reRNAs among the four reRNAs in PAM depletion assays within the 5′-RTTY PAM contexts (Figure 3D), although the fold-change of Cbt1 was lower (Figure S3H). None of the tested reRNAs yielded notable plasmid interference

Figure 3. TranCs exhibit nuclease activity guided by both crRNA and reRNA

(A) Schematics of the *E. coli* plasmid interference and PAM depletion assays. Candidate TranC systems cleave the *GFP-T1* target site encoded on the pTarget plasmid either with a specific PAM sequence (left, plasmid interference) or a PAM library (right, PAM depletion), resulting in loss of ampicillin (Amp) resistance and cell death.

(B) WebLogos of PAM motifs recognized by six TranCs from clade3, clade8-Cas12n, and clade11. TranC and TnpB names in this study were derived from the first two or three letters of their most specific available taxonomic annotations.

(C) Workflow for reRNA-guided activity testing. Covariance models derived from TnpB sister groups were used to identify candidate reRNAs, which were filtered by genomic copy number prior to experimental validation.

(D) Results of plasmid interference and PAM library depletion assays evaluating reRNA-guided dsDNA cleavage by TranCs. "T" and "NT" denote target and non-target plasmids, respectively. WebLogos were generated from PAMs depleted >10-fold, except for LaTranC with Fus, Lim, and Tet reRNAs, and Eu1TranC with Cbt1 reRNA, where a >1.5-fold depletion threshold was used due to weaker activity (Figures S3G and S3H).

(E) Top ten most-depleted PAM sequences from LaTranC PAM depletion assays using either crRNA or reRNA guides.

(F) Summary of guide RNA interchangeability observed in clade3, clade8-Cas12n, and clade11.

(G) Identification of mature tracrRNA and crRNA by RNA sequencing in *E. coli*. Read depth is shown in blue, and sharp changes in 5′ and 3′ read termini (red peaks) indicate RNA maturation sites. The sgRNA used in gene editing assays was constructed by fusing the identified tracrRNA and crRNA. The predicted secondary structure is shown above the gene diagram.

(H and I) Schematic (H) and editing efficiencies (I) of eLaTranC, an engineered variant with enhanced editing activity, guided by either sgRNA or reRNA at nine endogenous loci in HEK293T cells (n = 3, mean ± SEM). "*HEK-TTC-T1*" is an intergenic target site, while the remaining sites are located within genes.

See also Figure S3 and Data S1.

(Figure 3D). By contrast, none of the five reRNAs tested for clade8-Cas12n (IS607-derived) yielded detectable nuclease activity in both assays. Specifically, reRNAs from Bif1 TnpB, Bif2 TnpB, and TnpB were tested with MlCas12n and MiCas12n, while Bif1 TnpB and Bif3 TnpB were paired with CgCas12n (Figures S3B–S3D). To experimentally validate our computational predictions, and given the importance of the reRNA scaffold's 3′ end for cleavage activity,[20] we performed 3′ extension assays on five TranC members' reRNA scaffolds from clade8-Cas12n and clade11. In clade11, only the unextended reRNAs, not those with 3′ extensions, led to depletion at the genuine 5′-TTTC and 5′-RTTY PAMs, which confirms the accuracy of our 3′ end computational prediction (Figures S3E and S3F; Data S1.13). Nonetheless, all three Cas12n proteins remained inactive (Figures S3B–S3D). Therefore, five IS605-derived systems support both crRNA and reRNA guidance in *E. coli*. By contrast, the IS607-derived clade8-Cas12n lacks detectable reRNA-guided activity, potentially due to constraints from extensive protein-RNA coding overlap or some subtle but essential structural incompatibilities between the protein and reRNA (Figure 2F; see Discussion).

LaTranC gene editing with guide RNA interchangeability in human cells

To test guide RNA function in eukaryotic cells, we focused on the clade3 LaTranC due to its high sequence and structural similarity to ISDra2 TnpB (Figure 2C; Table S1). In mammalian genome-editing systems, the mature tracrRNA and crRNA are typically fused into a single-guide RNA (sgRNA).[10,11] We performed RNA-seq on *E. coli* expressing LaTranC and its intrinsic CRISPR array (Figure 3G). Deep sequencing identified a 173-nt tracrRNA paired with a 53-nt mature crRNA located immediately downstream of LaTranC (Figure 3G; Data S1.12), resembling the architecture of reRNAs downstream of TnpBs (Figure 2F). Subsequently, by fusing the mature tracrRNA and crRNA and optimizing truncation, we generated a 195-nt sgRNA active in human cells (Data S1.12). To increase the editing efficiency, we engineered LaTranC through five rounds of arginine scanning mutagenesis (Figures S3I–S3K; STAR Methods), generating eLaTranC, which showed an 8.7- to 56.8-fold increase in editing activity over wild-type LaTranC across four endogenous mammalian cell target sites (Figure S3K; Table S2). We next tested the guide RNA interchangeability of the eLaTranC system across nine endogenous target sites in HEK293T cells (*AGBL1-TTC-T1*, *AGBL1-TTC-T2*, *APEX1-TTC-T1*, *APOB-TTC-T1*, *APOB-TTC-T2*, *BCL2L13-TTC-T1*, *HEK-TTC-T1*, *VEGFA-TTC-T1*, and *VEGFA-TTC-T2*) with a 5′-TTC PAM (Figures 3H and 3I; Table S2). Deep sequencing revealed that the editing efficiencies of sgRNA-guided eLaTranC (3.1% to 29.0%, average 14.5%) consistently outperformed reRNA-guided editing (0.1% to 11.7%, average 3.4%) (Figures 3H and 3I). Together, these results indicate that LaTranC from clade3 can mediate gene editing under the guidance of either crRNA or reRNA.

The LaTranC complex has higher structural similarity to the ISDra2 TnpB complex than other characterized type V CRISPR systems

The nascent evolutionary status of TranC systems and their ability to utilize both reRNA and crRNA (tracrRNA-crRNA hybrid) suggest that their ternary complexes (TranC-sgRNA-DNA) may closely resemble those of TnpB-reRNA-DNA. To assess this, we first performed AlphaFold-based protein structural predictions on representative TranC-TnpB pairs from three representative clades. Each TranC and its sister TnpB exhibited highly similar domain architectures (Figures S4A and S4B). Consistent group-level analyses of 333 sister TnpBs further revealed high structural similarity with their corresponding TranCs (template modeling [TM] score > 0.5), significantly exceeding that observed between randomly paired Cas12a/b/c groups ($p < 0.0001$; Figures 4A–4C). Given the current scarcity of reliable tools for predicting the structure of protein-nucleic acid complexes, we then experimentally determined the cryoelectron microscopy (cryo-EM) structure of the LaTranC-sgRNA-DNA complex.

Although known type V CRISPR systems and TnpBs typically favor guide spacers of about 20–30 bp for DNA interference, TnpBs are highly compact and structurally incapable of generating or maintaining a complete R-loop required for DNA cleavage.[30,31,37,42–44] To gain insight into the structural transition from TnpB to LaTranC, we expressed and purified an LaTranC and sgRNA complex and reconstructed a complex generating a partial R-loop, with only 6-bp complementarity between the sgRNA spacer and the PAM-proximal target strand (2.96 Å; Data S2.1. The cryo-EM map revealed a ternary complex composed of a single LaTranC protein, the sgRNA, and double-stranded DNA (dsDNA) substrate in a partial R-loop conformation (Figures 4D and 4E).

LaTranC has six structural domains, including wedge (WED), recognition (REC), RuvC nuclease (RuvC), bridge helix (BH), target nucleic acid-binding (TNB), and C-terminal (CTD) domains, organized as in TnpBs (Figures 4D and 4E).[30,31] Like TnpBs and the known Cas12 proteins, it adopts a bi-lobed structure, with the recognition lobe (REC), comprising the REC and WED domains, responsible for recognizing a 5′-TTC PAM sequence (Figures S5A–S5D), and the nuclease lobe (NUC), comprising the RuvC, BH, TNB, and CTD domains, responsible for DNA cleavage (Figures 4D and 4E). The three motifs shared across TnpBs and Cas12 proteins (Figure 1B) are situated within the cleavage pocket of the NUC lobe (in the RuvC and TNB domains; Figure S5E).

We quantified the structural similarity between the ISDra2 TnpB complex and LaTranC, as well as Cas12 proteins (including Cas12f, Cas12m, Cas12i, Cas12a, Cas12b, Cas12e, Cas12j, and Cas12k), using established size-independent methods (TM score; Figures 4F and S5F).[45,46] The similarity between the LaTranC protein and its closely related ISDra2 TnpB protein, as measured by TM score, was 0.85, which is higher than any other Cas12 variant with experimentally determined structures (Figures 4F and S5F). Consistently, the only notable difference was an enlarged BH domain in LaTranC, which likely enhances nucleic acid interactions by increasing its positively charged surface (Figures S5G and S5H). By contrast, other Cas12 proteins contain either inactive mutations, form dimers, or have acquired additional domains (Figure 4G). Together, these findings demonstrate that TranC proteins such as LaTranC remain structurally close to their TnpB ancestors, supporting a model in which protein-level divergence played a limited role during the transition from mobile transposons to CRISPR-Cas immunity systems.

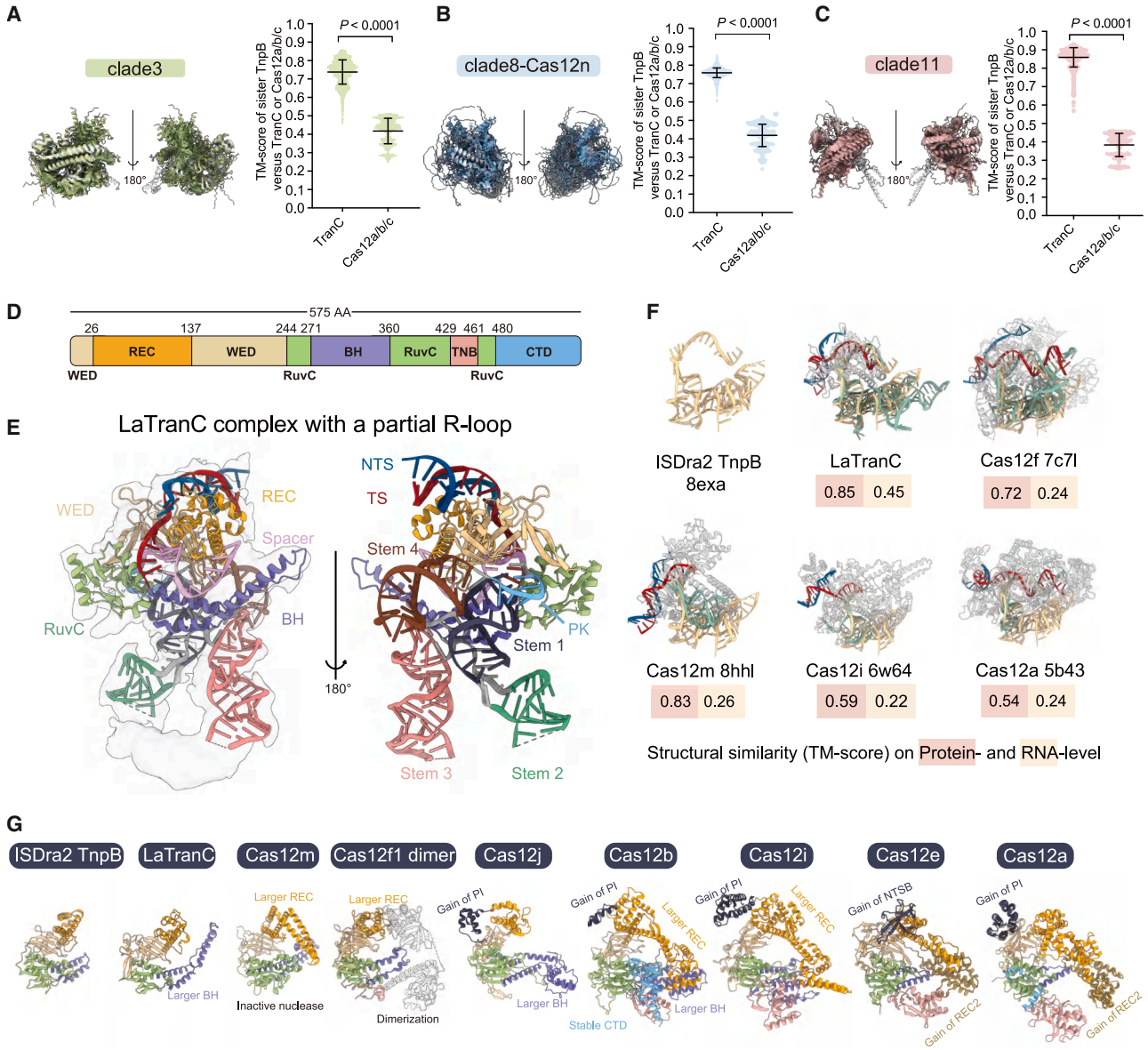


Figure 4. Comparative structural analyses of LaTranC, TnpB, and representative Cas12 proteins

(A–C) Structural comparisons between AlphaFold2-predicted models of TranCs and their sister TnpBs from clade3 (A), clade8-Cas12n (B), and clade11 (C). Cas12a/b/c representatives served as negative controls. n = 4,770, 7,182, and 1,599 for TnpB-TranC comparisons; n = 954, 798, and 246 for TnpB-Cas12a/b/c comparisons. Statistical significance was determined using the Mann-Whitney U test ($p < 0.05$). Data are shown as mean ± SD.

(D) Domain architecture of the LaTranC protein. The RuvC domain is split into three subdomains (RuvC-I, II, and III), a hallmark shared with Cas12 proteins.

(E) Atomic model of the LaTranC-sgRNA-DNA ternary complex resolved in a partial R-loop conformation (2.96 Å). Unsharpened cryo-EM maps are shown as gray outlines. Residues 423–464 and 479–575, mainly in the TNB and CTD domains, are disordered and not included in the model.

(F) Structural alignment of the ISDra2 TnpB complex with LaTranC and five Cas12 orthologs. Despite a high TM score, Cas12m lacks nuclease activity.

(G) Structural overlay of cryo-EM-resolved protein components from ISDra2 TnpB, LaTranC, and representative Cas12 effectors. Domain coloring corresponds to that shown in (A).

See also Figures S4 and S5 and Data S2.

reRNA splitting enabled the evolutionary emergence of type V CRISPR systems

We further compared RNA-level changes among TranC, TnpB, and Cas12s. The RNA component of the LaTranC cryo-EM complex showed the highest similarity to ISDra2 TnpB (RNA-level TM score: 0.45 for LaTranC, 0.22–0.26 for other Cas12 systems; Figures 4F and S5F). However, unlike TnpB complexes that utilize a single reRNA, the LaTranC system encodes a tracrRNA-crRNA hybrid with a semi-ring architecture comprised of four stems (stems 1–4) and a pseudoknot (PK) between the CRISPR repeats and stem 1 (Figures 5A and 5B). The stem 1-PK and stem 4 between TnpB and LaTranC display a high degree of structural

Please cite this article in press as: Jin et al., Functional RNA splitting drove the evolutionary emergence of type V CRISPR-Cas systems from transposons, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.004

CellPress

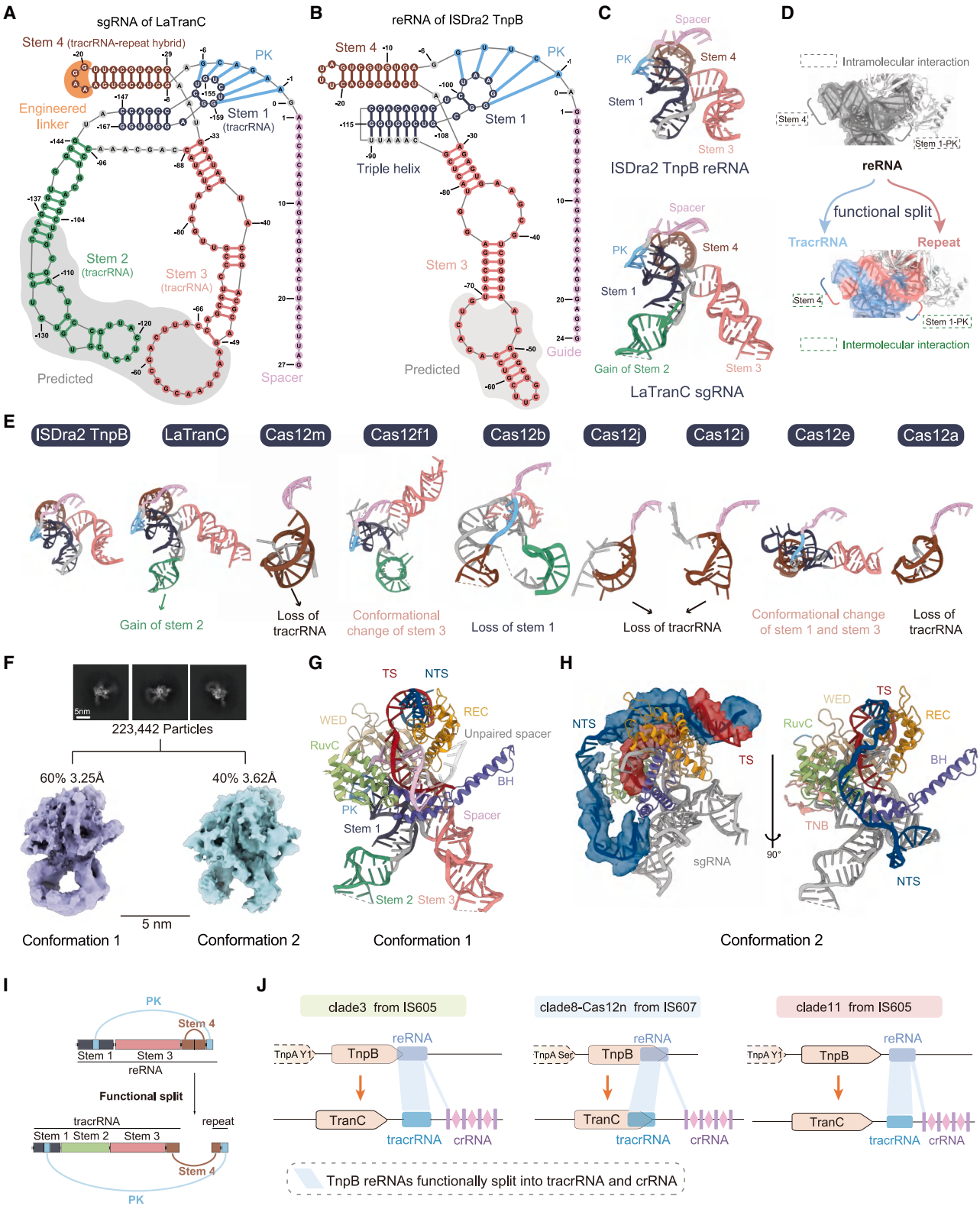


Figure 5. RNA-level changes underpin the transition from TnpBs to type V CRISPR systems

(A and B) Schematic representations of the guide RNAs in LaTranC (sgRNA, A) and ISDra2 TnpB (reRNA, B; PDB: 8EXA).[31] Color coding is consistent across panels: black for stem 1, green for stem 2, magenta for stem 3, brown for stem 4, blue for PK, and mauve for the spacer.

(legend continued on next page)

Please cite this article in press as: Jin et al., Functional RNA splitting drove the evolutionary emergence of type V CRISPR-Cas systems from transposons, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.004

preservation (Figures 5A–5C). Stem 1-PK interacts directly with the protein (Figures S6A and S6B), exhibiting high similarity between LaTranC and ISDra2 TnpB, but divergent configurations among various Cas12s (Figures 5C–5E).[30,31,36,37,43,44,47] Stem 4, which is formed by base pairing between the repeat and anti-repeat sequences in the LaTranC crRNA and tracrRNA, resembles that of stem 4 of the TnpB-reRNA (Figures 5A–5C). Therefore, stem 1-PK and stem 4 of the ISDra2 reRNA engage in an intramolecular interaction, while the two stems in LaTranC engage in an intermolecular interaction, as indicated by the tracrRNA and CRISPR repeat (Figure 5D), and this supports the hypothesis that the reRNA underwent an evolutionary split to become a tracrRNA-crRNA hybrid.[30,31] By contrast, stem 3 exhibits relatively lower conservation across ISDra2 TnpB and CRISPR-Cas12 systems and adopts different conformations (Figure 5E). Stem 2 is the least conserved and appears to have emerged only recently, as it lacks a counterpart in ISDra2 reRNA (Figures 5A and 5B). Sequence comparison across LaTranC homologs revealed a significant overrepresentation of point mutations in stem 2 (binomial test $p = 0.0016$; Figure S6C), suggesting low levels of functional constraint. To directly evaluate the functional importance of these two stems, we generated six sgRNA variants with different truncation sizes (Figure S6D). Individually truncating or deleting stem 2 and stem 3 of the sgRNA (v1-v5) did not significantly affect the editing activity of eLaTranC in HEK293T cells (Figure S6D). The shortest variant, sgRNA-v6, which has both stems deleted, had significantly decreased editing efficiency but remained active (Figure S6D). These results suggest that although these regions are individually dispensable for CRISPR function, they may collectively play auxiliary roles in RNA assembly or stability.

To further validate the structural findings obtained from the partial R-loop, we reconstituted a LaTranC ternary complex with a full R-loop and observed two distinct conformations (Figure 5F; Data S2.2). These conformations differed only slightly from the partial R-loop structure (Figure 4E). In conformation 1 (3.25 Å; ~60% of particles), the PAM-distal portion of the spacer (7A to 11G) base pairs with the target strand (21dT to 17dC; "d" indicates DNA; Figure 5G), and conformation 2 (3.62 Å; ~40% of particles) differed only in a few respects, mainly at the RNA level. The TNB domain density was more stabilized and interacted with the central regions of the non-target strand (14dG to 16dG; Figure 5H) as in Cas12 proteins.[27,44] Stem 3 was highly flexible, adopting a distinct conformation that interacts with the non-target strand at the PAM-distal end (Figures 5H and S6G), rather than forming a semi-ring with stem 2 (Figure 5G).

Beyond the functional split model observed in clade3 LaTranC (Figure 5I), we extended this observation to clade8-Cas12n and clade11 by comparing covariance-folded secondary structures of reRNAs and tracrRNA-crRNA hybrids (STAR Methods). These analyses revealed significant structural similarity between reRNAs and their corresponding crRNAs in clade8-Cas12n and clade11 (threshold E-values < 1e–5; all observed E-values < 1e–7) (Figures S7A–S7C) and showed the tracrRNA-crRNA hybrids in both clades closely resembled the reRNA structures from their corresponding TnpB sister clades (Figures S7D–S7F). Collectively, these results support a convergent evolutionary model in which the functional modularization of reRNA into tracrRNA and crRNA—rather than major protein-level divergence—was a key molecular innovation enabling the transition from transposon-based TnpB systems to CRISPR-Cas adaptive immunity across multiple evolutionary lineages (Figure 5J).

Artificial reRNA splitting is sufficient to convert TnpB-reRNA into a nascent CRISPR system

Based on the aforementioned structural data, we hypothesized that the functional splitting of the reRNA was sufficient to drive the transition from TnpB-reRNA to type V CRISPR systems. We therefore sought to test whether artificially engineering the reRNA could convert a TnpB-reRNA to a "nascent CRISPR system" composed of a TnpB protein functionally associated with a LaTranC CRISPR array.

We constructed a pEffector vector encoding ISDra2 TnpB, the CRISPR array derived from LaTranC, and a chimeric tracrRNA modified from the reRNA (Figure 6A). Specifically, to mimic the PK region, we replaced the ISDra2 reRNA stem 1 with the LaTranC tracrRNA stem 1 to complement the 3′ repeat sequence of the LaTranC CRISPR repeat, and to mimic the tracrRNA-crRNA hybrid, we modified the reRNA stem 4 to be complementary to the 5′ repeat sequence of the LaTranC CRISPR repeat (Figures 6B and 6C). This chimeric tracrRNA could, in principle, interact with the LaTranC CRISPR repeat at two regions of intermolecular interaction (Figure 6C). Subsequently, we performed both a PAM depletion assay and a plasmid interference assay to evaluate the activity of ISDra2 guided by the chimeric tracrRNA and LaTranC CRISPR array (Figure 6D). Wild-type ISDra2 TnpB-reRNA recognizes a 5′-TTGAT TAM sequence.[9] The PAM depletion assay confirmed that ISDra2 TnpB guided by a chimeric tracrRNA-CRISPR array retained its recognition of the 5′-TTGAT PAM and exhibited notable plasmid interference activity (Figure 6D). By contrast, both the wild-type ISDra2 reRNA and the LaTranC tracrRNA failed to establish a functional

(C) Comparative overview of RNA structures in LaTranC and ISDra2 TnpB. Structural elements are annotated and colored as in (A and B).
(D) Functional divergence of reRNA (gray) into tracrRNA (blue) and crRNA (red). In TnpB systems, stem 1-PK and stem 4 form intramolecular interactions (gray dashed boxes), and in LaTranC, these regions engage intermolecularly (green dashed boxes), consistent with tracrRNA-crRNA pairing.
(E) Structural alignment of guide RNA components from ISDra2 TnpB, LaTranC, and representative Cas12 systems based on cryo-EM data.
(F–H) Cryo-EM reconstructions of LaTranC ternary complexes. (F) Overview of two conformations: partial R-loop (G) and full R-loop (H) structures. In (H), the unsharpened cryo-EM density of the non-target DNA strand (NTS) is shown in blue.
(I) Structural model illustrating the functional split of reRNA.
(J) Summary of reRNA splitting across three representative TranC clades (3, 8-Cas12n, and 11). See Figure S7 for covariance-based structure comparisons of reRNAs and tracrRNA-crRNA hybrids in clade8-Cas12n and clade11.
See also Figures S6 and S7 and Data S2.

Please cite this article in press as: Jin et al., Functional RNA splitting drove the evolutionary emergence of type V CRISPR-Cas systems from transposons, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.004

CellPress

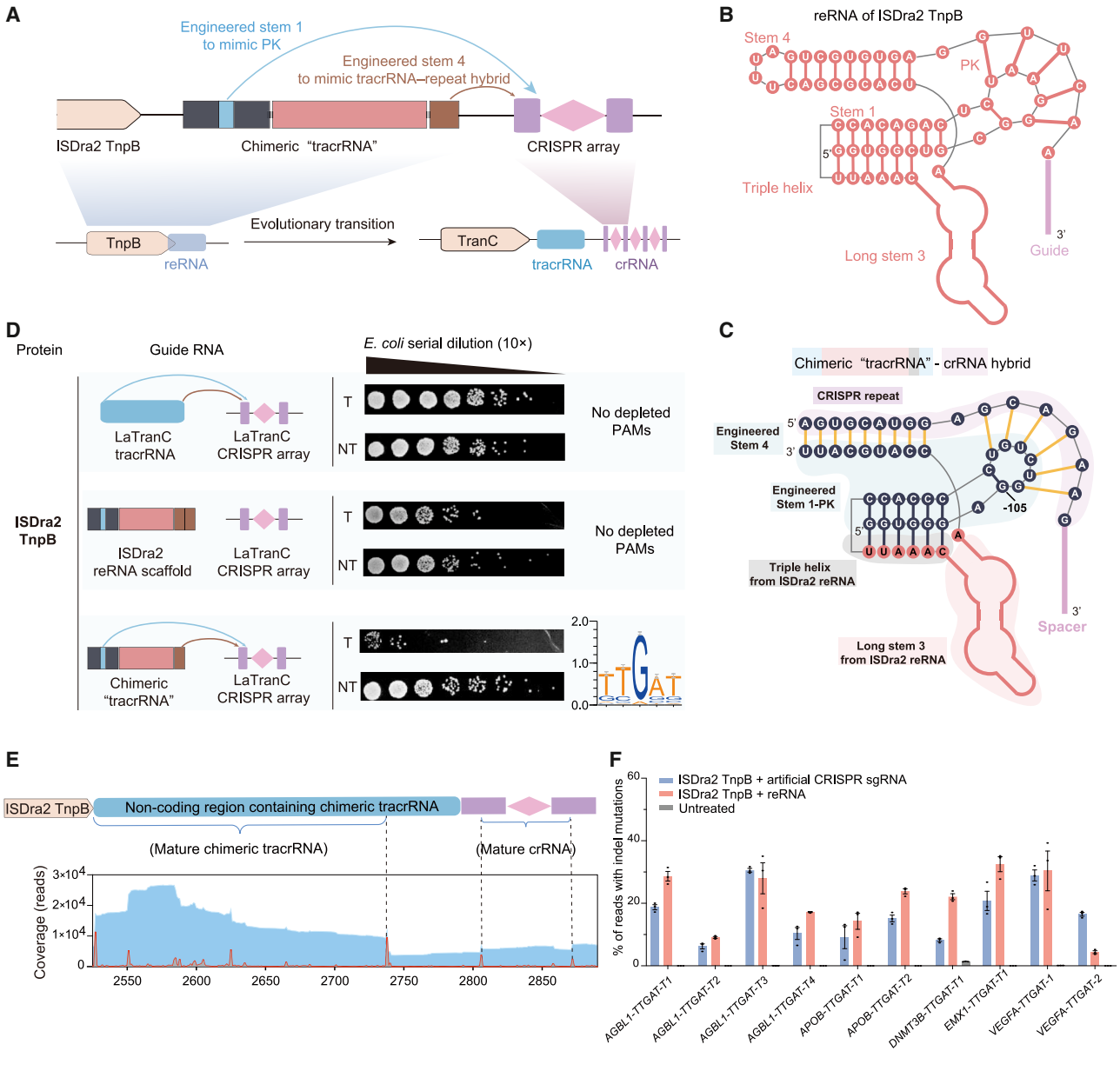


Figure 6. Engineering reRNA splitting to reconstitute a nascent CRISPR system from TnpB components

(A) Conceptual model illustrating experimental mimicry of the evolutionary transition from TnpB-based transposons to CRISPR-Cas systems.

(B and C) Structure comparison between the wild-type ISDra2 reRNA (B) and a chimeric "tracrRNA"-crRNA hybrid constructed by modifying ISDra2 reRNA (C). The stem 1-pseudoknot (PK) and stem 4 were highlighted as functional splitting sites based on cryo-EM comparisons (see Figure 5I). Therefore, stem 1 was replaced with the corresponding LaTranC tracrRNA stem to pair with the 3′ CRISPR repeat, resulting in the PK region. Stem 4 was redesigned to pair with the 5′ repeat. Consequently, the chimeric tracrRNA forms intermolecular interactions with the LaTranC repeats in two regions. The long stem 3 is conserved and not shown in detail.

(D) Functional validation of TnpB using a chimeric tracrRNA-crRNA hybrid in *E. coli* by interference assay (left) and PAM depletion assay (right). Wild-type ISDra2 reRNA and LaTranC tracrRNA served as negative controls.

(E) RNA-seq analysis of guide RNA processing, presented in the same format as Figure 3G.

(F) Genome-editing activity of ISDra2 TnpB in HEK293T cells guided by either wild-type reRNA or an artificial sgRNA (chimeric tracrRNA fused to LaTranC crRNA). Editing efficiencies across ten endogenous target sites containing a 5′-TTGAT PAM were measured by deep sequencing (n = 3, mean ± SEM).

See also Data S1.

Please cite this article in press as: Jin et al., Functional RNA splitting drove the evolutionary emergence of type V CRISPR-Cas systems from transposons, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.004

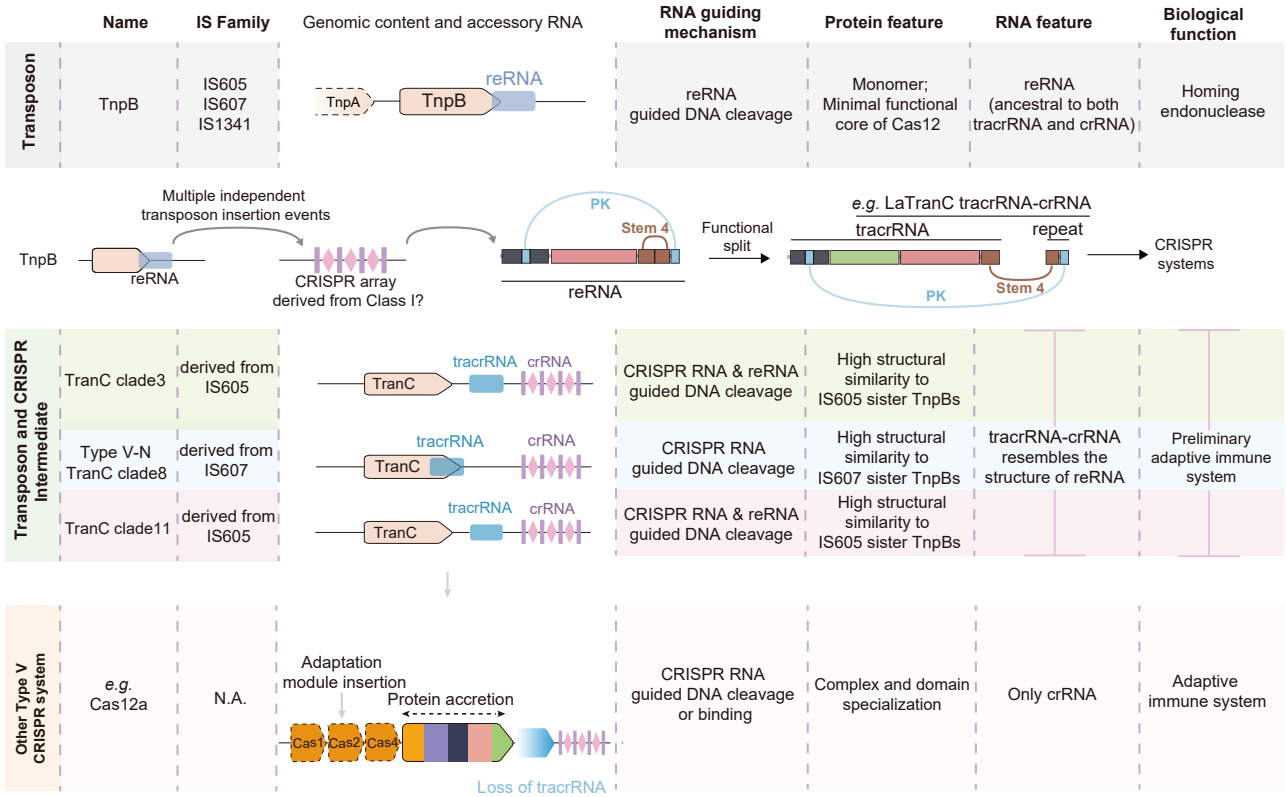


Figure 7. Comparative architecture of TnpB, TranC, and type V CRISPR-Cas systems

Top: TnpB nucleases function as RNA-guided effectors using a single structured guide RNA (reRNA). Middle: TranC systems, proposed here as evolutionary intermediates, display hybrid features of both TnpB and CRISPR systems. They retain a TnpB-like effector but feature modularized guide RNAs (tracrRNA and crRNA) and are found adjacent to co-opted CRISPR arrays.

Bottom: Canonical type V CRISPR-Cas systems (e.g., type V-A) encode large effector proteins and possess complete CRISPR loci, including Cas1, Cas2, and Cas4, supporting both adaptation and interference functions.

association between ISDra2 TnpB and the LaTranC CRISPR array (Figure 6D). RNA sequencing revealed maturation of the chimeric tracrRNA and CRISPR array into two transcripts (Figure 6E).

We next evaluated whether the chimeric tracrRNA-crRNA-guided ISDra2 TnpB would exhibit DNA targeting activity in human cells. After fusing the chimeric tracrRNA and LaTranC crRNA into an artificial sgRNA (Table S2), we compared the editing efficiencies of ISDra2 TnpB guided by a wild-type reRNA or an artificial CRISPR sgRNA across ten target sites in HEK293T human cells (*AGBL1-TTGAT-T1*, *AGBL1-TTGAT-T2*, *AGBL1-TTGAT-T3*, *AGBL1-TTGAT-T4*, *APOB-TTGAT-T1*, *APOB-TTGAT-T2*, *DNMT3B-TTGAT-T1*, *EMX1-TTGAT-T1*, *VEGFA-TTGAT-T1*, and *VEGFA-TTGAT-T2*) with 5′-TTGAT PAMs (Figure 6F). Deep sequencing results showed that these two guide RNAs induced similar levels of editing efficiency (artificial sgRNA: 6.2% to 30.6%, average 16.4%; reRNA: 4.4% to 32.5%, average 21.0%; Figure 6F). Notably, for four genome-editing systems in human cells, eLaTranC-sgRNA, eLaTranC-reRNA, TnpB-reRNA, and TnpB-artificial sgRNA, the indel mutation types were largely similar, mainly inducing deletions (< 50 bp) with rare insertion events (Data S1.15), while the LaTranC-sgRNA group induced slightly larger deletions. Overall, the generation of an artificial CRISPR association demonstrates that the functional splitting of reRNA into tracrRNA and crRNA is sufficient to drive the emergence of TranC systems.

DISCUSSION

Transposon domestication represents one of nature's major evolutionary innovations. This process promotes the birth of new genes, including the creation of new adaptive immune systems.[48–51] Here, we have shown how IS200/605-family TEs gave rise to diverse type V CRISPR-Cas12 systems via multiple, independent domestication events (Figure 7).

First, TranC systems, consistent with their nascent evolutionary status, display evolutionary instability and flexible RNA usage. This instability is evidenced by the frequent absence of CRISPR arrays and substitutions or loss of catalytic residues (Figures 2C–2E; Data S1.3–S1.5). This dynamic is consistent with the patterns observed in evolutionarily young genes, where some lineages retain young genes while others lose them.[52,53] As recently diverged derivatives of TnpB systems, five IS605-derived TranCs (from clade3 and clade11) can be activated by either a crRNA or a reRNA (Figures 3D and 3F). By contrast, the IS607-derived clade8-Cas12n relies solely on crRNA. While

Please cite this article in press as: Jin et al., Functional RNA splitting drove the evolutionary emergence of type V CRISPR-Cas systems from transposons, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.004

this may partly reflect a limited sample size or detection sensitivity, it may also result from clade-specific genomic organization where only clade8-Cas12n and its sister TnpBs exhibit extensive coding sequence overlap with tracrRNA/reRNA regions (Figures 2F and S2D). This co-localization implies a tight coupling between protein and guide RNA, potentially promoting functional associations with the tracrRNA while diminishing compatibility with reRNA.

Second, the functional splitting of reRNA into tracrRNA and crRNA seems to be sufficient for the transition from TnpB to CRISPR-Cas12 systems. In clade3, the LaTranC cryo-EM structure reveals striking conservation with ISDra2 TnpB (Figure 4F), but with a distinct guide RNA configuration: stem 1-PK and stem 4 shift from intramolecular interactions in TnpB-reRNA to intermolecular pairing in the tracrRNA-crRNA hybrid (Figure 5D). Additionally, computational analyses indicate that clade8 and clade11 retain high protein structural similarity to their specific sister TnpBs (Figures 4A–4C), with their tracrRNA-crRNA hybrids resembling the reRNAs of homologous TnpBs (Figure S7). Our engineering experiment further demonstrates that introducing a chimeric tracrRNA is sufficient to convert TnpB-reRNA into a functional TranC-like system (Figure 6). Importantly, in this study, "splitting" refers to functional and structural modularization into distinct tracrRNA and crRNA components, rather than a direct mutational transformation of reRNA sequences. This aligns with two proposed hypotheses for the origins of type V CRISPR arrays: *de novo* generation and co-option of ancestral Class 1 CRISPR arrays.[8,18,54] The CRISPR arrays associated with the experimentally active TranCs may have originated from ancestral Class 1 CRISPR arrays, as those from all eight TranCs are similar to known Class 1 CRISPR arrays (Table S1). In addition, the widespread presence of "isolated arrays" derived from Class 1 systems is consistent with the plausibility of such co-option events.[55] This scenario is also compatible with the Cas12 origination model, wherein a TnpB-encoding transposon integrates near a preexisting CRISPR array, subsequently establishing a functional association between the TnpB protein and the array.[5,18,19,26]

The engineered eLaTranC variant demonstrates endogenous gene editing activity and guide RNA interchangeability in human cells (Figure 3I). However, as a newly emerged and compact Cas12-like protein, its editing efficiency could be further improved. To this end, we developed a 90-nt mini-sgRNA, which is 53.8% shorter than the 195-nt wild-type sgRNA yet retains a similar level of editing activity (Figures S6E and S6F). This mini guide RNA may provide a platform for engineering more compact and efficient LaTranC-based gene editors. Furthermore, large language models (LLMs) have already been applied to the design of artificial genome-editing systems.[56] The guide RNA interchangeability between crRNA and reRNA provides a new perspective on protein-RNA interaction design, which could be leveraged to fine-tune AI (artificial intelligence) models for better modeling of complex protein-RNA dynamics.

Han et al. recently identified a diverse collection of TnpBs and candidate Cas12s (named CRISPR type V-Us) scattered across distinct TnpB branches.[21] To facilitate comparisons, we classified our TranC clades within the V-U nomenclature (Figures 2C–2E; Data S1.3–S1.5; Table S1), with clades 3, 11, 12, 13, and 14 corresponding to types V-U23, V-U18, V-U12, V-U34, and V-U37, respectively.

Altogether, our findings demonstrate that TnpB-encoded TEs evolved into diverse CRISPR-Cas12 systems through RNA-level rewiring. The high genomic abundance and diversity of TnpB loci, together with the relatively rapid evolutionary dynamics of RNA, created favorable conditions for multiple independent transitions toward CRISPR-Cas12 systems. This mode of recurrent emergence contrasts sharply with the apparent single-origin histories of Cas9, which is derived from IscB-encoded IS200/605 elements,[8] and Cas13, which originates from AbiF toxin-antitoxin systems.[57] These diverse evolutionary trajectories of RNA-guided nucleases highlight their unique evolutionary accessibility and promise continued discovery of new biological mechanisms and genome-editing platforms.

Limitations of the study

In this study, computational constraints made it infeasible to simultaneously apply all three mining strategies (motif-, profile-, and sequence-based searches) directly to massive datasets such as MGnify. Therefore, we initiated TranC mining using the combined approach on four smaller, high-quality, and diverse microbiome datasets. We anticipate that future evolutionary analyses leveraging increased computational resources and improved search methodologies will further expand the discovery and characterization of additional TranC systems.

Due to time constraints, we did not explore the detailed characteristics of other TranCs. Moreover, the editing activity of eLaTranC could be further optimized, and its performance in other species, as well as its off-target activity, warrants more comprehensive evaluation.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact: Caixia Gao (cxgao@genetics.ac.cn).

Materials availability

All unique/stable reagents generated in this study are available from the lead contact with a completed materials transfer agreement.

Data and code availability

- All deep sequencing data have been deposited in the National Center for Biotechnology Information BioProject under accession NCBI: PRJNA1017652.
- The electron density maps generated have been deposited in the Electron Microscopy Data Bank (EMDB) under accession numbers EMD-61434, EMD-61435, and EMD-61436 and will be publicly accessible as of the date of publication. Atomic coordinates and structure factors have been deposited in the Protein Data Bank under accession numbers PDB: 9JFO, PDB: 9JFP, and PDB: 9JFQ and will be publicly available upon publication. For those interested, the raw cryo-EM micrographs and movies employed in this study will be shared by the lead contact upon request.
- This paper does not report original code.

- All additional data supporting the findings of this study are available from the lead contact upon request.

ACKNOWLEDGMENTS

We are grateful to Y. Ai for providing HEK293T cells, Q. Ji for providing plasmids encoding Cas12n sequences, and B. Wu for providing plasmid p15A-pTac-dfncpf1. This work was supported by the Agriculture Science and Technology Major Project, the National Key Research and Development Program (2022YFF1002802), the National Natural Science Foundation of China (32230088, 32250012, and 32201224), the CAS Projects for Young Scientists in Basic Research (YSBR-080), and the New Cornerstone Science Foundation.

AUTHOR CONTRIBUTIONS

S.J., Y.E.Z., J.-J.G.L., and C.G. conceived the project and designed the experiments. Yunjia Li, Yuanqing Li, Y. Luo, Q.G., and S.J. performed the evolution and computational biology analyses. S.J., Z.Z., Y. Liu, Z.C., K.T.Z., G.Y., and H.L. performed the *in vivo* experiments. D.L. prepared the LaTranC complexes. S.Z. performed the cryo-EM analysis and built the atomic structures. S.J., Yunjia Li, Z.Z., S.Z., Yuanqing Li, and R.L. prepared the figures. C.G., Y.E.Z., J.-J.G.L., J.-L.Q., S.J., K.T.Z., and R.Z. wrote the manuscript with input from all authors.

DECLARATION OF INTERESTS

The authors have submitted patent applications based on the results reported in this paper. K.T.Z. is a founder and employee at Qi BioDesign. C.G. is a member of the *Cell* advisory board.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - *E. coli* transformation
  - Mammalian cell lines and culture conditions
- METHOD DETAILS
  - General experimental methods
  - Initial protein collection and curation
  - Identification of TranC candidates
  - Evolutionary analysis
  - Functional studies of TranCs based on metadata
  - Computational Structural Analyses
  - *E. coli* functional assays
  - Cryo-EM experiments and data analyses
  - Functional assays in human HEK293T cells
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2025.09.004.

Received: February 1, 2024
Revised: April 30, 2025
Accepted: September 4, 2025

REFERENCES

1. Frost, L.S., Leplae, R., Summers, A.O., and Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. Nat. Rev. Microbiol. *3*, 722–732. https://doi.org/10.1038/nrmicro1235.
2. Benler, S., and Koonin, E.V. (2022). Recruitment of mobile genetic elements for diverse cellular functions in prokaryotes. Front. Mol. Biosci. *9*, 821197. https://doi.org/10.3389/fmolb.2022.821197.
3. Koonin, E.V., Makarova, K.S., Wolf, Y.I., and Krupovic, M. (2020). Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. Nat. Rev. Genet. *21*, 119–131. https://doi.org/10.1038/s41576-019-0172-9.
4. Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science *337*, 816–821. https://doi.org/10.1126/science.1225829.
5. Koonin, E.V., and Makarova, K.S. (2022). Evolutionary plasticity and functional versatility of CRISPR systems. PLOS Biol. *20*, e3001481. https://doi.org/10.1371/journal.pbio.3001481.
6. Koonin, E.V., Gootenberg, J.S., and Abudayyeh, O.O. (2023). Discovery of diverse CRISPR-Cas systems and expansion of the genome engineering toolbox. Biochemistry *62*, 3465–3487. https://doi.org/10.1021/acs.biochem.3c00159.
7. Wu, W.Y., Adiego-Pérez, B., and van der Oost, J. (2024). Biology and applications of CRISPR-Cas12 and transposon-associated homologs. Nat. Biotechnol. *42*, 1807–1821. https://doi.org/10.1038/s41587-024-02485-9.
8. Altae-Tran, H., Kannan, S., Demircioglu, F.E., Oshiro, R., Nety, S.P., McKay, L.J., Dlakić, M., Inskeep, W.P., Makarova, K.S., Macrae, R.K., et al. (2021). The widespread IS200/IS605 transposon family encodes diverse programmable RNA-guided endonucleases. Science *374*, 57–65. https://doi.org/10.1126/science.abj6856.
9. Karvelis, T., Druteika, G., Bigelyte, G., Budre, K., Zedaveinyte, R., Silanskas, A., Kazlauskas, D., Venclovas, Č., and Siksnys, V. (2021). Transposon-associated TnpB is a programmable RNA-guided DNA endonuclease. Nature *599*, 692–696. https://doi.org/10.1038/s41586-021-04058-1.
10. Cong, L., Ran, F.A., Cox, D., Lin, S.L., Barretto, R., Habib, N., Hsu, P.D., Wu, X.B., Jiang, W.Y., Marraffini, L.A., et al. (2013). Multiplex genome engineering using CRISPR/Cas systems. Science *339*, 819–823. https://doi.org/10.1126/science.1231143.
11. Liu, G., Lin, Q., Jin, S., and Gao, C. (2022). The CRISPR-Cas toolbox and gene editing technologies. Mol. Cell *82*, 333–347. https://doi.org/10.1016/j.molcel.2021.12.002.
12. Gao, C. (2021). Genome engineering for crop improvement and future agriculture. Cell *184*, 1621–1635. https://doi.org/10.1016/j.cell.2021.01.005.
13. Wang, J.Y., and Doudna, J.A. (2023). CRISPR technology: A decade of genome editing is only the beginning. Science *379*, eadd8643. https://doi.org/10.1126/science.add8643.
14. Kapitonov, V.V., Makarova, K.S., and Koonin, E.V. (2015). ISC, a novel group of bacterial and archaeal DNA transposons that encode Cas9 homologs. J. Bacteriol. *198*, 797–807. https://doi.org/10.1128/JB.00783-15.
15. Jia, N., and Patel, D.J. (2022). Structure-based evolutionary relationship between IscB and Cas9. Cell Res. *32*, 875–877. https://doi.org/10.1038/s41422-022-00709-8.
16. Kato, K., Okazaki, S., Kannan, S., Altae-Tran, H., Esra Demircioglu, F., Isayama, Y., Ishikawa, J., Fukuda, M., Macrae, R.K., Nishizawa, T., et al. (2022). Structure of the IscB–ωRNA ribonucleoprotein complex, the likely ancestor of CRISPR-Cas9. Nat. Commun. *13*, 6719. https://doi.org/10.1038/s41467-022-34378-3.
17. Schuler, G., Hu, C., and Ke, A. (2022). Structural basis for RNA-guided DNA cleavage by IscB-omegaRNA and mechanistic comparison with Cas9. Science *376*, 1476–1481. https://doi.org/10.1126/science.abq7220.
18. Koonin, E.V., and Makarova, K.S. (2019). Origins and evolution of CRISPR-Cas systems. Philos. Trans. R. Soc. Lond. B Biol. Sci. *374*, 20180087. https://doi.org/10.1098/rstb.2018.0087.

19. Shmakov, S., Smargon, A., Scott, D., Cox, D., Pyzocha, N., Yan, W., Abudayyeh, O.O., Gootenberg, J.S., Makarova, K.S., Wolf, Y.I., et al. (2017). Diversity and evolution of class 2 CRISPR-Cas systems. Nat. Rev. Microbiol. *15*, 169–182. https://doi.org/10.1038/nrmicro.2016.184.

20. Xiang, G., Li, Y., Sun, J., Huo, Y., Cao, S., Cao, Y., Guo, Y., Yang, L., Cai, Y., Zhang, Y.E., et al. (2024). Evolutionary mining and functional characterization of TnpB nucleases identify efficient miniature genome editors. Nat. Biotechnol. *42*, 745–757. https://doi.org/10.1038/s41587-023-01857-x.

21. Altae-Tran, H., Shmakov, S.A., Makarova, K.S., Wolf, Y.I., Kannan, S., Zhang, F., and Koonin, E.V. (2023). Diversity, evolution, and classification of the RNA-guided nucleases TnpB and Cas12. Proc. Natl. Acad. Sci. USA *120*, e2308224120. https://doi.org/10.1073/pnas.2308224120.

22. He, S.S., Guynet, C., Siguier, P., Hickman, A.B., Dyda, F., Chandler, M., and Ton-Hoang, B. (2013). IS200/IS605 family single-strand transposition: mechanism of IS608 strand transfer. Nucleic Acids Res. *41*, 3302–3313. https://doi.org/10.1093/nar/gkt014.

23. Boocock, M.R., and Rice, P.A. (2013). A proposed mechanism for IS607-family serine transposases. Mobile DNA *4*, 24. https://doi.org/10.1186/1759-8753-4-24.

24. Žedaveinytė, R., Meers, C., Le, H.C., Mortman, E.E., Tang, S., Lampe, G. D., Pesari, S.R., Gelsinger, D.R., Wiegand, T., and Sternberg, S.H. (2024). Antagonistic conflict between transposon-encoded introns and guide RNAs. Science *385*, eadm8189. https://doi.org/10.1126/science.adm8189.

25. Meers, C., Le, H.C., Pesari, S.R., Hoffmann, F.T., Walker, M.W.G., Gezelle, J., Tang, S.P., and Sternberg, S.H. (2023). Transposon-encoded nucleases use guide RNAs to promote their selfish spread. Nature *622*, 863–871. https://doi.org/10.1038/s41586-023-06597-1.

26. Makarova, K.S., Wolf, Y.I., Iranzo, J., Shmakov, S.A., Alkhnbashi, O.S., Brouns, S.J.J., Charpentier, E., Cheng, D., Haft, D.H., Horvath, P., et al. (2020). Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. Nat. Rev. Microbiol. *18*, 67–83. https://doi.org/10.1038/s41579-019-0299-x.

27. Yamano, T., Nishimasu, H., Zetsche, B., Hirano, H., Slaymaker, I.M., Li, Y., Fedorova, I., Nakane, T., Makarova, K.S., Koonin, E.V., et al. (2016). Crystal structure of Cpf1 in complex with guide RNA and target DNA. Cell *165*, 949–962. https://doi.org/10.1016/j.cell.2016.04.003.

28. Xiao, R., Li, Z., Wang, S., Han, R., and Chang, L. (2021). Structural basis for substrate recognition and cleavage by the dimerization-dependent CRISPR-Cas12f nuclease. Nucleic Acids Res. *49*, 4120–4128. https://doi.org/10.1093/nar/gkab179.

29. Wang, J.Y., Pausch, P., and Doudna, J.A. (2022). Structural biology of CRISPR-Cas immunity and genome editing enzymes. Nat. Rev. Microbiol. *20*, 641–656. https://doi.org/10.1038/s41579-022-00739-4.

30. Nakagawa, R., Hirano, H., Omura, S.N., Nety, S., Kannan, S., Altae-Tran, H., Yao, X., Sakaguchi, Y., Ohira, T., Wu, W.Y., et al. (2023). Cryo-EM structure of the transposon-associated TnpB enzyme. Nature *616*, 390–397. https://doi.org/10.1038/s41586-023-05933-9.

31. Sasnauskas, G., Tamulaitiene, G., Druteika, G., Carabias, A., Silanskas, A., Kazlauskas, D., Venclovas, Č., Montoya, G., Karvelis, T., and Siksnys, V. (2023). TnpB structure reveals minimal functional core of Cas12 nuclease family. Nature *616*, 384–389. https://doi.org/10.1038/s41586-023-05826-x.

32. Yan, W.X., Hunnewell, P., Alfonse, L.E., Carte, J.M., Keston-Smith, E., Sothiselvam, S., Garrity, A.J., Chong, S.R., Makarova, K.S., Koonin, E. V., et al. (2019). Functionally diverse type V CRISPR-Cas systems. Science *363*, 88–91. https://doi.org/10.1126/science.aav7271.

33. Harrington, L.B., Burstein, D., Chen, J.S., Paez-Espino, D., Ma, E., Witte, I.P., Cofsky, J.C., Kyrpides, N.C., Banfield, J.F., and Doudna, J.A. (2018). Programmed DNA destruction by miniature CRISPR-Cas14 enzymes. Science *362*, 839–842. https://doi.org/10.1126/science.aav4294.

34. Strecker, J., Ladha, A., Gardner, Z., Schmid-Burgk, J.L., Makarova, K.S., Koonin, E.V., and Zhang, F. (2019). RNA-guided DNA insertion with CRISPR-associated transposases. Science *365*, 48–53. https://doi.org/10.1126/science.aax9181.

35. Wu, W.Y., Mohanraju, P., Liao, C.Y., Adiego-Pérez, B., Creutzburg, S.C. A., Makarova, K.S., Keessen, K., Lindeboom, T.A., Khan, T.S., Prinsen, S., et al. (2022). The miniature CRISPR-Cas12m effector binds DNA to block transcription. Mol. Cell *82*, 4487–4502.e7. https://doi.org/10.1016/j.molcel.2022.11.003.

36. Omura, S.N., Nakagawa, R., Südfeld, C., Villegas Warren, R., Wu, W.Y., Hirano, H., Laffeber, C., Kusakizako, T., Kise, Y., Lebbink, J.H.G., et al. (2023). Mechanistic and evolutionary insights into a type V-M CRISPR–Cas effector enzyme. Nat. Struct. Mol. Biol. *30*, 1172–1182. https://doi.org/10.1038/s41594-023-01042-3.

37. Takeda, S.N., Nakagawa, R., Okazaki, S., Hirano, H., Kobayashi, K., Kusakizako, T., Nishizawa, T., Yamashita, K., Nishimasu, H., and Nureki, O. (2021). Structure of the miniature type V-F CRISPR-Cas effector enzyme. Mol. Cell *81*, 558–570.e3. https://doi.org/10.1016/j.molcel.2020.11.035.

38. van der Oost, J., Jore, M.M., Westra, E.R., Lundgren, M., and Brouns, S. J.J. (2009). CRISPR-based adaptive and heritable immunity in prokaryotes. Trends Biochem. Sci. *34*, 401–407. https://doi.org/10.1016/j.tibs.2009.05.002.

39. van Houte, S., Ekroth, A.K.E., Broniewski, J.M., Chabas, H., Ashby, B., Bondy-Denomy, J., Gandon, S., Boots, M., Paterson, S., Buckling, A., et al. (2016). The diversity-generating benefits of a prokaryotic adaptive immune system. Nature *532*, 385–388. https://doi.org/10.1038/nature17436.

40. Chen, W.Z., Ma, J.C., Wu, Z.W., Wang, Z.P., Zhang, H.Y., Fu, W.H., Pan, D., Shi, J., and Ji, Q.J. (2023). Cas12n nucleases, early evolutionary intermediates of type V CRISPR, comprise a distinct family of miniature genome editors. Mol. Cell *83*, 2768–2780.e6. https://doi.org/10.1016/j.molcel.2023.06.014.

41. Karvelis, T., Bigelyte, G., Young, J.K., Hou, Z., Zedaveinyte, R., Budre, K., Paulraj, S., Djukanovic, V., Gasior, S., Silanskas, A., et al. (2020). PAM recognition by miniature CRISPR-Cas12f nucleases triggers programmable double-stranded DNA target cleavage. Nucleic Acids Res. *48*, 5016–5023. https://doi.org/10.1093/nar/gkaa208.

42. Stella, S., Alcón, P., and Montoya, G. (2017). Structure of the Cpf1 endonuclease R-loop complex after target DNA cleavage. Nature *546*, 559–563. https://doi.org/10.1038/nature22398.

43. Yang, H., Gao, P., Rajashankar, K.R., and Patel, D.J. (2016). PAM-Dependent Target DNA Recognition and Cleavage by C2c1 CRISPR-Cas Endonuclease. Cell *167*, 1814–1828.e12. https://doi.org/10.1016/j.cell.2016.11.053.

44. Liu, J.-J., Orlova, N., Oakes, B.L., Ma, E., Spinner, H.B., Baney, K.L.M., Chuck, J., Tan, D., Knott, G.J., Harrington, L.B., et al. (2019). CasX enzymes comprise a distinct family of RNA-guided genome editors. Nature *566*, 218–223. https://doi.org/10.1038/s41586-019-0908-x.

45. Xu, J., and Zhang, Y. (2010). How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics *26*, 889–895. https://doi.org/10.1093/bioinformatics/btq066.

46. Zhang, C.X., Shine, M., Pyle, A.M., and Zhang, Y. (2022). US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. Nat. Methods *19*, 1109–1115. https://doi.org/10.1038/s41592-022-01585-1.

47. Su, M.J., Li, F., Wang, Y.J., Gao, Y., Lan, W.Q., Shao, Z.W., Zhu, C., Tang, N., Gan, J.H., Wu, Z.W., et al. (2024). Molecular basis and engineering of miniature Cas12f with C-rich PAM specificity. Nat. Chem. Biol. *20*, 180–189. https://doi.org/10.1038/s41589-023-01420-4.

48. Joly-Lopez, Z., and Bureau, T.E. (2018). Exaptation of transposable element coding sequences. Curr. Opin. Genet. Dev. *49*, 34–42. https://doi.org/10.1016/j.gde.2018.02.011.

49. Schrader, L., and Schmitz, J. (2019). The impact of transposable elements in adaptive evolution. Mol. Ecol. *28*, 1537–1549. https://doi.org/10.1111/mec.14794.

50. Feschotte, C., Jiang, N., and Wessler, S.R. (2002). Plant transposable elements: where genetics meets genomics. Nat. Rev. Genet. *3*, 329–341. https://doi.org/10.1038/nrg793.

51. Alzohairy, A.M., Gyulai, G., Jansen, R.K., and Bahieldin, A. (2013). Transposable elements domesticated and neofunctionalized by eukaryotic genomes. Plasmid *69*, 1–15. https://doi.org/10.1016/j.plasmid.2012.08.001.

52. Ross, B.D., Rosin, L., Thomae, A.W., Hiatt, M.A., Vermaak, D., de la Cruz, A.F.A., Imhof, A., Mellone, B.G., and Malik, H.S. (2013). Stepwise evolution of essential centromere function in a Drosophila neogene. Science *340*, 1211–1214. https://doi.org/10.1126/science.1234393.

53. Yang, H.W., He, B.Z., Ma, H.J., Tsaur, S.C., Ma, C.Y., Wu, Y., Ting, C.T., and Zhang, Y.E. (2015). Expression profile and gene age jointly shaped the genome-wide distribution of premature termination codons in a Drosophila melanogaster population. Mol. Biol. Evol. *32*, 216–228. https://doi.org/10.1093/molbev/msu299.

54. Koonin, E.V., and Krupovic, M. (2023). New faces of prokaryotic mobile genetic elements: guide RNAs link transposition with host defense mechanisms. Curr. Opin. Syst. Biol. *36*, 100473. https://doi.org/10.1016/j.coisb.2023.100473.

55. Shmakov, S.A., Utkina, I., Wolf, Y.I., Makarova, K.S., Severinov, K.V., and Koonin, E.V. (2020). CRISPR arrays away from Cas genes. CRISPR J. *3*, 535–549. https://doi.org/10.1089/crispr.2020.0062.

56. Ruffolo, J.A., Nayfach, S., Gallagher, J., Bhatnagar, A., Beazer, J., Hussain, R., Russ, J., Yip, J., Hill, E., Pacesa, M., et al. (2025). Design of highly functional genome editors by modelling CRISPR-Cas sequences. Nature *645*, 518–525. https://doi.org/10.1038/s41586-025-09298-z.

57. Zilberzwige-Tal, S., Altae-Tran, H., Kannan, S., Wilkinson, M.E., Vo, S.C. D.T., Strebinger, D., Edmonds, K.K., Yao, C.J., Mears, K.S., Shmakov, S. A., et al. (2025). Reprogrammable RNA-targeting CRISPR systems evolved from RNA toxin-antitoxins. Cell *188*, 1925–1940.e20. https://doi.org/10.1016/j.cell.2025.01.034.

58. Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D. M., Meng, E.C., and Ferrin, T.E. (2004). UCSF chimera - A visualization system for exploratory research and analysis. J. Comput. Chem. *25*, 1605–1612. https://doi.org/10.1002/jcc.20084.

59. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

60. Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A.J., Bambrick, J., et al. (2024). Addendum: Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature *636*, E4. https://doi.org/10.1038/s41586-024-08416-7.

61. Bailey, T.L. (2021). STREME: accurate and versatile sequence motif discovery. Bioinformatics *37*, 2834–2840. https://doi.org/10.1093/bioinformatics/btab203.

62. Steinegger, M., and Söding, J. (2018). Clustering huge protein sequence sets in linear time. Nat. Commun. *9*, 2542. https://doi.org/10.1038/s41467-018-04964-5.

63. Eddy, S.R. (2011). Accelerated profile HMM searches. PLOS Comput. Biol. *7*, e1002195. https://doi.org/10.1371/journal.pcbi.1002195.

64. van Kempen, M., Kim, S.S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C.L.M., Söding, J., and Steinegger, M. (2024). Fast and accurate protein structure search with Foldseek. Nat. Biotechnol. *42*, 243–246. https://doi.org/10.1038/s41587-023-01773-0.

65. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. *37*, 1530–1534. https://doi.org/10.1093/molbev/msaa015.

66. Jin, S., Lin, Q.P., Gao, Q., and Gao, C.X. (2023). Optimized prime editing in monocot plants using PlantPegDesigner and engineered plant prime editors (ePPEs). Nat. Protoc. *18*, 831–853. https://doi.org/10.1038/s41596-022-00773-9.

67. Shmakov, S., Abudayyeh, O.O., Makarova, K.S., Wolf, Y.I., Gootenberg, J.S., Semenova, E., Minakhin, L., Joung, J., Konermann, S., Severinov, K., et al. (2015). Discovery and functional characterization of diverse Class 2 CRISPR-Cas systems. Mol. Cell *60*, 385–397. https://doi.org/10.1016/j.molcel.2015.10.008.

68. Wu, Z.W., Zhang, Y.F., Yu, H.P., Pan, D., Wang, Y.J., Wang, Y.N., Li, F., Liu, C., Nan, H., Chen, W.Z., et al. (2021). Programmed genome editing by a miniature CRISPR-Cas12f nuclease. Nat. Chem. Biol. *17*, 1132–1138. https://doi.org/10.1038/s41589-021-00868-6.

69. Urbaitis, T., Gasiunas, G., Young, J.K., Hou, Z.L., Paulraj, S., Godliauskaite, E., Juskeviciene, M.M., Stitilyte, M., Jasnauskaite, M., Mabuchi, M., et al. (2022). A new family of CRISPR-type V nucleases with C-rich PAM recognition. EMBO Rep. *23*, e55481. https://doi.org/10.15252/embr.202255481.

70. Pausch, P., Al-Shayeb, B., Bisom-Rapp, E., Tsuchida, C.A., Li, Z., Cress, B.F., Knott, G.J., Jacobsen, S.E., Banfield, J.F., and Doudna, J.A. (2020). CRISPR-Cas Phi from huge phages is a hypercompact genome editor. Science *369*, 333–337. https://doi.org/10.1126/science.abb1400.

71. Sun, A., Li, C.-P., Chen, Z., Zhang, S., Li, D.-Y., Yang, Y., Li, L.-Q., Zhao, Y., Wang, K., Li, Z., et al. (2023). The compact Casπ (Cas12l) 'bracelet' provides a unique structural platform for DNA manipulation. Cell Res. *33*, 229–244. https://doi.org/10.1038/s41422-022-00771-2.

72. Bigelyte, G., Young, J.K., Karvelis, T., Budre, K., Zedaveinyte, R., Djukanovic, V., Van Ginkel, E., Paulraj, S., Gasior, S., Jones, S., et al. (2021). Miniature type V-F CRISPR-Cas nucleases enable targeted DNA modification in cells. Nat. Commun. *12*, 6191. https://doi.org/10.1038/s41467-021-26469-4.

73. Burstein, D., Harrington, L.B., Strutt, S.C., Probst, A.J., Anantharaman, K., Thomas, B.C., Doudna, J.A., and Banfield, J.F. (2017). New CRISPR–Cas systems from uncultivated microbes. Nature *542*, 237–241. https://doi.org/10.1038/nature21059.

74. Harrington, L.B., Ma, E.B., Chen, J.S., Witte, I.P., Gertz, D., Paez-Espino, D., Al-Shayeb, B., Kyrpides, N.C., Burstein, D., Banfield, J.F., et al. (2020). A scoutRNA is required for some Type V CRISPR-Cas systems. Mol. Cell *79*, 416–424.e5. https://doi.org/10.1016/j.molcel.2020.06.022.

75. Strecker, J., Jones, S., Koopal, B., Schmid-Burgk, J., Zetsche, B., Gao, L.Y., Makarova, K.S., Koonin, E.V., and Zhang, F. (2019). Engineering of CRISPR-Cas12b for human genome editing. Nat. Commun. *10*, 212. https://doi.org/10.1038/s41467-018-08224-4.

76. Zetsche, B., Gootenberg, J.S., Abudayyeh, O.O., Slaymaker, I.M., Makarova, K.S., Essletzbichler, P., Volz, S.E., Joung, J., van der Oost, J., Regev, A., et al. (2015). Cpf1 Is a single RNA-guided endonuclease of a Class 2 CRISPR-Cas system. Cell *163*, 759–771. https://doi.org/10.1016/j.cell.2015.09.038.

77. Huang, Y., Niu, B.F., Gao, Y., Fu, L.M., and Li, W.Z. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics *26*, 680–682. https://doi.org/10.1093/bioinformatics/btq003.

78. Richardson, L., Allen, B., Baldi, G., Beracochea, M., Bileschi, M.L., Burdett, T., Burgin, J., Caballero-Pérez, J., Cochrane, G., Colwell, L.J., et al. (2023). MGnify: the microbiome sequence data analysis resource in 2023. Nucleic Acids Res. *51*, D753–D759. https://doi.org/10.1093/nar/gkac1080.

79. Parks, D.H., Chuvochina, M., Rinke, C., Mussig, A.J., Chaumeil, P.-A., and Hugenholtz, P. (2022). GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res. *50*, D785–D794. https://doi.org/10.1093/nar/gkab776.

80. Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z. J., Pollard, K.S., Sakharova, E., Parks, D.H., Hugenholtz, P., et al. (2021). A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat. Biotechnol. *39*, 105–114. https://doi.org/10.1038/s41587-020-0603-3.

81. Bai, Y., Müller, D.B., Srinivas, G., Garrido-Oter, R., Potthoff, E., Rott, M., Dombrowski, N., Münch, P.C., Spaepen, S., Remus-Emsermann, M., et al. (2015). Functional overlap of the Arabidopsis leaf and root microbiota. Nature *528*, 364–369. https://doi.org/10.1038/nature16192.

82. Levy, A., Salas Gonzalez, I., Mittelviefhaus, M., Clingenpeel, S., Herrera Paredes, S., Miao, J., Wang, K., Devescovi, G., Stillman, K., Monteiro, F., et al. (2017). Genomic features of bacterial adaptation to plants. Nat. Genet. *50*, 138–150. https://doi.org/10.1038/s41588-017-0012-9.

83. Hyatt, D., Chen, G.L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinf. *11*, 119. https://doi.org/10.1186/1471-2105-11-119.

84. Li, T., and Yin, Y.B. (2022). Critical assessment of pan-genomic analysis of metagenome-assembled genomes. Brief. Bioinform. *23*, bbac413. https://doi.org/10.1093/bib/bbac413.

85. Bailey, T.L., and Gribskov, M. (1998). Combining evidence using p-values: application to sequence homology searches. Bioinformatics *14*, 48–54. https://doi.org/10.1093/bioinformatics/14.1.48.

86. Bai, J., Pennill, L.A., Ning, J., Lee, S.W., Ramalingam, J., Webb, C.A., Zhao, B., Sun, Q., Nelson, J.C., Leach, J.E., et al. (2002). Diversity in nucleotide binding site-leucine-rich repeat genes in cereals. Genome Res. *12*, 1871–1884. https://doi.org/10.1101/gr.454902.

87. Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. *30*, 3059–3066. https://doi.org/10.1093/nar/gkf436.

88. Bland, C., Ramsey, T.L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N.C., and Hugenholtz, P. (2007). CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinf. *8*, 209. https://doi.org/10.1186/1471-2105-8-209.

89. Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Néron, B., Rocha, E.P.C., Vergnaud, G., Gautheret, D., and Pourcel, C. (2018). CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Res. *46*, W246–W251. https://doi.org/10.1093/nar/gky425.

90. Schoch, C.L., Ciufo, S., Domrachev, M., Hotton, C.L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., et al. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database. *2020*, baaa062. https://doi.org/10.1093/database/baaa062.

91. Siguier, P., Perochon, J., Lestrade, L., Mahillon, J., and Chandler, M. (2006). ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. *34*, D32–D36. https://doi.org/10.1093/nar/gkj014.

92. Wang, Y., Guo, M., Yang, N., Guan, Z., Wu, H., Ullah, N., Asare, E., Shi, S., Gao, B., and Song, C. (2023). Phylogenetic relationships among TnpB-containing mobile elements in six bacterial species. Genes *14*, 523. https://doi.org/10.3390/genes14020523.

93. Makarova, K.S., Wolf, Y.I., Alkhnbashi, O.S., Costa, F., Shah, S.A., Saunders, S.J., Barrangou, R., Brouns, S.J.J., Charpentier, E., Haft, D.H., et al. (2015). An updated evolutionary classification of CRISPR-Cas systems. Nat. Rev. Microbiol. *13*, 722–736. https://doi.org/10.1038/nrmicro3569.

94. Edgar, R.C. (2022). Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. Nat. Commun. *13*, 6968. https://doi.org/10.1038/s41467-022-34630-w.

95. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973. https://doi.org/10.1093/bioinformatics/btp348.

96. Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2-Approximately Maximum-Likelihood Trees for Large Alignments. PLOS One *5*, e9490. https://doi.org/10.1371/journal.pone.0009490.

97. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin, L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods *14*, 587–589. https://doi.org/10.1038/Nmeth.4285.

98. Saito, M., Xu, P., Faure, G., Maguire, S., Kannan, S., Altae-Tran, H., Vo, S., Desimone, A., Macrae, R.K., and Zhang, F. (2023). Fanzor is a eukaryotic programmable RNA-guided endonuclease. Nature *620*, 660–668. https://doi.org/10.1038/s41586-023-06356-2.

99. Russel, J., Pinilla-Redondo, R., Mayo-Muñoz, D., Shah, S.A., and Sørensen, S.J. (2020). CRISPRCasTyper: automated identification, annotation, and classification of CRISPR-Cas loci. CRISPR J. *3*, 462–469. https://doi.org/10.1089/crispr.2020.0059.

100. Pourcel, C., Touchon, M., Villeriot, N., Vernadet, J.P., Couvin, D., Toffano-Nioche, C., and Vergnaud, G. (2020). CRISPRCasdb a successor of CRISPRdb containing CRISPR arrays and cas genes from complete genome sequences, and tools to download and query lists of repeats and spacers. Nucleic Acids Res. *48*, D535–D544. https://doi.org/10.1093/nar/gkz915.

101. Gregory, A.C., Zablocki, O., Zayed, A.A., Howell, A., Bolduc, B., and Sullivan, M.B. (2020). The gut virome database reveals age-dependent patterns of virome diversity in the human gut. Cell Host Microbe *28*, 724–740.e8. https://doi.org/10.1016/j.chom.2020.08.003.

102. Unterer, M., Khan Mirzaei, M.K., and Deng, L. (2021). Gut Phage Database: phage mining in the cave of wonders. Signal Transduct. Target. Ther. *6*, 193. https://doi.org/10.1038/s41392-021-00615-2.

103. Arndt, D., Grant, J.R., Marcu, A., Sajed, T., Pon, A., Liang, Y., and Wishart, D.S. (2016). PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res. *44*, W16–W21. https://doi.org/10.1093/nar/gkw387.

104. Camargo, A.P., Nayfach, S., Chen, I.A., Palaniappan, K., Ratner, A., Chu, K., Ritter, S.J., Reddy, T.B.K., Mukherjee, S., Schulz, F., et al. (2023). IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. Nucleic Acids Res. *51*, D733–D743. https://doi.org/10.1093/nar/gkac1037.

105. Wang, R.H., Yang, S., Liu, Z., Zhang, Y., Wang, X., Xu, Z., Wang, J., and Li, S.C. (2024). PhageScope: a well-annotated bacteriophage database with automatic analyses and visualizations. Nucleic Acids Res. *52*, D756–D761. https://doi.org/10.1093/nar/gkad979.

106. Zhang, Y.C., Bhosle, A., Bae, S., McIver, L.J., Pishchany, G., Accorsi, E. K., Thompson, K.N., Arze, C., Wang, Y., Subramanian, A., et al. (2022). Discovery of bioactive microbial gene products in inflammatory bowel disease. Nature *606*, 754–760. https://doi.org/10.1038/s41586-022-04648-7.

107. Tisza, M.J., and Buck, C.B. (2021). A catalog of tens of thousands of viruses from human metagenomes reveals hidden associations with chronic diseases. Proc. Natl. Acad. Sci. USA *118*, e2023202118. https://doi.org/10.1073/pnas.2023202118.

108. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

109. Ross, B.D., Verster, A.J., Radey, M.C., Schmidtke, D.T., Pope, C.E., Hoffman, L.R., Hajjar, A.M., Peterson, S.B., Borenstein, E., and Mougous, J.D. (2019). Human gut bacteria contain acquired interbacterial defence systems. Nature *575*, 224–228. https://doi.org/10.1038/s41586-019-1708-z.

110. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021).

Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.

111. Li, D., Zhang, S., Lin, S., Xing, W., Yang, Y., Zhu, F., Su, D., Chen, C., and Liu, J.G. (2024). Cas12e orthologs evolve variable structural elements to facilitate dsDNA cleavage. Nat. Commun. *15*, 10727. https://doi.org/10.1038/s41467-024-54491-9.
112. Mann, M., Wright, P.R., and Backofen, R. (2017). IntaRNA 2.0: enhanced and customizable prediction of RNA-RNA interactions. Nucleic Acids Res. *45*, W435–W439. https://doi.org/10.1093/nar/gkx279.
113. Tagashira, M., and Asai, K. (2022). ConsAlifold: considering RNA structural alignments improves prediction accuracy of RNA consensus secondary structures. Bioinformatics *38*, 710–719. https://doi.org/10.1093/bioinformatics/btab738.
114. Nawrocki, E.P., and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics *29*, 2933–2935. https://doi.org/10.1093/bioinformatics/btt509.
115. Rivas, E., Clements, J., and Eddy, S.R. (2017). A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. Nat. Methods *14*, 45–48. https://doi.org/10.1038/nmeth.4066.
116. Weinberg, Z., and Breaker, R.R. (2011). R2R–software to speed the depiction of aesthetic consensus RNA secondary structures. BMC Bioinf. *12*, 3. https://doi.org/10.1186/1471-2105-12-3.
117. Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. *35*, 1026–1028. https://doi.org/10.1038/nbt.3988.
118. Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152. https://doi.org/10.1093/bioinformatics/bts565.
119. Yamano, T., Zetsche, B., Ishitani, R., Zhang, F., Nishimasu, H., and Nureki, O. (2017). Structural basis for the canonical and non-canonical PAM recognition by CRISPR-Cpf1. Mol. Cell *67*, 633–645.e3. https://doi.org/10.1016/j.molcel.2017.06.035.
120. Liu, L., Chen, P., Wang, M., Li, X.Y., Wang, J.Y., Yin, M.L., and Wang, Y.L. (2017). C2c1-sgRNA complex structure reveals RNA-guided DNA cleavage mechanism. Mol. Cell *65*, 310–322. https://doi.org/10.1016/j.molcel.2016.11.040.
121. Wu, D., Guan, X., Zhu, Y., Ren, K., and Huang, Z. (2017). Structural basis of stringent PAM recognition by CRISPR-C2c1 in complex with sgRNA. Cell Res. *27*, 705–708. https://doi.org/10.1038/cr.2017.46.
122. Zhang, B., Lin, J.Y., Perčulija, V., Li, Y., Lu, Q.H., Chen, J., and Ouyang, S.Y. (2022). Structural insights into target DNA recognition and cleavage by the CRISPR-Cas12c1 system. Nucleic Acids Res. *50*, 11820–11833. https://doi.org/10.1093/nar/gkac987.
123. Tsuchida, C.A., Zhang, S., Doost, M.S., Zhao, Y., Wang, J., O'Brien, E., Fang, H., Li, C.P., Li, D., Hai, Z.Y., et al. (2022). Chimeric CRISPR-CasX enzymes and guide RNAs for improved genome editing activity. Mol. Cell *82*, 1199–1209.e6. https://doi.org/10.1016/j.molcel.2022.02.002.
124. Li, Z., Zhang, H., Xiao, R.J., Han, R.J., and Chang, L.F. (2021). Cryo-EM structure of the RNA-guided ribonuclease Cas12g. Nat. Chem. Biol. *17*, 387–393. https://doi.org/10.1038/s41589-020-00721-2.
125. Zhang, H., Li, Z., Xiao, R.J., and Chang, L.F. (2020). Mechanisms for target recognition and cleavage by the Cas12i RNA-guided endonuclease. Nat. Struct. Mol. Biol. *27*, 1069–1076. https://doi.org/10.1038/s41594-020-0499-0.
126. Huang, X., Sun, W., Cheng, Z., Chen, M.X., Li, X.Y., Wang, J.Y., Sheng, G., Gong, W.M., and Wang, Y.L. (2020). Structural basis for two metal-ion catalysis of DNA cleavage by Cas12i2. Nat. Commun. *11*, 5241. https://doi.org/10.1038/s41467-020-19072-6.
127. Pausch, P., Soczek, K.M., Herbst, D.A., Tsuchida, C.A., Al-Shayeb, B., Banfield, J.F., Nogales, E., and Doudna, J.A. (2021). DNA interference states of the hypercompact CRISPR-CasPhi effector. Nat. Struct. Mol. Biol. *28*, 652–661. https://doi.org/10.1038/s41594-021-00632-3.
128. Querques, I., Schmitz, M., Oberli, S., Chanez, C., and Jinek, M. (2021). Target site selection and remodelling by type V CRISPR-transposon systems. Nature *599*, 497–502. https://doi.org/10.1038/s41586-021-04030-z.
129. Huang, J.Y., Lin, Q.P., Fei, H.Y., He, Z.X., Xu, H., Li, Y.J., Qu, K.L., Han, P., Gao, Q., Li, B.S., et al. (2023). Discovery of deaminase functions by structure-based protein clustering. Cell *186*, 3182–3195.e14. https://doi.org/10.1016/j.cell.2023.05.041.
130. Pollastri, G., Baldi, P., Fariselli, P., and Casadio, R. (2002). Prediction of coordination number and relative solvent accessibility in proteins. Proteins *47*, 142–153. https://doi.org/10.1002/prot.10069.
131. Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure – pattern-recognition of hydrogen-bonded and geometrical features. Biopolymers *22*, 2577–2637. https://doi.org/10.1002/bip.360221211.
132. Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. *31*, 3406–3415. https://doi.org/10.1093/nar/gkg595.
133. Chen, Z.H., Sun, J.Y., Guan, Y., Li, M., Lou, C.B., and Wu, B. (2021). Engineered DNase-inactive Cpf1 variants to improve targeting scope for base editing in E. coli. Synth. Syst. Biotechnol. *6*, 326–334. https://doi.org/10.1016/j.synbio.2021.09.002.
134. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.
135. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/Nmeth.1923.
136. Zheng, S.Q., Palovcak, E., Armache, J.P., Verba, K.A., Cheng, Y.F., and Agard, D.A. (2017). MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. Nat. Methods *14*, 331–332. https://doi.org/10.1038/nmeth.4193.
137. Punjani, A., Rubinstein, J.L., Fleet, D.J., and Brubaker, M.A. (2017). cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. Nat. Methods *14*, 290–296. https://doi.org/10.1038/Nmeth.4169.
138. Zivanov, J., Nakane, T., Forsberg, B.O., Kimanius, D., Hagen, W.J.H., Lindahl, E., and Scheres, S.H.W. (2018). New tools for automated high-resolution cryo-EM structure determination in RELION-3. eLife *7*, e42166. https://doi.org/10.7554/eLife.42166.
139. Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta Crystallogr. D Biol. Crystallogr. *60*, 2126–2132. https://doi.org/10.1107/S0907444904019158.
140. Liebschner, D., Afonine, P.V., Baker, M.L., Bunkóczi, G., Chen, V.B., Croll, T.I., Hintze, B., Hung, L.W., Jain, S., McCoy, A.J., et al. (2019). Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. Acta Crystallogr. D Struct. Biol. *75*, 861–877. https://doi.org/10.1107/S2059798319011471.
141. Miller, S.M., Wang, T., Randolph, P.B., Arbab, M., Shen, M.W., Huang, T.P., Matuszek, Z., Newby, G.A., Rees, H.A., and Liu, D.R. (2020). Continuous evolution of SpCas9 variants compatible with non-G PAMs. Nat. Biotechnol. *38*, 471–481. https://doi.org/10.1038/s41587-020-0412-8.
142. Anzalone, A.V., Randolph, P.B., Davis, J.R., Sousa, A.A., Koblan, L.W., Levy, J.M., Chen, P.J., Wilson, C., Newby, G.A., Raguram, A., et al. (2019). Search-and-replace genome editing without double-strand breaks or donor DNA. Nature *576*, 149–157. https://doi.org/10.1038/s41586-019-1711-4.
143. Medini, D., Donati, C., Tettelin, H., Masignani, V., and Rappuoli, R. (2005). The microbial pan-genome. Curr. Opin. Genet. Dev. *15*, 589–594. https://doi.org/10.1016/j.gde.2005.09.006.

Please cite this article in press as: Jin et al., Functional RNA splitting drove the evolutionary emergence of type V CRISPR-Cas systems from transposons, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.004

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
FastT1 Competent Cells	Vazyme	Cat# C505-02
DH5α Electro-Cells	Takara	Cat# 9027
Chemicals, peptides, and recombinant proteins		
Yeast extract	OXOID	Cat# LP0021B
Tryptone	OXOID	Cat# LP0042B
Ampicillin Sodium	Macklin	Cat# A6265
Kanamycin Sulfate	Solarbio	Cat# AK177
PureYield™ Plasmid Miniprep System	Promega	Cat# A1222
DMEM (1X)+GlutaMax	Gibco	Cat# 10569044
FBS Fetal Bovine Serum, Qualified	Gibco	Cat# 10091148
TrypLE Express	Gibco	Cat# 12605-010
PBS pH 7.4 basic (1X)	Gibco	Cat# C10010500BT
Opti-MEM	Gibco	Cat# 31985-070
Lipofectamine 2000	Invitrogen	Cat# 11668019
75 cm^2 Cell Culture Flask	NEST	Cat# 708003
48-well Clear TC-treated Plates	Corning	Cat# 3548
Trypan Blue stain 0.4%	Invitrogen	Cat# T10282
Countess cell counting chamber slides	Invitrogen	Cat# C10283
Critical commercial assays		
Mycoplasma Detection Kit	Transgen	Cat# FM311-01
Triumfi Mouse Tissue Direct PCR Kit	Genesand	Cat# SD312
Phanta Max Master Mix	Vazyme	Cat# P525-01
GeneJET Gel Extraction Kit	Thermo Scientific	Cat# K0692
ClonExpress II One Step Cloning Kit	Vazyme	Cat# C112
TIANprep Mini Plasmid Kit	TIANGEN Biotech	Cat# DP103
Deposited data		
Deep amplicon sequencing data	This paper	Accession ID NCBI: PRJNA1017652
Experimental models: Cell lines		
HEK293T	ATCC	Cat# CRL-3216
Oligonucleotides		
Primers used in this paper, see Table S2	This paper	N/A
Recombinant DNA		
pEffector-LaTranC-CRISPR RNA	This paper	Addgene # 246928
pEffector-RuTranC-CRISPR RNA	This paper	N/A
pEffector-Clo1TranC-CRISPR RNA	This paper	N/A
pEffector-MiCas12n-CRISPR RNA	This paper	N/A
pEffector-AnTranC-CRISPR RNA	This paper	N/A
pEffector-Eu1TranC-CRISPR RNA	This paper	N/A
pEffector-ISDra2 TnpB-artificial CRISPR RNA	This paper	Addgene # 246929
pTarget-TTC	This paper	Addgene # 246930
pCMV-LaTranC	This paper	Addgene # 246931
pCMV-eLaTranC	This paper	Addgene # 246932
phU6-LaTranC-sgRNA	This paper	Addgene # 246933

(*Continued on next page*)

Please cite this article in press as: Jin et al., Functional RNA splitting drove the evolutionary emergence of type V CRISPR-Cas systems from transposons, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.004

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
phU6-LaTranC-mini-sgRNA	This paper	Addgene # 246934
Software and algorithms		
GraphPad Prism	GraphPad Prism software	N/A
Adobe Illustrator	Adobe	N/A
UCSF Chimera	Pettersen et al.[58]	https://www.cgl.ucsf.edu/chimera/
AlphaFold2	Jumper et al.[59]	https://deepmind.google/technologies/alphafold/
AlphaFold3	Abramson et al.[60]	https://alphafoldserver.com/
STREME	Bailey[61]	https://meme-suite.org/meme/
MMSeqs2	Steinegger et al.[62]	https://github.com/soedinglab/MMseqs2
HMMER	Eddy[63]	http://hmmer.org/
Foldseek	van Kempen et al.[64]	https://search.foldseek.com/
IQ-TREE2	Minh et al.[65]	https://github.com/iqtree/iqtree2
Code for gene editing efficiency analysis	Jin et al.[66]	https://github.com/ReiGao/GEanalysis

EXPERIMENTAL MODEL AND SUBJECT DETAILS

E. coli transformation

FastT1 *E. coli* competent cells were used for plasmid amplification. DH5α *E. coli* competent cells were used for PAM depletion and DNA interference assays. Transformed *E. coli* cells were grown at 37°C in Lysogeny Broth (LB) medium supplemented with appropriate antibiotics.

Mammalian cell lines and culture conditions

Human HEK293T cells (ATCC, CRL-3216) were cultured in Dulbecco's Modified Eagle's medium (DMEM, Gibco) supplemented with 10% (vol/vol) fetal bovine serum (FBS, Gibco) and 1% (vol/vol) Penicillin-Streptomycin (Gibco) in a humidified incubator at 37°C with 5% CO_2.

METHOD DETAILS

General experimental methods

Unless otherwise noted, DNA was amplified by PCR using 2 × Phanta Max Master Mix (Vazyme Biotech) or 2 × TransStart FastPfu PCR SuperMix (TransGen), while plasmids were assembled using a ClonExpress II One Step Cloning Kit (Vazyme Biotech). The *E. coli* DH5α cells used in the bacterial assays were obtained from Takara (*E. coli* DH5α Electro-Cells). The *E. coli* cells used for vector construction were cultured in LB liquid or on agar medium with appropriate antibiotics (ampicillin: 100 mg/L; kanamycin: 50 mg/L) at 37°C. All plasmids used in this study were constructed by Gibson assembly or mismatch PCR. The primers used for bacterial and human cell assays and sequences of target sites are listed in Table S2.

Initial protein collection and curation

We obtained amino acid sequences comprising 1,712 TnpB homologs and 90 Type V-U CRISPR effectors (uncharacterized Type V CRISPRs) from Shmakov et al.,[19] along with 416 Type V-U CRISPR effectors and 5,752 TnpBs from Han et al.[21] Additionally, ISDra2 TnpB sequence was obtained from Sasnauskas et al.[31] Furthermore, we collected data from several studies that compiled various Cas12 proteins, ranging from Cas12a to Cas12n.[32,33,35,40,44,67–76] We applied CD-HIT (v4.8.1)[77] with the parameters "-i all_proteins -o merged_proteins -G 0.5 -aS 0.5" to eliminate redundancy among proteins. In total, we acquired 7,012 TnpBs, 114 classified Cas12 proteins, and 294 V-Us. These protein sequences were used for phylogenetic analysis in Figure 1**C**. Detailed protein information is provided in Table S1.

A collection of representative TnpB and Cas12 protein sequences was assembled to search for TranC candidates. The objective was to employ an impartial benchmark representing TnpB and Cas12 proteins. To achieve this, a comprehensive set of amino acid sequences was constructed, encompassing a diverse range of TnpB and Cas12 types. It comprised 50 randomly selected TnpB sequences (10 from each of the Typical, Derived, RIIr-3, RIIr-5, and RIIIr-4 categories) from Han et al.,[21] ensuring that none contained the ambiguous amino acid "X". Furthermore, the 114 classified Type V CRISPR effectors described earlier, which are non-redundant proteins, were incorporated. This set ensemble of 164 protein sequences was employed for three TranC mining methods, including motif search, profile search, and sequence search, facilitating the accurate and reliable identification of TranC candidate proteins.

Identification of TranC candidates

Compilation and annotation of microbial datasets

Considering the substantial size and inherent redundancy of metagenomic sequence data,[78] we focused our initial TranC candidate mining efforts on four high-quality, representative microbiome datasets: GTDB,[79] Almeida et al.,[80] Bai et al.,[81] and Levy et al.[82] The GTDB (RS220, released 24th April, 2024)[79] provides standardized and well-annotated genome sequences generated from publicly available microbial sequencing data, enabling access to high-quality, diverse, and non-redundant microbial genomes. The data from Almeida et al. corresponds to version 2 of the Unified Human Gastrointestinal Genome (UHGG) catalog released by MGnify,[78,80] whereas the datasets from Bai et al. and Levy et al. were derived from plant microbiomes. For these latter three datasets, we annotated open reading frames (ORFs) using Prodigal (v2.6.3),[83] applying the "-p meta" parameter specifically designed for metagenomic data annotation, which accommodates the diverse taxonomic composition typically observed in metagenomic sequences.[84]

Motif search

We detected conserved motifs shared between TnpB and Cas12 proteins using the STREME algorithm.[61] The reference dataset of TnpB and Cas12 proteins used in this analysis is described in "***Initial protein collection and curation***" above. Three motifs were identified (Figure 1B): Motif 1 is located in the first segment of the tri-split RuvC domain; Motif 2 was identified within the TNB domain; and Motif 3 is located at the junction between TNB and RuvC domain. Subsequently, these motifs were subjected to the QFAST algorithm against the microbial genome and metagenome dataset described above to search the candidate proteins harboring at least one motif with an E-value significance threshold of 1e-5.[85] In detail, the motifs were represented as position-dependent probability matrices, which describe the likelihood of each possible amino acid at each position in the motif. For sequences containing ambiguous characters, QFAST calculates a weighted average of the log-odds scores for matching characters, using their background frequencies as weights. The algorithm assumes that the distribution of the product of p-values from motif scores can be approximated by the distribution that results from multiplying independent random variables, each of which is uniformly distributed between 0 and 1. The overall significance of the product of n independent p-values is determined by calculating the statistical distribution of these products, with an overall p-value threshold of 0.01, indicating significant motif enrichment. Among the 164 TnpB and Cas12 proteins used for motif discovery, 153 proteins harbor at least one motif. Proteins harboring at least one motif were considered successful hits, which aid in the identifying proteins with disrupted motifs or rearranged domains.[7,86] This analysis led to the identification of 1,804,743 initial non-redundant proteins.

Profile search

We further constructed the Hidden Markov Models (HMMs) representing conserved regions shared between TnpBs and Cas12s. The dataset of TnpB and Cas12 proteins used here is consistent with the motif search. Based on multiple sequence alignments (MSAs) of amino acid sequences generated by MAFFT (v7.505),[87] we focused on the relatively conserved RuvC and TNB domains in TnpB and Cas12 proteins to construct three HMMs. Specifically, we focused on three regions: (1) the conserved N-terminal region of RuvC (RuvC-I and part of RuvC-II), (2) the TNB domain and conserved C-terminal region of RuvC (TNB-RuvC-III), and (3) the full-length of the RuvC-TNB domain. These three HMMs were subsequently employed in a profile search with the HMMER suite (v3.3.2) against the microbial genome and metagenome dataset described above to search the candidate proteins harboring at least one HMM, using a significance threshold of 1e-5.[63] This approach yielded 99,445 initial non-redundant candidate proteins for further analysis.

Sequence search

A BLASTp-based sequence search was performed against the aforementioned microbial genome and metagenome dataset, utilizing a significance threshold of 1e-5. The query dataset of TnpB and Cas12 proteins used in the other two search methods was applied here. This search was limited to proteins of a minimum length of 300 amino acids and a maximum sequence identity of 90% with the input proteins, in order to avoid redundancy. This screening process yielded 62,855 initial non-redundant candidates.

Protein filtration

Based on the non-redundant results obtained from the above three searching methods, we first obtained the proteins located more than 1 kilobase away to contig termini to ensure functionality and minimize annotation errors, and then applied seven filtering criteria to construct a conservative candidate dataset: 1) sequence divergence from known classified Cas12 proteins (identity $\leq$ 0.35); 2) presence of a CRISPR array within a 5-kilobase flanking region detected by MinCED (v0.4.2)[88]; 3) utilizing MinCED with the parameter "-minNR 4", we discerned CRISPR arrays with a stringent threshold of at least four repeats, given that CRISPR-Cas systems featuring fewer repeats are likely non-functional[89]; 4) protein sizes falling within the range of 300 to 625 residues, akin to TnpBs[20]; 5) exclusion of proteins associated with adaptation modules (Cas1, Cas2, and Cas4) within 50-kilobase flanking region (see details in next section); 6) elimination of proteins potentially representing intact transposons (see details in next section); 7) removal of protein redundancy, retaining only one representative per homolog group (inter-homology identity and coverage exceeding 50%).

Based on the results of the above filtering methods, we initially identified 130 candidates. To further expand the candidate pool, we conducted a second round of tBLASTn-based search in the NCBI NT database, using an E-value threshold of 1e-5, with the initial candidates serving as bait. We applied the same filtering criteria as before, ultimately identifying 146 candidates, each linked to the host species taxonomy information retrieved from the NCBI Taxonomy database.[90] Finally, the candidate CRISPR arrays were manually verified to ensure accuracy. Please refer to Table S1 for the detailed list of these TranCs.

Transposon and adaptation modules identification

TnpB proteins that establish a functional coupling with CRISPR array are expected to have lost their transposition activity.[5,18,19,26] Therefore, when we identified the candidate TranCs, we filtered out all intact transposons adjacent to CRISPR arrays, defined by the

presence of a TnpA transposase within 1.5 kb. Leveraging the ISfinder database,[91] we retrieved sequences for 157 Y1 TnpA transposases and 49 serine TnpA recombinases. Employing both sequence (BLASTp) and profile (HMMER) searches with an E-value significance threshold of 1e-5, we identified putative TnpA ORFs positioned 1.5-kilobase flanking the ORF of each candidate, accounting for their various genetic arrangements.[92,93] The adaptation proteins, including Cas1, Cas2, and Cas4, were identified *via* HMM profile searches. We employed HMM models previously generated[21,93] and the HMMER software suite (v3.3.2) with a significance threshold of 1e-5 for this purpose.

Evolutionary analysis

Reconstruction of a TnpB and Cas12 supertree

Our analysis involved a total of 7,012 TnpBs, 114 classified Type V CRISPR effectors, 294 V-Us, and 146 candidates (see Table S1). Due to considerable sequence divergence, we employed a specialized pipeline to construct a phylogenetic tree. MSAs were generated using MUSCLE (v5.1)[94] with the parameter "-super5", which offers high sensitivity for large, deeply diverged homologs. The resulting MSA was automatically trimmed using trimAl (v1.4.rev22)[95] with the parameter "-gt 0.5" and subsequently manually corrected to remove unclear or gapped regions, ensuring that the catalytic residues D, E, and D were conserved across the sequences. To address potential topological bias due to long-branch attraction, we created an initial draft tree using FastTree2 (v2.1.11).[96] According to a previous study,[21] singleton nodes with excessively long branches (branch length $\geq$ 1.5) were removed unless they displayed significant homology to other TnpBs or Cas12 alignments (HMMER with an E-value significance threshold of 1e-5). For the final comparison, we performed five independent runs using IQTREE2 (v2.2.0)[65] with the following parameters: VT+F+R10 substitution parameters (determined to be optimal for this alignment *via* ModelFinder v2.3.2[97]), 2,000 bootstraps, -nstop 5, –ninit 100, –ntop 100, –nbest 20, and the –bnni option to reduce bootstrap overconfidence in the case of model violations. The tree with the optimal likelihood score from the five parallel runs was selected as the final tree. Bootstrap support values for the best tree were computed using the "–bnni" option to account for model violations during bootstrap determination. Given the extensive investigation of ISDra2 TnpB and its position in the most prominent TnpB phylogenetic group,[9,31,98] we selected a clade containing this TnpB protein as the root of the tree. The consistency of TnpB placement in our phylogenetic tree aligns with a 2023 study that provided a panoramic view of TnpB evolution.[21] The resulting topology reflects similar relationships among the five known isoforms of TnpB. Additionally, TnpBs are observed to intermingle with various subtypes, such as Cas12n, in agreement with the previously documented independent origins of Cas12 subtypes.[21,33,35,40] Phylogenetic trees and DNA sequence alignments have been provided in Data S3.

Independent phylogenetic analyses of TranC clades

Based on the supertree, we identified 25 TranC candidate clades, each containing at least two TranC candidate proteins coupled with CRISPR arrays (Figure 1C; Table S1). These clades may represent early transitional forms from transposons to CRISPR systems. However, the TnpB family has extreme diversity and is an order of magnitude more diverse than the IscB family. Furthermore, there were multiple transitions from TnpB to Cas12 in the origination of Type V systems. Therefore, a single phylogenetic analysis (Figure 1C) might not fully capture the complete evolutionary landscape. In order to depict evolutionary scenarios of these 25 clades, both sequence-based and structure-based homology searches were employed with the objective of identifying as many homologs as possible. For sequence searching, we conducted a phmmer search (E-value < 1e-5) for each TranC candidate in MGnify (one of the largest metagenomic resources of public datasets, which contains over 2.4 billion non-redundant sequences).[78] A total of 764,800 initial homologs were identified in this step. Concurrently, the "TM-align" mode of Foldseek was employed, in conjunction with protein structures predicted using AlphaFold3,[60,64] to traverse all available protein structure databases. The databases included BFVD 2023_02, AlphaFold/UniProt50 v4, AlphaFold/Swiss-Prot v4, AlphaFold/Proteome v4, BFMD 20240623, CATH50 4.3.0, MGnify-ESM30 v1, PDB100 20240101, and GMGCL 2204. Hits were initially screened using a TM-score cutoff of 0.5 or higher. A total of 485,750 initial homologs were identified in this step.

We then performed a two-step filtration of the hits obtained from both sequence-based and structure-based homology searches. First, we removed homologs with low coverage and similarity by applying thresholds of coverage $\geq$ 0.70 and identity $\geq$ 0.30. Next, within the joint homolog set generated through both sequence- and structure-based searches, we removed redundancy by grouping similar proteins (identity > 90%) using MMSeqs2 (v14.7e284)[62] and selecting a representative protein from each group. To reduce the likelihood of false positives due to low-quality metagenomic assemblies, we excluded samples where the contig end was positioned less than 500 bp away from the ORF end. Following these filtering steps, a total of 3,276 homologs were obtained.

Subsequently, we constructed MSAs using MAFFT (v7.505) with E-INS-I mode, which is suitable for sequences containing large non-comparable regions. This alignment included the identified homologs and the samples for initial defining each TranC candidate clade. Finally, 25 reliable phylogenetic trees for each clade were constructed using IQTREE2, following the same parameter settings as described above. Among these 25 TranC candidate clades, six with a clear TnpB sister group (maximum-likelihood branch support $\geq$80%) were defined as TranC clades. Consequently, each TranC clade comprises a TranC group and a TnpB sister group sharing a common ancestor. Although occasional CRISPR array associations were observed for a few TnpB sister group members, over 95% lacked CRISPR association, suggesting that they predominantly represent ancestral TnpB systems. Phylogenetic trees and sequence alignments have been provided in Data S3.

Functional studies of TranCs based on metadata

Outline of functional and divergence analyses

To fully understand the transition from transposon systems to the CRISPR system, we performed a series of comparative analyses focusing on *bona fide* TranC clades (Figures 2C-2E and S2A-S2C; Data S1.3–S1.5) and representative non-TranC clades (Data S1.6–S1.11). These analyses included assessments of transposon features, CRISPR features, and catalytic activity, providing insights into the evolutionary divergence and functional specialization of TranC proteins.

For transposon features relevant to element mobility, we performed the following analyses: (1) The TnpA association analysis. This analysis was conducted in accordance with the "Transposon and adaptation modules identification" section above. (2) The left/right end (LE/RE) sequences analysis. We first retrieved 108, 55, and 248 terminal sequences respectively for the IS605, IS607, and IS1341 transposons from ISfinder database.[91] We then conducted BLASTn searches (E-value < 1e-5) within a 1.5-kilobase region surrounding the ORF of each candidate to identify the conservation of LE/RE sequences. The sequence identities of the successful searching results were annotated adjacent to each phylogenetic tree. (3) The TnpB copy number analysis. To evaluate the TnpB copy number within each contig, we conducted BLASTp searches using the candidate protein as a query, employing two sets of thresholds: identity $\geq$ 0.70 and coverage $\geq$ 0.70, as well as identity $\geq$ 0.95 and coverage $\geq$ 0.95, to generate two independent results. (4) The reRNA covariance analysis. This analysis was described in an independent section "Analysis of TnpB reRNAs" below. (5) The identification of co-occurrences of other TnpBs within the genome. Two HMM profiles, PF01385 and PF07282, were employed to detect additional TnpB homologs within the same contig through an HMMER search (E-value < 1e-5).

For CRISPR features, we performed the following analyses: (1) The adaptation module characterization. This analysis was conducted in accordance with the "Transposon and adaptation modules identification" section above. (2) CRISPR array profiling and co-occurrence analysis of other arrays. MinCED (v0.4.2)[88] was used to identify CRISPR arrays located near ORFs (< 500 bp) and co-occurring arrays across the contig level. (3) The origin of repeat sequences was predicted through comparative sequence analysis. The repeat sequences were subjected to BLASTn searches against the CRISPRCasTyper (v1.8.0)[99] and CRISPRCasdb (2022 release)[89,100] databases. All outputs with an E-value significance threshold of 1e-5 were listed in Table S1. (4 and 5) The analyses of PAM sequences, spacer sources, and tracrRNA-crRNA covariance are presented in separate sections below: "Prediction of PAM sequences and spacer source" and "Analysis of tracrRNA-crRNA hybrids in TranCs." (6) For identification of co-occurrences of other Cas12s within the contig, we collected 40 HMMs[99] representing different Cas12 subtypes for profile searches (E-value < 1e-5) to identify co-occurring Cas12s within the contig.

For catalytic activity, we assessed the conservation of nuclease catalytic sites by carefully examining the unpruned MSA of individual clades. To illustrate, the catalytic sites of ISDra2 TnpB are D191, E278, and D361.[30,31] Although variations may occur in these sites or their surrounding sequences, they generally remain conserved. Notably, D361 is situated within the "DRDXN" motif in ISDra2 TnpB, and in "derived TnpB" group, it is predominantly present within the "NADXN" motif.[21] Furthermore, we found that some of TranC proteins lack part of the catalytic site, due to encoding a considerably shorter ORF compared to other homologs. For example, we manually examined the genomic locus of a clade3 TranC member from subclade4, designated "MGYG000001704_01928", using the LaTranC protein as a reference. We identified a frameshift, likely caused by an indel, that disrupted the ancestrally intact ORF. Given that all subclade4 homologs encode this truncated ORF, the indel likely occurred in their common ancestor.

Prediction of PAM sequences and spacer source

Using spacer sequences as queries, we conducted BLASTn searches against multiple databases, Human Gut Virome Database (GVD),[101] Gut Phage Database (GPD),[102] PHAge Search Tool Enhanced Release (PHASTER),[103] NCBI RefSeq Viral and Plasmid Genomes, Integrated Microbial Genomes/Virus (IMG/VR),[104] and PhageScope.[105] Given the relatively short length of spacers (20-30 bp), we applied a stringent cut-off of 85% identity and 85% coverage to enhance search reliability. For candidates with multiple hits, redundant matches were removed, particularly when the same orthologous locus appeared in related viral lineages. Spacer hits to viral metagenome databases were all classified as viral-derived spacers, although that such databases can contain a small proportion of host-derived sequences. After identifying hits corresponding to distinct spacers from the same candidate, we extracted the 5-bp upstream (5′-flanking) sequences and utilized this information to construct PAM motifs.

Meta-transcriptome analysis

In the absence of a centralized database for meta-transcriptome data, we undertook an exhaustive literature review to identify studies that have generated meta-transcriptome data across diverse ecological contexts. In total, we obtained 1,576 meta-transcriptomic datasets from 27 projects, including 1,107 from human gut microbiota, 218 from marine microbiota, and 251 from soil microbiota datasets derived from published studies and NCBI SRA database.[106,107] Using BWA-MEM,[108] we mapped reads to TranC loci, along with their flanking 5-kilobase regions. Since these meta-transcriptome datasets were unpaired with the metagenome datasets encoding TranCs, we required > 50% read coverage in BWA-MEM to ensure confident identification of homologs, following conventional practices in meta-transcriptome data mining.[109] The results were processed into BAM files, and read depth was calculated using the Samtools (v1.17)[110] depth command.

Analysis of tracrRNA-crRNA hybrids in TranCs

Based on prior knowledge of Type V CRISPR systems (*e.g.*, Cas12b, Cas12e, Cas12f, Cas12k and Cas12n)[33,34,40,67,111] and the meta-transcriptome analyses above, it was suggested that the tracrRNAs of the TranCs may be positioned between the 3′-terminus of the effector ORF and the upstream region of CRISPR arrays. Therefore, to identify the tracrRNA, we extracted the sequence starting 300 bp upstream of the last nucleotide of the candidate genes and extending to the first CRISPR repeat downstream. IntaRNA

(v3.4.1)[112] was used with parameters "–outOverlap Q" and "–outMode C" to identify the hybrid stem formed at the 3′ end of the tracrRNA and the CRISPR repeat region. The region was extended by 15 bp downstream to define the 3′ end boundary of the tracrRNA. Based on this identified end, a 250 bp segment upstream was designated as the putative tracrRNA. The tracrRNA sequences were then aligned using MAFFT (v7.505)[87] with the "E-INS-I" algorithm. Subsequently, a preliminary clade-specific tracrRNA consensus structure was inferred using ConsAliFold (v0.1.3)[113] with default parameters. An initial covariance model was constructed with the cmbuild program of Infernal (v1.1.5)[114] and subsequently calibrated using cmcalibrate. Subsequently, this preliminary model was used for iterative scanning of the TranC members from the corresponding TranC clade by the CMsearch program with an E-value threshold of < 1e-5, the identified sequences were concatenated to the corresponding repeat sequences with a "AAAAA" linker to obtain the tracrRNA-crRNA hybrids. Manual inspection was performed to ensure the confident 3′ terminal of tracrRNA, which can pair to CRISPR repeat. Finally, the tracrRNA-crRNA hybrids were used to generate the final CRISPR RNA covariance models for each representative TranC clade. These models were then verified with R-scape at an E-value threshold of 1e-5.[115] R2R (v1.0.6)[116] was used to draw consensus structure diagrams and assess secondary structure conservation (parameters: –GSC-weighted-consensus 3 0.97 0.9 0.75 4 0.97 0.9 0.75 0.5 0.1). Covariance model files and RNA alignments have been provided in Data S3.

Analysis of TnpB reRNAs

The reRNAs consist of a scaffold sequence encoded by the transposon and a guide sequence located just outside the transposon's right end. The reRNA scaffold is conserved across different transposon copies, while the flanking guiding sequences vary.[9,24,25] However, most members of the TnpB sister groups were derived from metagenomic data, and the incomplete assemblies (with contig lengths typically shorter than 10 kb) may have limited the recovery of multiple transposon copies. To address this, we further identified highly conserved homologs of these sister TnpBs (with similarity > 90%) across the entire MGnify database using "mmseqs search" program of MMseqs2 software,[117] with the aim of detecting more transposition events for reRNA covariance model construction. Using this expanded dataset of TnpBs, we extracted sequences spanning 300 bp upstream and 300 bp downstream of the last nucleotide of the TnpB ORFs. These sequences were clustered using CD-HIT (v4.8.1)[118] with 95% length and alignment coverage to remove duplicates. The MAFFT (v7.505) with the "E-INS-I" pattern was further used for sequence alignment. The alignments were manually inspected to locate the transposon's right end, typically found at the junction between highly conserved and poorly conserved sequence regions, marked by a pronounced drop in mean pairwise identity across sequences. The 250-bp sequence window upstream of the 3′ end of the reRNA scaffold was extracted. The subsequent covariance model construction followed the same methodology as described in "***Analysis of tracrRNA-crRNA hybrids in TranCs***". Covariance model files and RNA alignments have been provided in Data S3.

For experimental examining reRNA-guided activity of TranCs, we also used these reRNA covariance models for detecting the specific reRNA sequences (Figure 3C). Using CMsearch with an E-value threshold of < 1e-5, we scanned for reRNA sequences across loci of clade-specific, closely related TnpBs identified through sequence- and structure-based homology searches. Identified transposons were ranked based on three criteria: (1) copy number (≥ 2 copies, indicative of potential mobility, as previously reported,[21] (2) sequence similarity (> 40% similarity to TranCs), and (3) structural similarity (TM-score > 0.6). The majority of TnpB and TranC sequences were derived from metagenomic datasets such as MGnify.[78] MGYP entries are MGnify accession identifiers for representative sequences of non-redundant protein clusters. Therefore, a single MGYP accession may correspond to multiple similar sequences. For example, MGYP004455556179 contains several DNA sequences, among which MGYP004455556179_1 and MGYP004455556179_9 were designated Ana1 TnpB and Ana2 TnpB, respectively. Both were included in reRNA analysis and experimental testing (Figures 3D, S3E and S3F). The representative protein sequence Ana1 TnpB was used for phylogenetic analysis (Figure S2C). For other MGYP samples, a DNA sequence from each corresponding MGYP accession was selected based on CMsearch results for experimental testing. From this analysis, 12 reRNAs meeting all thresholds were selected for functional validation of TranC activity (Table S2).

Conservation analyses of LaTranC tracrRNA stems

During our exploration of clade3 homologs, we detected two MGnify representative sequences exhibiting substantial similarity (> 85%) to LaTranC. Within the MGnify homolog clusters corresponding to these sequences, there are numerous members, but only four were associated with sufficiently long contigs that included the 3′ flanking noncoding regions. These four noncoding sequences, together with LaTranC's own tracrRNA, were employed in the construction of an MSA. We then examined the distribution of substitutions across four stem regions to evaluate their conservation. The test was performed using a simplified Binomial Test. Specifically, we assumed that the probability of retention at each site is homogeneous, under the null hypothesis of a uniform retention probability of 0.25. The nucleotides corresponding to different stems were treated as independent observations sampled across multiple times, allowing us to assess whether the probability of mutations occurring at different stems deviates significantly from the expected value.

Computational Structural Analyses

Compilation of public protein structure data

We assembled a collection of reported protein structures, including ISDra2 TnpB (PDB: 8EXA[31]), Cas12a (PDB: 5B43,[27] 5XUT[119]), Cas12b (PDB: 5WQE,[120] 5WTI[121]), Cas12c (PDB: 7VYX[122]), Cas12e (PDB: 6NY2,[44] 7WB1[123]), Cas12f (PDB: 7C7L[37]), Cas12g (PDB: 6XMF[124]), Cas12i (PDB: 6W64,[125] 6LU0[126]), Cas12j (PDB: 7M5O[127]), Cas12k (PDB: 7PLA[128]), Cas12l (PDB: 7YOJ[71]) and

Cas12m (PDB: 8HHL[36]). In cases where a particular Cas subtype had multiple structures available, one representative structure was selected at random.

Prediction of protein structures

To predict protein structures, we utilized AlphaFold2 (v2.2.0)[59] to predict protein structures, unless otherwise stated. All the predicted structures exhibited an average per-residue confidence metric pLDDT $\geq$ 70, signifying high reliability.[59] To construct a structure-based tree from a streamlined dataset, the pairwise structural alignments were computed with TM-score and adjusted by residue distance measures.[45,46] These values were assembled into a similarity matrix, normalized with the min-max method, and clustered by UPGMA.[129]

Positively charged amino acids counting

Leveraging the protein structures, we computed the numbers of surface-exposed amino acids. Accessible surface areas (ASA)[130] were calculated using the DSSP algorithm,[131] considering amino acids with a relative ASA greater than 0.25 as surface-exposed. Positively charged amino acids were defined as arginine (R), histidine (H), and lysine (K).

Protein and RNA structure comparison

We used USalign (v20220626)[46] for pairwise comparisons of protein structures and of RNA structures, separately. We also employed the TM-score parameter to quantify structural similarity.[45] Protein structure visualizations were performed using UCSF Chimera.[58] While TM-score is generally considered size-independent,[45] we observed that guided structural comparisons yielded more precise structural overlays for RNA. Specifically, although full protein structures were compared, we focused on substructure comparisons when analyzing RNAs of TnpB and Type V CRISPR systems. Informed by RNA secondary structure predictions with mfold (v3.6),[132] we identified substructure breakpoints and chose substructures within ISDra2 TnpB that corresponded to crRNA, specifically encompassing portions of stem 4 and the pseudoknot (PK) region.

E. coli functional assays

PAM depletion assays

The p15a-based pEffector vectors were employed for the expression of candidate TranC proteins and non-coding RNAs, which were constructed based on the p15a-pTac-*dfncpf1* plasmid, generously provided by the Bian Wu lab at the Chinese Academy of Sciences (CAS). As shown in Figure S3A, the pEffector vectors comprise Ptac-driven TranC and Pj23119-driven guide RNA expression cassettes. For CRISPR RNA, the RNA components included "repeat-spacer-repeat" mini-CRISPR arrays, along with the non-coding region positioned between the TranC ORF and its adjacent proteins, potentially containing the tracrRNA sequence (Table S2). The repeat sequences used here were derived from the first repeat located immediately downstream (3′) of the protein-coding sequence. For all Pj23119-driven reRNA expression cassettes, a 5′ HH ribozyme and a 3′ HDV ribozyme were included to ensure reRNA maturation. The TranC protein sequences, which were codon-optimized for expression in *E. coli*, as well as their corresponding guide RNA sequences, were commercially synthesized by Nanjing GenScript Biotech Co., Ltd., and Sangon Biotech (Shanghai) Co., Ltd. For MlCas12n and CgCas12n, the protein coding sequences were constructed based on the p15a_MlCas12n and p15a_CgCas12n plasmids generously provided by the Quanjiang Ji lab at ShanghaiTech University. The corresponding guide RNAs were adopted from previously reported mature CRISPR RNAs.[40] Details of TranC coding sequences can be found in Table S1. All these sequences were cloned into the pEffector vector using a ClonExpress II One Step Cloning Kit (Vazyme Biotech).

The pMB1-based pTarget plasmid library was constructed, featuring five randomized nucleotides positioned immediately upstream of the *GFP*-T1 target sequence (5'-NNNNN-target, please refer to Table S2) and an AmpR expression cassette to confer ampicillin antibiotic resistance. This construction followed a previously reported protocol with minor modifications.[133] Forward and reverse primers introducing the five "NNNNN" nucleotides were synthesized by BGI, Beijing. Subsequently, the pTarget plasmid library was electroporated into *E. coli* DH5α cells already carrying the pEffector plasmid. The transformed cells underwent recovery in SOC liquid medium for 30 minutes and were then plated on LB agar supplemented with kanamycin and ampicillin, incubated overnight at 37°C. After harvesting the resulting colonies and extracting the plasmids (TIANGEN Biotech.), the regions containing randomized PAMs were amplified and subjected to deep sequencing (Novogene, Beijing). The analysis of deep sequencing data and the identification of PAM sequences adhered to well-established protocols.[76] In summary, the recognition of a specific five nucleotides as a PAM triggered plasmid cleavage, resulting in the loss of ampicillin resistance and subsequent *E. coli* cell death. The loss of the specific five nucleotides was subsequently discerned by analyzing the sequencing data. During PAM depletion analysis, TTAT PAMs were excluded because an endogenous TnpB encoded in *E. coli* DH5α was computationally predicted to recognize similar PAM motifs, which could potentially confound the depletion results.

In total, the CRISPR RNA-guided activities of LaTranC, RuTranC, Clo1TranC, MiCas12n, MlCas12n, CgCas12n, AnTranC, and Eu1TranC were identified using PAM depletion assays. We employed two strategies to optimize the pEffector vector. In one approach, the coding sequences of TranC members were incorporated into the Pj23119-driven expression cassette, as needed, to avoid potential protein-RNA overlap that might result in incomplete tracrRNA. In another approach, an additional reverse-oriented Pj23119 promoter was introduced downstream of the CRISPR loci, as needed, to enable bidirectional RNA expression. Despite applying both optimization strategies, Clo2TranC and Eu2TranC showed no detectable CRISPR RNA-guided activity. For Clo3TranC, competent cells were not successfully prepared in DH5α due to poor cell growth. All tested protein sequences are provided in Table S1, and the corresponding CRISPR loci with detailed optimization information are provided in Table S2.

Please cite this article in press as: Jin et al., Functional RNA splitting drove the evolutionary emergence of type V CRISPR-Cas systems from transposons, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.004

Small RNA sequencing

Small RNA sequencing was performed according to a previously reported protocol with minor modifications.[68] Total RNA from *E. coli* cells harboring pEffector plasmid was extracted using an EasyPure RNA Kit (TransGen Biotech Co.). Subsequent RNA purification and dephosphorylation steps were carried out using DNase I (NEB), phenol/chloroform, and T4 polynucleotide kinase (NEB). Illumina deep sequencing libraries were constructed according to the manufacturer's instructions (Synbio Technologies. LLC). Sequenced reads were subjected to quality control using Trimmomatic (v0.39)[134] and then mapped to pEffector using Bowtie2 (v2.4.5).[135] We utilized Samtools to compute the read depth. The number of reads initiating or terminating at each position (5′ and 3′ termini) was used to infer RNA maturation sites. This inference was supported by predicted base-pairing interactions in the repeat-anti-repeat region.

Plasmid DNA interference assay

As described in the PAM depletion assay, pTarget plasmids containing both a specific PAM and the *GFP-T1* target sequence (designated as "T", Target) or lacking both (designated as "NT", Non-target) were electroporated into *E. coli* cells already containing the pEffector vector. Transformed cells were allowed to recover in SOC medium for 30 minutes, followed by serial dilutions and plating on LB agar supplemented with ampicillin and kanamycin, and incubated overnight at 37°C.

Cryo-EM experiments and data analyses

Purification of the LaTranC ribonucleoprotein

LaTranC-coding sequence was cloned to the protein purification vector pJJGL001 (Addgene #: 180605) with the original cargo removed (Table S1). T7 promoter and sgRNA were inserted into the pACYC184 vector (GenBank Accession #: X06403). These two plasmids were co-transformed into Rosetta *E. coli* cells. The cells were cultured in TB medium supplemented with ampicillin (100 μg/ml) and chloramphenicol (37 μg/ml) at 37°C until optical density (OD600) reached 1.0-1.2. Then, expression was induced with 0.4 mM IPTG, and cells were grown overnight at 16°C. The cells were harvested by centrifugation and resuspended in lysis buffer (20 mM HEPES pH 7.5 at 25°C, 500 mM NaCl, 10% (v/v) glycerol, 25 mM imidazole, 2 mM $MgCl_2$) supplemented with 4 mM beta-mercaptoethanol and 1 mM phenylmethylsulfonyl fluoride (PMSF). The cell suspension was lysed by sonication and centrifuged at 12,000 rpm for 60 minutes at 4°C. The supernatant was loaded onto the pre-equilibrated Ni-NTA (QIAGEN) agarose beads three times. Then the resin was washed with 15 ml lysis buffer, 15 ml washing buffer (20 mM HEPES pH 7.5, 1 M NaCl, 10% (v/v) glycerol, 25 mM imidazole, 2 mM $MgCl_2$), and 15 ml tag cleavage buffer (20 mM HEPES pH 7.5, 500 mM NaCl, 10% (v/v) glycerol, 2 mM $MgCl_2$, 2 mM TCEP). The 10×His tag was cleaved by TEV protease overnight at 4°C. The flow-through was applied to an anion-exchange column (Mono QTM 4.6/100 PE) and eluted using a gradient of NaCl concentration from 0.2 to 1.0 M. Fractions containing LaTranC protein and sgRNA were collected, flash-frozen in liquid nitrogen and stored at -80°C. Peak fractions were pooled and further purified by size-exclusion chromatography (Superdex increase 200 10/300 GL; GE Healthcare) gel filtration column equilibrated with SEC buffer (20 mM HEPES, pH 7.5, 150 mM NaCl, 0.25% glycerol, 2 mM $MgCl_2$, 1 mM TCEP, at 25°C). The RNP fractions validated by polyacrylamide gel electrophoresis (PAGE) were aliquoted, flash-frozen in liquid nitrogen, and stored at -80°C for structure determination. The sequences of the LaTranC protein and sgRNA constructs are listed in Table S1 and Table S2, respectively.

Preparation of cryo-EM samples and data collection

The target dsDNA substrate used for cryo-EM samples was generated by annealing two oligonucleotides (sequences are provided in Table S1). For the 6-nt match ternary complex, the ternary complex was assembled by mixing the purified LaTranC-sgRNA ribonucleoprotein (RNP) complex with the dsDNA substrate at a molar ratio of 1:1.5 (RNP: dsDNA). The mixture was further purified after incubation by size-exclusion chromatography using a Superdex 200 Increase 3.2/300 column (GE Healthcare) equilibrated in reconstitution buffer (20 mM HEPES pH 7.5 at 25°C, 150 mM NaCl, 0.25% glycerol, 1 mM TCEP). Peak fractions corresponding to the ternary complex were collected, cross-linked with Bis (sulfosuccinimidyl) suberate sodium salt (BS3, Sigma-Aldrich), and used for cryo-EM sample preparation. For the 27-nt match ternary complex, the complex was assembled in the same manner as described above. However, following incubation with the dsDNA substrate, the mixture was directly applied to cryo-EM sample preparation without further purification. LaTranC-sgRNA-dsDNA complex (~0.7 μM) was cross-linked with Bis (sulfosuccinimidyl) suberate sodium salt (BS3, Sigma-Aldrich) and applied to Quantifoil Au 1.2/1.3, 300 mesh graphene oxide grids, which were glow-discharged in a Harrick Plasma for 15 s after a 2 min evacuation. Subsequently, grids were blot-dried with a pair of 55 mm filter papers (Ted Pella) for 3 s at 8°C and 100% humidity, followed by flash-freezing in liquid ethane using an FEI Vitrobot Mark IV. The same grid treatment was applied to other cryo-EM samples. Cryo-EM data were acquired on a Titan Krios electron microscope operating at 300 kV, equipped with a Gatan K3 direct electron detector featuring a Gatan Quantum energy filter. Micrographs were recorded in counting mode at a nominal magnification of 105,000× or 81,000×, resulting in physical pixel sizes of 0.8433 Å or 1.0825 Å. Defocus values were set between -1.3 μm and -1.5 μm, and the total exposure time for each movie stack resulted in a total accumulated dose of 50 electrons per $Å^2$ which was fractionated into 32 frames. Detailed data collection parameters can be found in Data S2.

Image processing and 3D reconstruction

Raw dose-fractionated image stacks were processed, including binning, alignment, dose-weighting, and summation, using MotionCor2.[136] CTF estimation, particle picking, 2D reference-free classification, initial model generation, heterogeneous refinement, non-uniform refinement, and local resolution estimation were performed using cryoSPARC.[137] Particle picking was also conducted using the topaz pipeline. Multiple rounds of 3D reference-based classification were carried out within RELION-3.[138] Further details regarding data processing can be found in Data S2.

Model building and refinement

Manual model building of RNA and DNA structures was carried out in Coot,[139] utilizing the refined cryo-EM density maps. Subsequently, complete models were refined against the EM maps in real space using Phenix,[140] employing secondary structure and geometric restraints. The final models underwent validation using Phenix. Comprehensive information on the structural validation of the final models is also presented in Data S2.

Functional assays in human HEK293T cells

Construction of vectors in human cells

The CMV-driven pCMV-LaTranC plasmid was employed for the expression of LaTranC protein. A codon-optimized LaTranC variant, featuring N-terminal and C-terminal nuclear localization signal (NLS) sequences tailored for human, was commercially synthesized by Nanjing GenScript. This synthesized LaTranC fragment was introduced into a CMV-driven Cas9 expression vector[141] through Gibson assembly, resulting in the generation of pCMV-LaTranC. For the expression of LaTranC sgRNA, the human U6-driven phU6-LaTranC-sgRNA vector was utilized. The sgRNA scaffold sequence was commercially synthesized by Nanjing GenScript and subsequently cloned into the human U6-driven sgRNA expression vector[142] *via* Gibson assembly, yielding phU6-LaTranC-sgRNA. The construction process of ISDra2 reRNA and artificial sgRNA were similar to LaTranC sgRNA, and an HDV ribozyme located at the 3′ end of the guide sequence was used for maturation, as previously described.[9]

Human cell culture and genome editing

HEK293T cells were cultured in DMEM medium (Invitrogen) supplemented with 10% FBS (Invitrogen) and maintained in a CO_2 incubator at 37°C with 5% CO_2 (Thermo Fisher). Cells were routinely tested for mycoplasma contamination using the Mycoplasma Detection Kit (TransGen Biotech, China). Prior to transfection, cells were treated with TrypLE Express (Gibco). Approximately 4.5×10^4 cells were seeded in 48-well plates, and 16 to 24 hours after seeding, when cells reached approximately 80% confluence, transfection was carried out. Transfection involved the use of 0.75 μl lipofectamine 2000 (Thermo Fisher Scientific) containing 750 ng LaTranC plasmid DNA and 250 ng sgRNA plasmid DNA. Plasmids used for transfection were prepared using a PureYield™ Plasmid Miniprep System (Promega). After a 72-hour incubation period, the culture medium was removed, and cells were washed with 1 × PBS (Thermo Fisher). Genomic DNA was extracted using a Triumfi Mouse Tissue Direct PCR kit (Genes and Biotech Co., Ltd, Beijing).

Amplicon deep sequencing and data analysis

The targeted loci were amplified from genomic DNA and subjected to deep sequencing. Specifically, two rounds of PCR were performed. In the first round, the target amplicon was amplified using 2 × Phanta Max Master Mix (Vazyme Biotech) with site-specific primers (Table S2). In the second round of PCR, the index sequences were added for library construction.[66,142] All PCR products were assessed by electrophoresis in 1.5% agarose gels. After two rounds of PCR, the final PCR products were purified using a GeneJET Gel Extraction Kit (Thermo Fisher). DNA concentrations were measured using fluorometry (Qubit, Thermo Fisher) or spectrophotometry (NanoDrop 2000, Thermo Fisher) and subsequently sequenced according to the Illumina manufacturer's protocol. The amplicon deep sequencing data were further analyzed as previously described.[66] Specifically, for each deep sequencing sample, reads were first merged. The analysis window was defined as a 27-nt region covering the spacer. Eight-nt sequences immediately upstream and downstream of this window were extracted as flanking index sequences. Only merged reads containing both flanking index sequences were retained, and the intervening sequence within the analysis window was extracted and aligned to the reference to identify indel mutations.

Arginine mutagenesis screen

To determine the mutation sites, we primarily considered structural conservation. Utilizing USalign,[46] a 3D structural alignment was conducted across known protein sequences of Cas12s, ISDra2 TnpB, and LaTranC. Considering the deep divergence of these homologs, we defined putative core regions as those shared across at least 30% homologs by following the previous work.[143] Sites located within these regions were selected for mutagenesis. In addition to this strategy, we also took domain annotation and secondary structure information into account, prioritizing sites within critical domains (*e.g.*, RuvC) or folded regions. A total of 353 single-residue sites meeting these criteria were selected for mutagenesis (Table S2), and a pooled plasmid library of LaTranC single-arginine mutants at these positions was synthesized commercially (Nanjing GenScript).

To obtain high-efficiency variants, a screening process was conducted in five rounds. In the first to fourth rounds, the post-transfection incubation time was reduced to 24-48 hours. Before screening, the pooled plasmid library was transformed into *E. coli* competent cells and plated on LB agar plates for overnight growth. To manage labor costs, pairs of colonies were mixed and cultured in liquid LB medium overnight. Plasmids were then extracted from mixed samples and subsequently transformed into HEK293T cells. Editing efficiencies at the *VEGFA-TTC-T1* target site were evaluated using the aforementioned deep amplicon sequencing.

In the initial screening round, we identified the top 10 high-efficiency samples, each with a pair of mutated plasmids, through PCR amplification and Sanger sequencing of the corresponding genomic DNA. In total, 17 distinct mutations were identified. In the second round of screening, pairs of these mutations were randomly combined, and the dual mutants with high editing efficiency were selected for the next round of screening. In the third to fifth rounds, an additional mutation from the pool of 17 Arginine mutations identified from the previous round and other rational selected mutations from putative core regions was introduced into the high-efficiency mutants. These mutations were introduced into the coding sequence of LaTranC using mismatch PCR. Despite encountering various challenges, such as mismatch PCR failures, a total of 138 variants with three mutations, 57 variants with four mutations, and

43 variants with five mutations were generated and tested. After completing five rounds of screening, the final eLaTranC variant with five Arginine mutations was successfully identified.

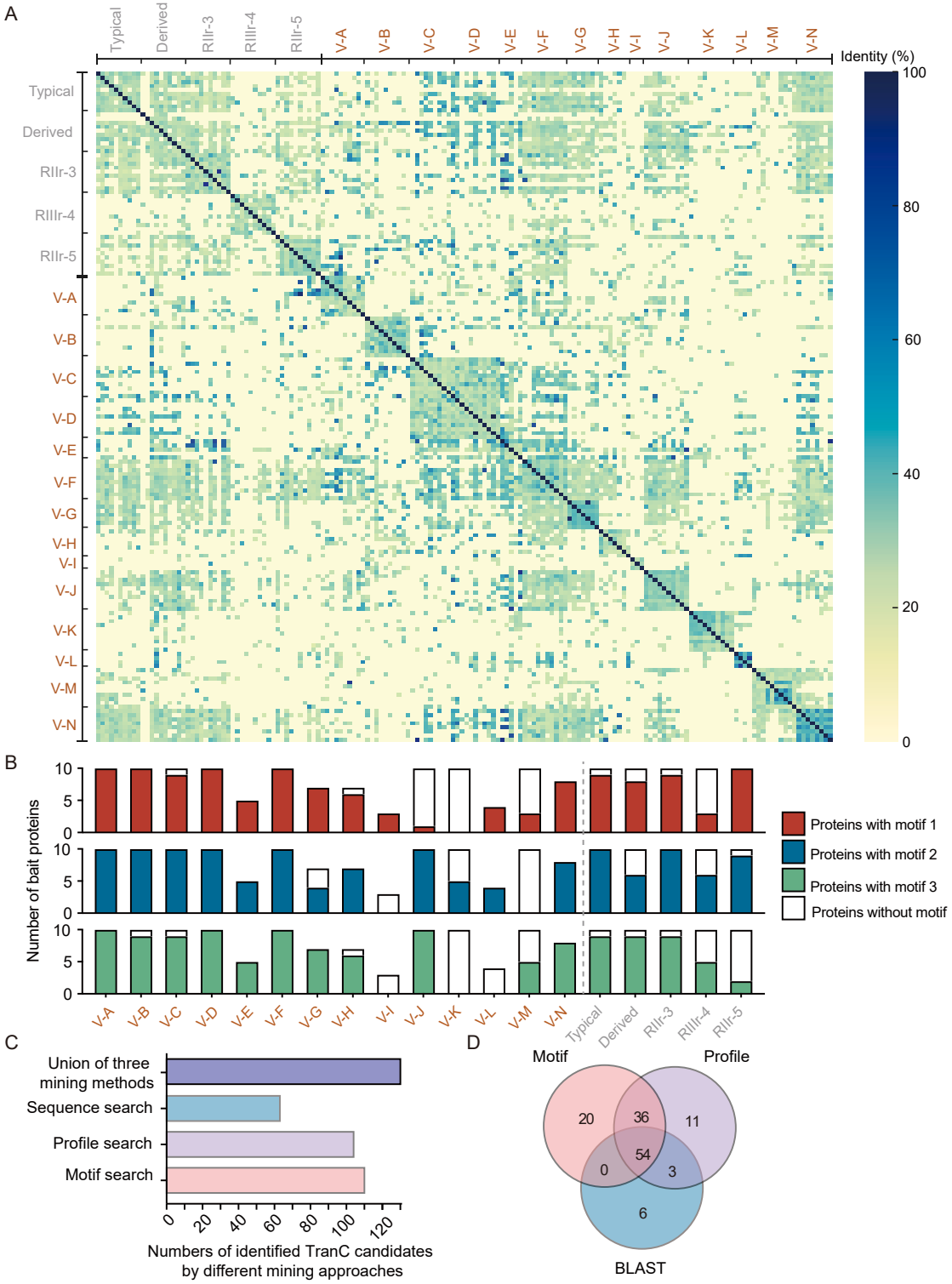
The guide RNA interchangeability

Guide RNAs—including TranC CRISPR RNAs and TnpB-derived reRNAs—were synthesized by Nanjing GenScript. Their sequences are listed in Table S2. The plasmids expressing these guide RNAs were generated by Gibson assembly.

QUANTIFICATION AND STATISTICAL ANALYSIS

All datasets were analyzed using GraphPad Prism 8 (GraphPad Software) for basic quantification and plotting. Statistical differences were assessed using two-tailed Student's *t*-tests, Mann-Whitney *U* tests, or Binomial tests, depending on the context.

Supplemental figures



(legend on next page)

Figure S1. Three mining methods for identifying TranC candidates, related to Figure 1

(A) Heatmap showing sequence similarity among Cas12 and TnpB proteins. The dataset includes 50 TnpB and 114 Cas12 proteins, which served as references for identifying TranC candidates. Based on pairwise sequence identity matrices, categories are listed on both axes, and color intensity represents identity levels from 0% to 100%. Exact protein sequences are provided in Table S1.

(B) Motif distribution observed in TnpB and Cas12 proteins. Each subtype harbors at least one conserved motif. The protein sequences used in this analysis are identical to those in (A).

(C and D) The number of TranC candidates identified by each of the three mining methods. (C) shows the number of candidates identified by each method, while (D) presents their overlap in a Venn diagram. After applying the three mining methods and subsequent protein filtration (see STAR Methods), a total of 130 TranC candidates were identified. These 130 proteins were then used as queries for a second-round tBLASTn search against the NCBI NT database, yielding 146 final candidates. Each approach exhibited specific detection biases, as detailed in Data S1.1.

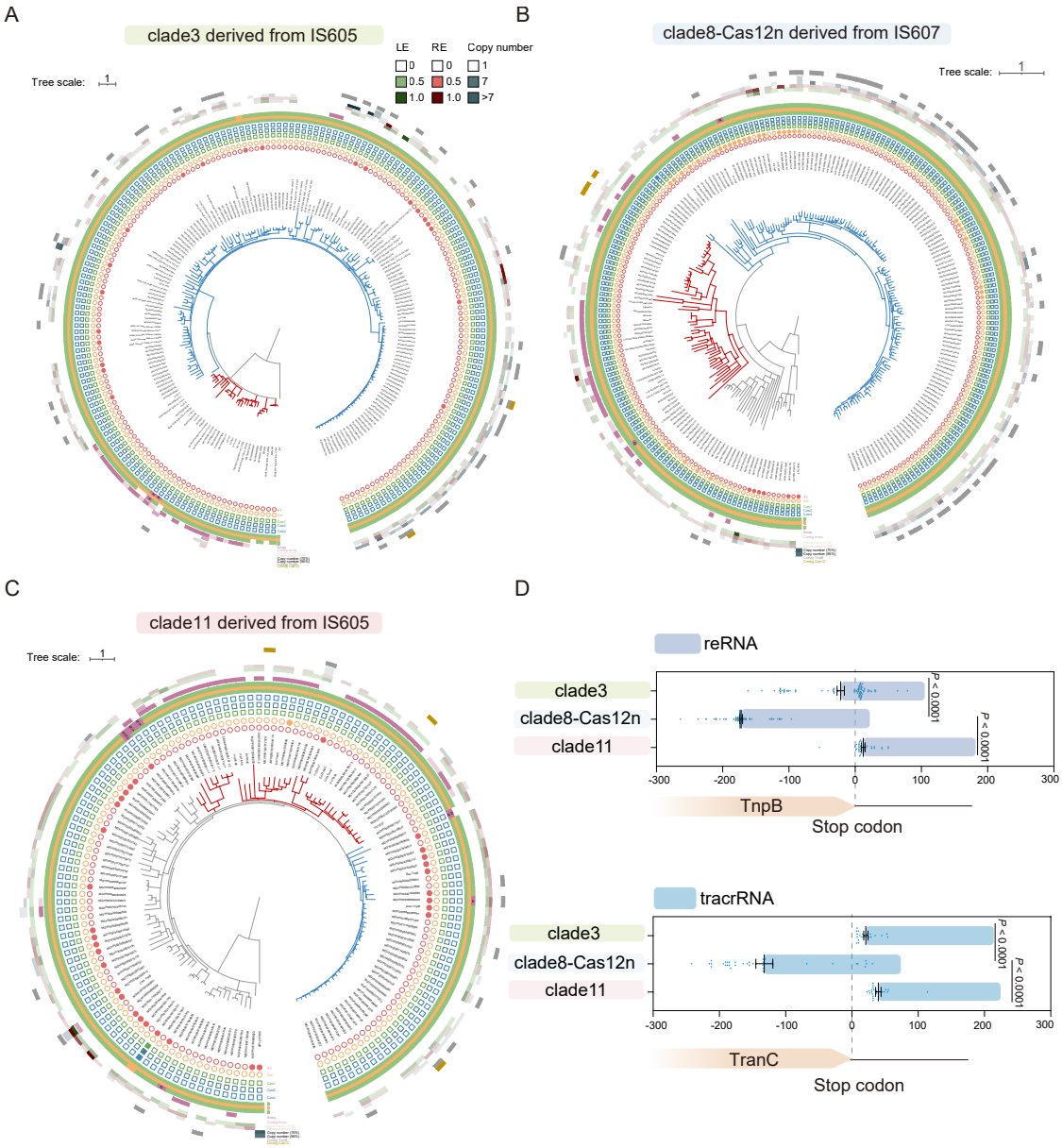
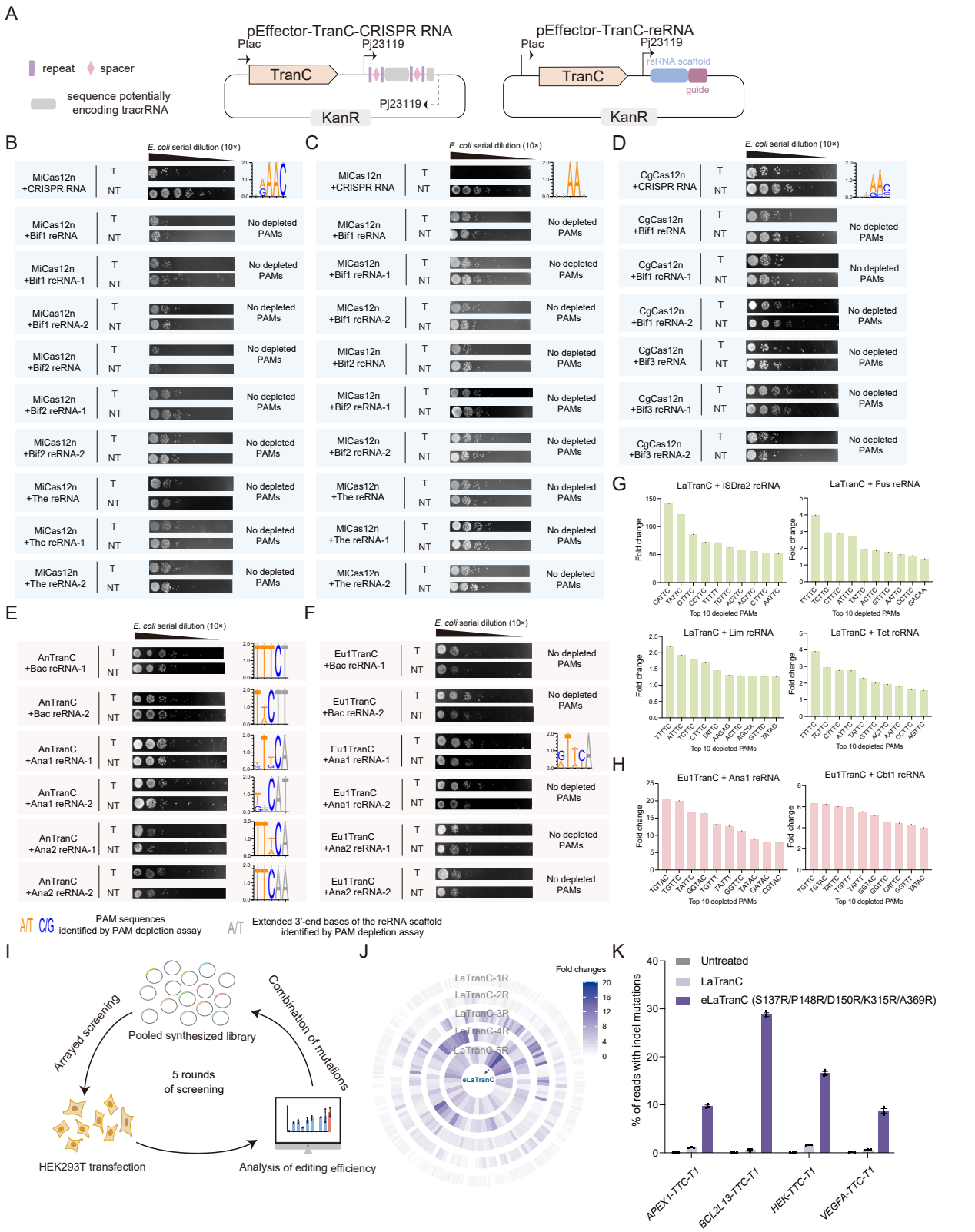


Figure S2. Extended phylogenetic and gene organization analyses of TranCs, related to Figure 2

(A–C) Phylogenetic trees of three representative TranC clades. Members of the TranC group are shown in red, and members of the TnpB sister group are shown in blue. The adjacent panels display annotations of their functional and evolutionary features: associated TnpA transposase types (Y1 or serine TnpA, labeled "Y1" and "Ser"); linkage to Cas1/Cas2/Cas4 proteins; presence of catalytic residues in the RuvC domain (D, E, and D); proximity of CRISPR arrays within 5 kb ("CRISPR array") or located elsewhere on the same contig ("contig array"); preservation of transposon terminal sequences, including the left end (LE) and right end (RE), which are color-scaled by sequence identity ("reservation of LE/RE"); protein copy numbers on the contig under two thresholds ("copy number (70%)" and "copy number (95%)"); and co-occurrence of additional TnpB or Cas12 homologs on the same contig ("contig TnpB" and "contig Cas12").

(D) Comparison of the overlap between protein-coding and guide RNA regions across the three TranC clades. Genomic regions encompassing TnpB loci lacking CRISPR arrays and TranC loci containing CRISPR arrays were analyzed for potential RNA elements (Table S1). This scatterplot illustrates the positions of guide RNAs relative to their corresponding proteins. Each data point represents the position of the 5′ end of either a reRNA or a tracrRNA. The last nucleotide of the stop codon in the TnpB and TranC coding regions is defined as position "0" on the *x* axis for plotting purposes. The bars indicate the average lengths and positions of reRNAs and tracrRNAs across different clades. Statistical significance was determined using the Mann-Whitney *U* test. $p < 0.05$ was considered statistically significant. Data are presented as mean ± SEM. $n = 91$ (clade 3 TnpB), 21 (clade 3 TranC), 131 (clade 8-Cas12n TnpB), 37 (clade 8-Cas12n TranC), 39 (clade 11 TnpB), and 22 (clade 11 TranC).



(legend on next page)

Figure S3. Extended TranC *E. coli* PAM depletion and plasmid interference assays, as well as LaTranC engineering for enhancing editing activity, related to Figure 3

(A) Schematic of the pEffector vector used in *E. coli*. The effector protein and various guide RNAs are heterologously expressed in *E. coli*. "Ptac" and "Pj23119" denote two commonly used *E. coli* promoters. The gray box marks the sequence potentially encoding the tracrRNA, including the native flanking regions of the TranC ORF (see Table S1).

(B–F) Extended results of *E. coli* PAM depletion and plasmid interference assays used to investigate the crRNA- and reRNA-guided dsDNA nuclease activity of TranCs from clade 8-Cas12n (B–D) and clade 11 (E–F). The sequences of reRNA scaffold were designed with 1 or 2 nt extension at the 3′ terminus (termed "reRNA-1" and "reRNA-2"). (E) and (F) present the results for clade 11 reRNAs with +1 and +2 nucleotide scaffold extensions at the 3′ end, providing expanded data corresponding to Figure 3D. Clade 11 served as a positive control for validating the 3′ end of the reRNA scaffold prediction because the same computational method was applied to predict the reRNA termini for both clade 8-Cas12n and clade 11. The results confirmed the accuracy of our 3′ end assignment: only the unextended reRNA scaffolds revealed the genuine PAM sequences (Figure 3D), whereas reRNA scaffolds carrying artificial 3′ extensions displayed PAM sequences shifted by the added nucleotides. A detailed schematic illustrating the rationale and design of the reRNA scaffold extension strategy is provided in Data S1.13. "T" and "NT" represent the pTarget plasmid with or without a target site, respectively. The PAM sequences depleted by more than 10-fold were used to generate a WebLogo. All related guide RNA sequences are listed in Table S2.

(G and H) Detailed fold changes in PAM depletion assays for two TranC proteins guided by different reRNAs. These two panels and Figure 3E expand upon the data presented in Figure 3D (the result of LaTranC + ISDra2 reRNA is also shown in Figure 3E). For instance, LaTranC guided by ISDra2 reRNA (G) and Eu1TranC guided by Ana1 reRNA (H) demonstrated robust PAM depletion, with specific PAMs exhibiting > 10-fold changes compared with the control PAM library. By contrast, LaTranC guided by Fus, Lim, and Tet reRNAs (G) and Eu1TranC guided by Cbt1 reRNA (H) exhibited limited PAM depletion, with no specific PAMs showing > 10-fold depletion. Nevertheless, they revealed conserved PAM motifs consistent with those detected in crRNA-guided activity. For these groups, PAMs depleted by >1.5-fold were used to generate the WebLogo in Figure 3D.

(I) Schematic overview of LaTranC optimization through arginine mutagenesis screening in HEK293T cells.

(J) A heatmap representing the editing efficiencies of LaTranC variants across five rounds of screening. $n = 1$ for the first round, and $n = 3$ for rounds two through five. All experiments were conducted at the endogenous *VEGFA-TTC-T1* site with 5′-TTC PAM in HEK293T cells.

(K) Comparison of the editing efficiencies between wild-type LaTranC and the engineered variant (eLaTranC) at four endogenous target sites with 5′-TTC PAM in HEK293T cells ($n = 3$, mean ± SEM).

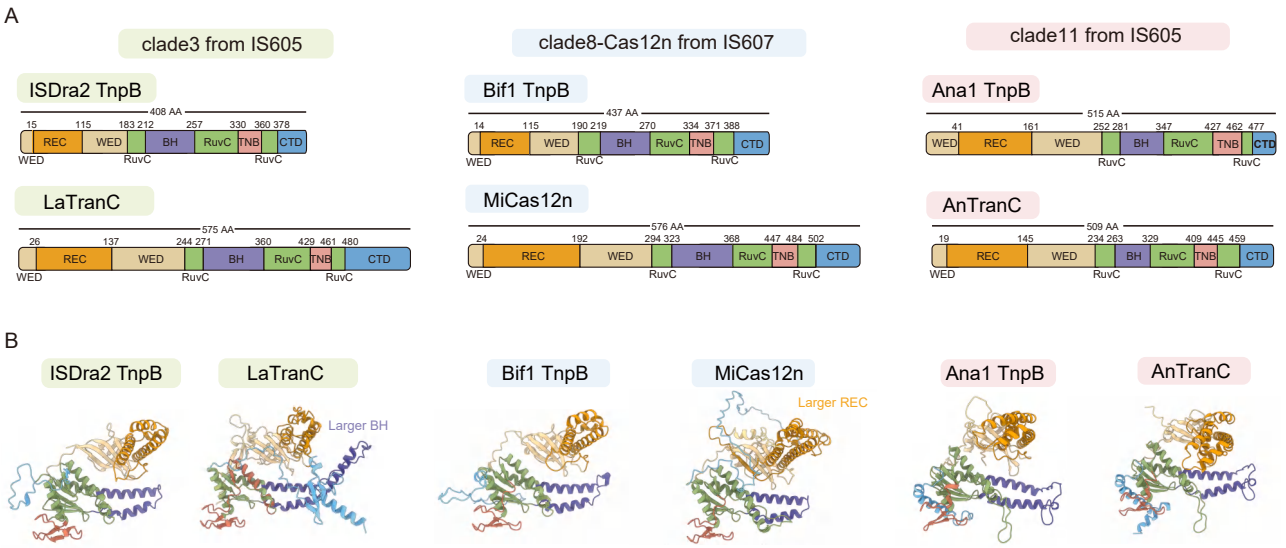


Figure S4. Comparison of the protein structures of three representative TranC clades, related to Figure 4

(A) Domain organizations of TnpB and TranC proteins from three representative TranC clades.

(B) Structural comparison based on AlphaFold models between TnpB and TranC proteins from three representative TranC clades.

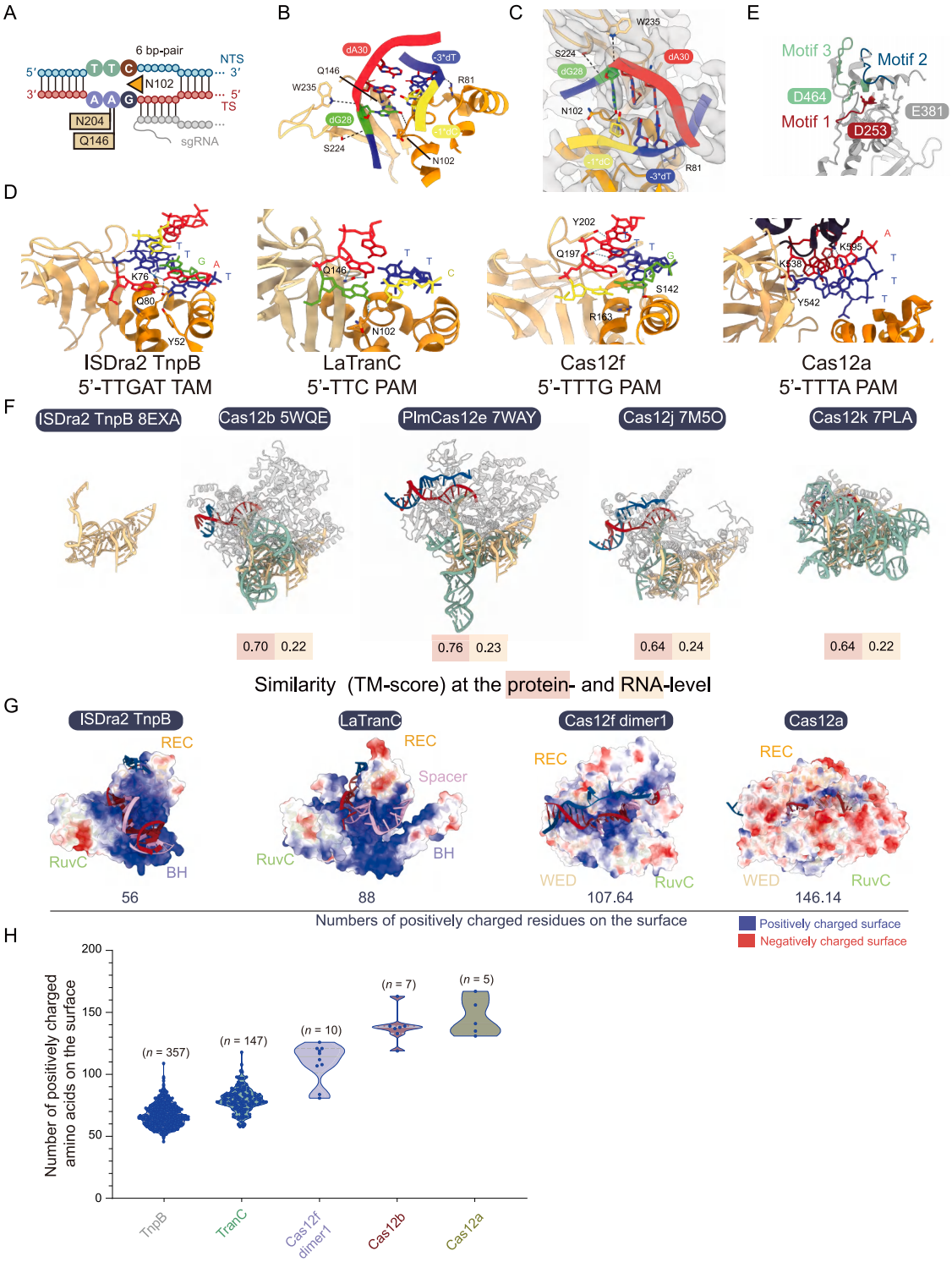


Figure S5. Target DNA recognition features of LaTranC and extended structural comparison data of LaTranC, TnpB, and Cas12 proteins, related to Figure 4

(A–C) Schematic representation (A), atomic model (B), and close-up view of the density (C) of the PAM region and surrounding residues involved in LaTranC-mediated DNA recognition. NTS, non-target strand. TS, target strand.

(D) Structural comparison of TAM recognition by TnpB and PAM recognition by LaTranC and other type V CRISPR systems.

(legend continued on next page)

(E) Localization of three conserved motifs identified in Figure 1B within the LaTranC structure. Catalytic residues (D253, E381, and D464) are labeled. The first and third residues are situated in the first and third motifs, respectively, while the second residue is not located within the three conserved motifs.
(F) Comparative analysis between ISDra2 TnpB and representative Cas12s, expanding upon Figure 4F.
(G) Visualization of the positively charged surface in the cryo-EM structures of TnpB, LaTranC, and Cas12 proteins.
(H) Distribution of positively charged residues on the surface across different protein groups.

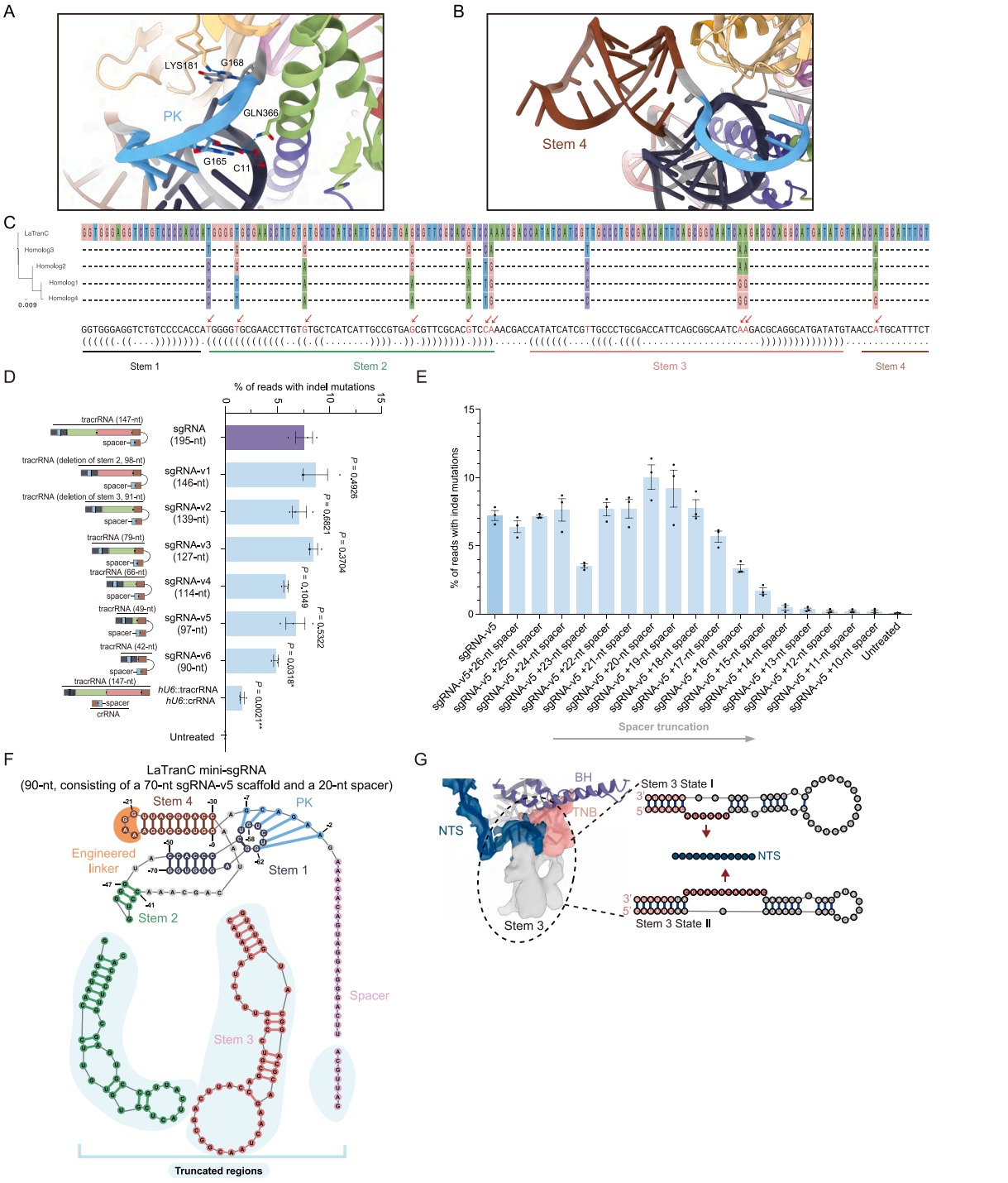


Figure S6. Stem-level characteristics of LaTranC crRNA, related to Figure 5

(A and B) Close-up views of the stem 1-PK region (A) and stem 4 region (B) of the sgRNA in complex with the LaTranC protein. (A) shows base-specific interactions observed between the PK region of the sgRNA and the LaTranC protein.

(C) Conservation analysis of LaTranC tracrRNA homologs. Variable nucleotides are marked in red. The accession numbers of the data sources are as follows: homolog 1 (ERR1190949), homolog 2 (SRR2155482), homolog 3 (SRR5091510), and homolog 4 (ERR1190901).

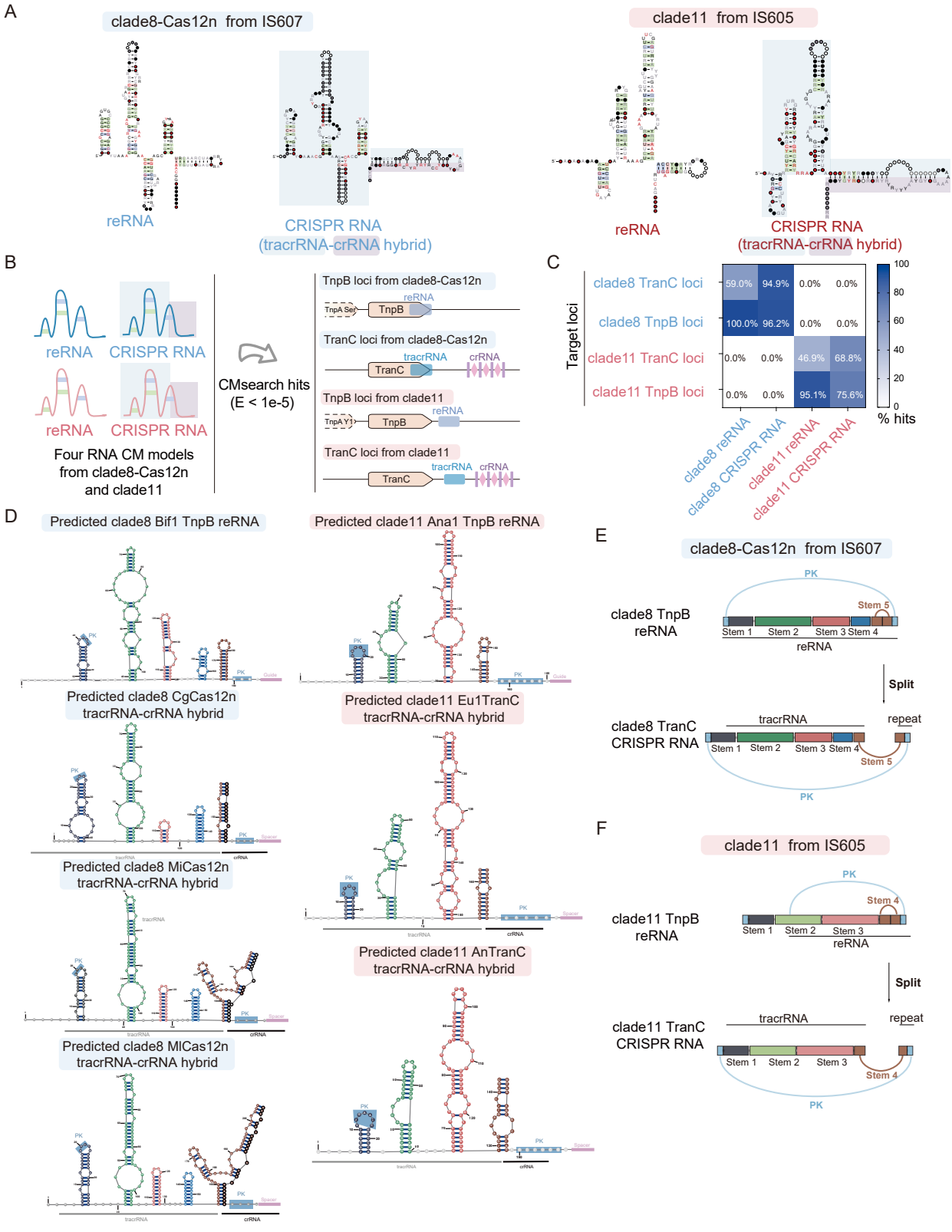
(D and E) Comparison of editing efficiencies of eLaTranC utilizing different sgRNA versions. In (D), six truncated sgRNAs, as well as separate expression of tracrRNA and crRNA using two human U6 promoters (hU6), are colored as in Figure 5A. The editing activities of eLaTranC guided by these seven guide RNAs were compared with that of the wild-type sgRNA. (E) Shows the editing efficiencies of sgRNA-v5 and its truncated-spacer variants. All experiments were conducted at

(legend continued on next page)

the endogenous *VEGFA-TTC-T1* site with 5′-TTC PAM in HEK293T cells. Data are presented as mean ± SEM (n = 3), and statistical significance was evaluated using the Student's t test ($p < 0.05$*, < 0.01**).

(F) Schematic of the LaTranC sgRNA-v5 with 20-nt spacer.

(G) Cryo-EM map showing a flexible non-target strand interaction region in sgRNA stem 3 (left) and two predicted secondary structures (right). NTS, non-target strand. Note that the focal region is largely unresolved, possibly due to the existence of multiple states.



(legend on next page)

Figure S7. The guide RNA structure characteristics of clade 8 and clade 11, related to Figure 5

(A) Structures of covariance-folded reRNAs and tracrRNA-crRNA hybrids from TranC clades 8 and 11. For crRNAs, the tracrRNA and crRNA (excluding the spacer) regions are colored in blue and purple, respectively.

(B and C) Homology analysis using RNA covariance models. (B) Presents a schematic overview of the homology analysis between reRNAs from TnpB loci and crRNAs (tracrRNA-crRNA hybrids) from TranC loci based on RNA covariance models. Genomic regions encompassing TnpB loci lacking CRISPR arrays and TranC loci containing CRISPR arrays were analyzed for potential RNA elements (Table S1). Four RNA covariance models were constructed from sequences of clades 8-Cas12n and 11, and each was applied using the CMsearch program (E value < $1e^{-5}$, Infernal suite) to search TnpB and TranC loci from these two clades. The results of these searches are summarized in (C). Strong intra-clade similarity was observed among reRNAs and crRNAs, as models built from either crRNAs or reRNAs could successfully identify TnpB and TranC loci within the same clade. For example, the clade 8 crRNA model detected 37 out of 39 (94.9%) TranC loci and 126 out of 131 (96.2%) TnpB loci within clade 8-Cas12n but failed to detect any loci from clade 11. Similarly, the clade 11 RNA models showed strong homology within clade 11 but not across clades.

(D) Predicted secondary structures of guide RNAs from representative TranC members and specific sister TnpBs. RNA stems are color-coded within each structure, and PK regions are highlighted in blue boxes, consistently located in stem 1 and at the 3′ terminus of all guide RNA molecules.

(E and F) Schematic models illustrating the functional splitting of guide RNAs in clade 8-Cas12n (E) and clade 11 (F) TranCs.