PNAS

Supporting Information for
Diversity, evolution, and classification of the RNA-guided nucleases TnpB and Cas12

Han Altae-Tran[1,2,3,4,5], Sergey A. Shmakov[6], Kira S. Makarova[6], Yuri I. Wolf[6], Soumya Kannan[1,2,3,4,5], Feng Zhang[1,2,3,4,5,*], Eugene V. Koonin[6,*]

Email: zhang@broadinstitute.org, koonin@ncbi.nlm.nih.gov

This PDF file includes:

Supporting text
Figures S1 to S8
Tables S3, S5
SI References

Other supporting materials for this manuscript include the following:

Tables S1-S2, S4, S6
Additional Files 1-11

All additional files and tables can be found at Zenodo under the DOI:
10.5281/zenodo.8339301

Table of Contents

Supporting Information Text

TnpB and Cas12 curation. For the purpose of comprehensive identification of TnpBs and Cas12s, a representative set of TnpB sequences was obtained using the HHblits with 8 iterations. The sequences were aligned using mafft, and two contiguous regions were extracted from the alignment: 1) conserved N-terminal domain and RuvC-I and 2) RuvC-II, ZF, and RuvC-III. These aligned regions were converted into two TnpB HMM profiles for HMMER and HHSearch (1–3). Additional Cas12 profiles were obtained (4) covering Cas12a, Cas12b, Cas12c, Cas12d, Cas12e, Cas12g, Cas12h, Cas12i, V-U1-5. These profiles were employed to search a custom genomic database that was constructed by combining all publicly available, non-embargoed data from JGI, and all publicly available data from NCBI, and NCBI WGS.

A genomic database was prepared by combining all publicly available, non-embargoed data from JGI, and all publicly available data from NCBI, and NCBI WGS. All potential ORFs were computed as follows. ORFs on all contigs were predicted with TGA, TAG, and TAA stop codons, and ATG start codons (with a minimum size of 55 aa), allowing for alternative start codons GTG (if it produced a minimum ORF size of 200 aa), TTG (if it produced a minimum ORF size of 200 aa) or CTG (if it produced a minimum ORF size of 300 aa). In the case of multiple potential start sites with different start codons for the same potential ORF, all potential start sites were listed in order of increasing corresponding ORF size. A best start site was iteratively selected by traversing the list and selecting a new best start site based on the previous one with the first item in the list selected as the initial best. GTG/TTG start sites were accepted as best if they would make the ORF 20 aa longer than the current best start site, while CTG was accepted only if it would make the protein 40 aa longer than the current best start site. Any ATG start site larger than the current best was automatically selected. ORFs sharing the same stop location and strand (+/-) as an existing protein annotation were discarded in favor of the existing annotation. All ORFs were then searched against the TnpB or Cas12 profiles (described above) using hmmsearch (v3) (3). ORFs were considered candidate ORFs if they contained a profile match with a minimum bit score of 18. The 10kb region around each ORF was then extracted into a corresponding genomic "frame" for further analysis. For each frame, all genes were predicted using Genmark (v2) (5). When the stop site of a predicted gene coincided with the stop site of a candidate ORF, the start site of the candidate ORF was then refined by updating it to be the start site of the coinciding predicted gene from Genmark. Candidate ORFs were considered partial if the ORF start or end was within 200 bp of the contig edge.

All candidate ORFs were then clustered into micro-clusters using MMseqs2 (6) with a minimum sequence identity of 0.9, minimum coverage of 0.85, and the standard cascaded clustering algorithm. Within each cluster, a representative was selected as follows. All sequences within a cluster with a length greater than or equal to the 80^{th} percentile of lengths within the cluster were considered for cluster representative. Sequences were then iterated over in the order found in the MMSeqs2 cluster. Two quantities were tracked for determining representative sequences: 1) ORF's distance to the edge of its contig (m_d) and 2) whether or not the ORF translation contains ambiguities in the form of X amino

acids (has_x) were tracked. A best representative sequence was set initially as Null. Then through iteration, the best representative was replaced with the current sequence if its m_d is greater than or equal to the current best m_d. Throughout iteration if the current best representative sequence does not have an X amino acid, then only sequences without X amino acids will be considered for replacing the current best representative sequence. The sequence set as the best representative at the end of iteration was considered the representative sequence for the micro-cluster.

Micro-clusters were then filtered by removing micro-clusters with representative sequences that are considered partial (as determined above). The passing redundancy reduced micro-cluster representative sequences were then clustered into protein clusters further using MMSeqs2 using a minimum sequence identity of 0.5 and minimum coverage of 0.7. All micro-cluster representatives within each 50% protein cluster were then aligned using MAFFT. These cluster-wise alignments were then searched against the 2 TnpB HHM profiles using HMM-HMM profile comparisons with hhalign (1). Clusters passed the HMM-HMM filtering stage if they contained an hhalign bitscore of 17 or higher to either of the 2 TnpB HHM profiles and also had at least one micro-cluster representative sequence with length of 120 or higher.

Comprehensive TnpB + Cas12 phylogenetic analysis. For comprehensive phylogenetic analysis of TnpB and Cas12, sequences were clustered using MMSeq2, representative sequences from each cluster were realigned using muscle5, and the phylogenetic tree was constructed using IQTree2 (6–8). In detail, this was performed as follows. Representative sequences for the 50% protein clusters for full protein alignment were selected as follows. For each cluster, all 90% representative sequences without X amino acids were selected, and of the remaining sequences, the sequence with the 90^{th} percentile length was used as the representative for the entire protein cluster. The representative sequences for the 50% protein clusters were then aligned using muscle (super5). Sequences were removed from the alignment if they did not have sufficient coverage to 2 of the 3 following regions: 1) RuvC-I, 2) RuvC-II, and 3) RuvC-III. The resulting alignment was then trimmed with trimAl with the gap threshold parameter set at -gt 0.5. The alignment was then manually trimmed further to remove large blocks of low conservation, ensuring the inclusion of the conserved N-terminal domain, RuvC-I, RuvC-II, the conserved zinc finger, and RuvC-III. To reduce the possibility of long branch attraction in the final tree, the following procedure was used. FastTree was then used to create a draft tree, and singleton nodes with excessively long branches with branch lengths ≥1.25 were automatically removed unless their alignment to other TnpBs/Cas12s contained clear homology. This process was repeated with FastTree (9) until excessively long branches were not present. The final alignment was used in 5 independent IQTree2 (8) runs with the parameters: VT+F+G substitution parameters (determined to be optimal for this alignment via IQTree2's modeltest program), 2000 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 best likelihood score out of the 5 runs was selected as the best tree. For the best tree, bootstrap support values were computed with the -bnni option to control for model violations in bootstrap determination. The full tree building process completed with approximately 100k CPU hours. For annotating lengths of each

protein along the tree, the full muscle5 (7) alignment was processed separately. Based on manual inspection, sequences containing large, non-homologus (≥50aa) N-terminal extensions beyond the N-terminal most conserved region of the alignment were trimmed down to the conserved region to reduce artifacts due to incorrectly predicted start sites. Similarly, non-homologous C-terminal extensions beyond the C-terminal most conserved region of the alignment (≥50aa) were also trimmed down to avoid artifacts due to sequencing errors resulting in incorrectly predicted stop sites. Large internal insertions are reflective of TnpBs that are possibly inserted into other genes, but were not trimmed or processed. TreeCluster was used to generate automatic branches (designed by s_id) that were manually refined into minor branches and major clades (10).

CRISPR prediction. CRISPR arrays were predicted on each of the 10kbp windows around each TnpB/Cas12. 4 different CRISPR finders were used: 1) PILERCR with minarray=3, mincons=0.9, maxspacer=80, minspacerratio=0.45, and minid=0.84. 2) CRT with minNR=3, minRL=22, maxRL=60, minSL=16, maxSL=60, and searchWL=9. 3) CRISPRDetect with min_repeats=3, min_repeatation=3, max_repeat_dist=500, and min_repeat_length=22. 4) CRISPRFinder with minDR=22, minSP=16, pm=0.4, px=2.5, s=70, -bDT=1. Predictions from multiple CRISPR finders are redundant, so overlapping CRISPRs were resolved in the following way: all CRISPR arrays that were overlapping were grouped. Within each group, the arrays were sorted lexographically in descending order according to the following criteria (number of repeats, CRISPR program priority, length of the CRISPR bounds, and the CRISPR array start position), where the CRISPR program priority was 1) CRISPRDetect, 2) CRT, 3) PILERCR, and 4) CRISPRFinder (highest priority). The top CRISPR array according to the sorted list was taken from each group to eliminate overlapping CRISPR arrays.

CRISPR spacer search. All predicted CRISPRs within 10kbp of Cas12 effectors were enumerated and given unique IDs. All spacers surrounded by two DRs were then extracted. Spacers below 16bp were filtered out. Spacers were then searched using BLAST against a Combined Prokaryotic Plasmid and Phage database generated from NCBI plasmid and phage sequences with a total database size of 4072513320 and a minimum expect value of 1e-3. Hits within 40 bp of their contig edge were truncated due to inability to resolve self-hits. Target hits were then expanded 300bp upstream and downstream on the target contig. The consensus DR was then blasted against the expanded region, and if any DR hit to the expanded region with a bit score of 40 or more and a distance of 50 or less was found, then the entire hit was discarded due to it being a likely related CRISPR array in another contig. The resulting hits were considered significant hits.

CRISPR spacer hits were then aggregated by Cas12 subtype (or U designation as follows). All spacers were subsequently clustered with MMSeqs2 with --min-seq-id 0.9 -s 7.5 and -c 0.7 to reduce redundancy. For each Cas12 microcluster id, all spacers from loci belonging to the cluster were considered and grouped into their respective spacer clusters. Then for each spacer cluster, the best scoring hit (according to E-value) was taken as the representative target for that specific spacer cluster. Then, for each microcluster id, the number of spacers, number of spacer hits to viruses, and number of

spacer hits to phage were tabulated considering only each representative spacer cluster and its respective representative best scoring target.

Tabulation of locus features. For each row, locus features were tabulated as follows. The following variables were set to false: hascas1, hascas2, hascas4. The locus bounds L and R were set to the protein of interest (POI) start and POI end coordinates in the respective 10kbp window. To reduce false positive associations, the locus bounds were then updated iteratively while incorporating information about the presence of various known associated genes as follows. Repeat until the locus bounds stop changing: 1) iterate through all proteins within 1kbp of the locus bounds [L,R]. If any such protein has a hit to a Cas1 profile with bit score ≥ 18.0, update the locus bounds to include the new protein and set hascas1=True. If any such protein has a hit to a Cas2 profile with bit score ≥ 18.0, update the locus bounds to include the new protein and set hascas2=True. If any such protein has a hit to a Cas4 profile with bit score ≥ 18.0 and does not have a hit to a Cas1 profile with a larger bit score than the Cas4 profile hit's bit score, update the locus bounds to include the new protein and set hascas4=True. 2) For any CRISPR, if the distance is within 10kbp of the locus bounds, set the locus bounds to include the CRISPR start and end. If the CRISPR is upstream of the POI's start position + 200bp, then the locus is considered to have an upstream CRISPR. If there is a CRISPR downstream of the POI's end position – 200bp, then the locus is considered to have a downstream CRISPR. Loci may have both upstream and downstream CRISPRs. The iterations stop once the locus bounds no longer change.

The second stage involved incorporation of mobilome information. For each protein in the 10kbp window, the following was performed: 1) if the protein is within 1000bp of the POI and has a hit to TnpA (Y1), then the locus is considered to have TnpA. 2) if the protein is within 1000bp of the POI and has a hit to SerineRecombinase, then the locus is considered to have a serine recombinase. 3) if the protein is within 1000bp of the POI and has a hit to DDE, then the locus is considered to have a DDE. 2) if the protein is within 1000bp of the POI and has a hit to TyrosineRecombinase, then the locus is considered to have a Tyrosine Recombinase.

Tabulation of Cas12 and TnpB taxonomic features. For all calculations of Cas12 and TnpB taxonomic features, Cas12s were defined as the set of V-A, V-B, V-C, V-D, V-E, V-F, V-G, V-H, V-I, V-K, V-M, V-U2, V-U3, and V-U4. TnpBs were defined to include all other clusters included in the tree provided that their fractions of CRISPR association were below 0.1. To calculate the prevalence of TnpB and Cas12 in all NCBI genomes, we first only analyzed the intersection of the genomes used in this study with the NCBI genomes for which taxonomy data is available. Plasmids were analyzed separately on the basis of the contig name as opposed to genome name because plasmids may appear inside existing prokaryotic genome assemblies. The prevalence of TnpB and Cas12s were computed by calculating the number of entities (archaeal genomes, bacterial genomes, bacteriophage genomes, plasmids) with the effector divided by the total number of entities. For tabulation according to phylum, phyla information was automatically parsed from the NCBI taxonomy lineage per genome, and normalization was performed over all phyla. For each large branch, we also computed the distribution of TnpBs across

Archaea, Bacteria, Viruses, and Plasmids. To perform this for each large branch, we first removed plasmid TnpB contig matches from Bacterial and Archaeal contig matches to avoid double counting. Then for each large branch we counted the number of unique full taxon lineage names of all the matches from that branch in the Archaea, Bacterial, Virus, and Plasmid categories, then normalized by the total number of counts across all categories. Due to the small counts for various Cas12s, we were unable to perform this type of normalization for Cas12s. Instead of normalizing across the number of full taxon lineage names like for Tnpbs, for Cas12s, normalizations were performed over cluster ids (each cluster id that had a match towards one of Archaea, Bacteria, Viruses, Plasmids was counted as 1 towards that category).

CRISPR and Cas1/Cas2/Cas4 associations were tabulated per Cas12 subtype (and the grouped V-U* branches) by calculating the fraction of non-redundant loci (with a distance to contig edge of ≥ 1000) that contain the specified quantity for the association (CRISPR array, Cas1, Cas2, Cas4 gene respectively). For CRISPR length distribution calculations, the median CRISPR DR length from each non-redundant cluster was used (dropping 0s) and then grouped by subtype.

The M-div mobility metric was computed as follows. For each microcluster (90% sequence id cluster), a maximum of 2000 loci were sampled, prioritizing loci with larger TnpB to contig edge distances first. 5000 bp windows were extracted upstream and downstream from the TnpB, keeping the upstream and downstream windows in separate lists. Only windows with a minimum size of 2000 were retained for further analysis. For the upstream and downstream windows separately the following was performed. Megablast was used with a word size of 16 to detect homology between the windows. A matrix of e-values were created from the corresponding pairwise megablast searches. The TnpBs from the passing windows were aligned using MAFFT (2) and used to construct a matrix of pairwise protein sequence identity. If e-values were above 1e-5 in the megablast search, the locus was considered rearranged. For all loci pairs considered to be rearranged relative to one another, the corresponding sequence divergence (1 minus the sequence identity) for the two TnpBs in the pair as determined via MAFFT was considered the “percentage sequence divergence before mobilization.” For each microcluster, the minimum “percentage sequence identity before mobilization” was used as the final M-div metric for the microcluster, with a maximum allowable value of 0.1. Microcluster M-div metrics were aggregated into M-div metrics per cluster (50% cluster) by taking the minimum M-div value of all microclusters in the cluster.

Structure prediction. All structures presented were conducted using AlphaFold2 (multimer version) with the ColabFold notebook (11, 12).

Analysis of TnpB family in complete genomes. 24,757 completely sequenced prokaryotic genomes were obtained from the NCBI GenBank [https://ftp.ncbi.nlm.nih.gov/genomes/] in November 2021. Protein sequences, annotated in these genomes, were analyzed using PSI-BLAST (13) search (e-value cutoff of 0.0001 and effective database size of 10^7) with NCBI CDD profile database (14) and previously described CRISPR-Cas protein profiles (4, 15, 16) as queries.

A TnpB profile from an earlier work (16) was used as a PSI-BLAST query in a search against this database using an e-value cutoff of 0.01. The protein sequences detected in this search were clustered using MMSEQS2 with the similarity cutoff of 0.5, aligned with MUSCLE5 and passed through several rounds of HHSEARCH-HHALIGN cluster merging (15). The cluster alignments thus obtained were used again as queries in a PSI-BLAST search against the same database, followed by manual curation of the lower-scoring hits. This procedure identified 15,519 TnpB sequences from completely sequenced genomes. Additionally, 394 sequences of previously characterized Cas12f and Cas12m proteins and their homologs from the previous studies (15, 17, 18) were included in the set. Metadata for this set is provided in Table S6. We then used the above procedure to align the complete set of 15,913 sequences. We used this alignment as a PSI-BLAST query in a search against all sequences in the same set. The query footprints, corresponding to the alignable conserved core of this set, were extracted from the full-length proteins, and aligned by running the same procedure to convergence. This final alignment was used as an input for FastTree program (9) to construct an approximate Maximum Likelihood phylogenetic tree with the WAG evolutionary model and gamma-distributed site rates (Additional File 8). The same program was used to calculate support values.

RuvC I, II and III sites were identified in the TnpB footprint alignments. The presence of the Asp, Glu and Asp catalytic residues in the respective conserved positions was used to classify the RuvC domain as active. The assignments were corrected manually based on identification of catalytic sites rearrangements as described in the text.
TnpB neighborhood islands contain 10 up- and downstream annotated genes for each instance of tnpB gene in the completely sequenced prokaryotic genomes database (Additional File 9). All proteins were annotated using CDD profiles as described above. CRISPR arrays within the neighborhoods were identified using the minCED tool (19) (https://github.com/ctSkennerton/minced).

Mobility analysis in complete genomes. 15,913 TnpB sequences were clustered with MMSeqs2 (6) with 0.8 and 0.98 sequence similarity thresholds, method “cluster”, 0.333 coverage, 0.1 e-value and cluster-mode 2. These clusters were used to estimate TnpB mobility in the genomes with permissive and strict threshold respectively. TnpB sequences were defined as mobile if the same or another TnpB sequence of the same TnpB cluster (separately for 0.8 and 0.98 thresholds) is present in the same genome. A cluster was classified as mobile if at least one of its members is present in more than one instance in at least one genome. The same approach was used to calculate mobility values for other transposases. These families were identified using CDD assignments as follows:

transposon family	**CDD profile**
IS1 family transposase	NF033558
IS3 family transposase	NF033516
IS4 family transposase (1)	NF033592
IS4 family transposase (2)	NF033591
IS4 family transposase (3)	NF033590

IS5 family transposase (1) NF033578
IS5 family transposase (2) NF033579
IS5 family transposase (3) NF033580
IS5 family transposase (4) NF033581
IS6 family transposase (1) NF033587
IS6 family transposase (2) NF033588
IS21 family transposase NF033546
IS30 family transposase NF033563
IS91 family transposase NF033538
IS110 family transposase NF033542
IS200/IS605 family transposase NF033573
IS256 family transposase NF033543
IS481 family transposase NF033577
IS607 family transposase NF033518
IS630 family transposase NF033545
IS66 family transposase NF033517
IS701 family transposase NF033540
IS982 family transposase NF033520
IS1182 family transposase NF033551
IS1249 family transposase NF033544
IS1380 family transposase NF033539
IS1595 family transposase NF033547
IS1634 family transposase NF033559
ISAs1 family transposase NF033564
ISAzo13 family transposase NF033519
ISH3 family transposase NF033541
ISKra4 family transposase NF033572
ISL3 family transposase NF033550
ISLre2 family transposase NF033529
ISNCY family transposase (1) NF033593
ISNCY family transposase (2) NF033594
Tn3 family transposase NF033527

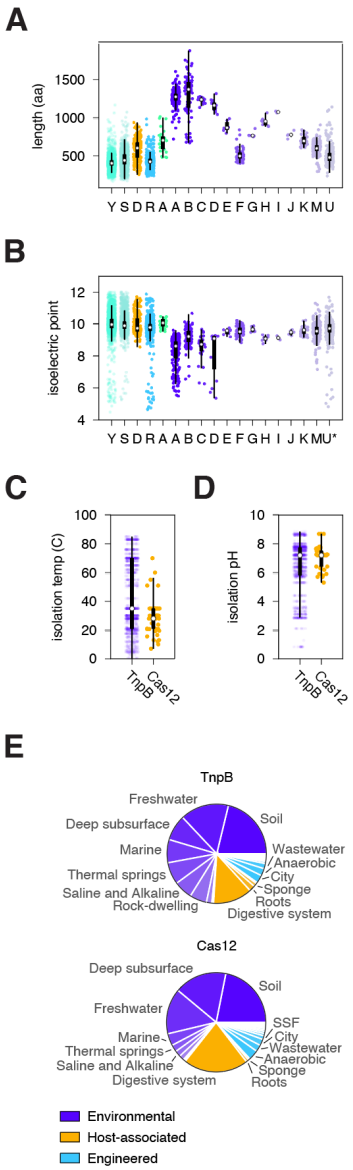

Fig. S1. Additional properties and comparisons of TnpB and Cas12

A) Distribution of trimmed protein lengths for TnpB Y1-associated (Y), Serine Recombianse associated (S), DDE transposon-associated (D), RIIr-5 (R), Asgard-associated (A), Cas12 A,B,…M (A,B,…M), unknown Cas12 types grouped together (U*). Box and whisker plots shown; median (white circle), 25th and 75th percentiles (thick vertical black line), interquartile range (thin vertical black line). **B)** Same as A), except showing isoelectric point distributions. Box and whisker plot as in A).
C) Isolation temperature of various TnpBs and Cas12s from metagenomic samples. Box and whisker plots as in A. Unpaired t-test of significance demonstrating Cas12s are found at lower temperature samples, t-test ***p=1.4e-4. **D)** Isolation pH of various TnpBs vs Cas12s from metagenomic samples. Box and whisker plot as in A). **E)** Distribution of TnpBs and Cas12s across metagenomic samples.

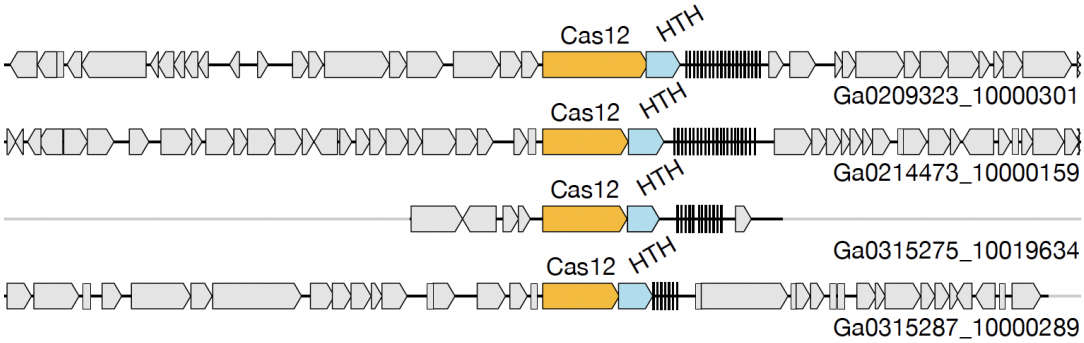

Fig. S2. V-U24 and its association with HTH domains

Genomic accession information is located on the bottom right for each contig.

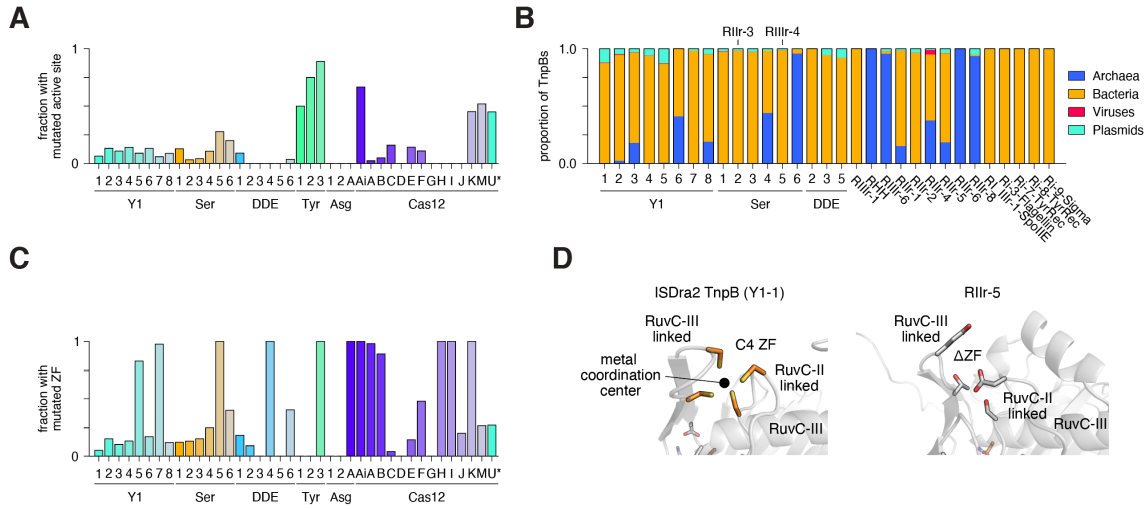

Fig. S3. Catalytic rearrangements and inactivations of TnpBs

A) Fraction of TnpB representative sequences (at 50% sequence identity) with catalytic residues that differ from the most common arrangement in the given TnpB/Cas12 group. **B)** Phylogenetic distribution of TnpB groups and Cas12 subtypes split across Archaea, Bacteria, Viruses, and Plasmids. **D)** Fraction of TnpB representative sequences (at 50% sequence id) without the complete 4 cysteine motif in the TnpB/Cas12 Zinc Finger. **E)** Representative structures of the TnpB Zinc Finger domain.

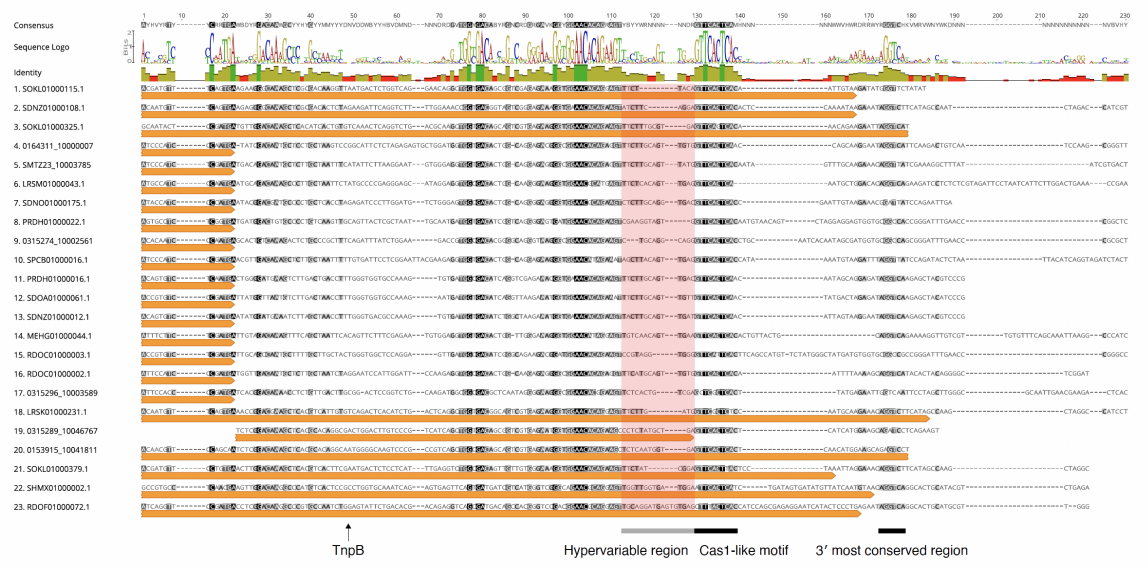

Fig. S4. aCas1-Asgard locus alignment with MAFFT

Alignment of various aCas1-related TnpB loci from Asgard Archaea. Orange gene symbol reflects the TnpB gene bounds at the 3' end. Conservation extends past the 3' of the TnpB genes.

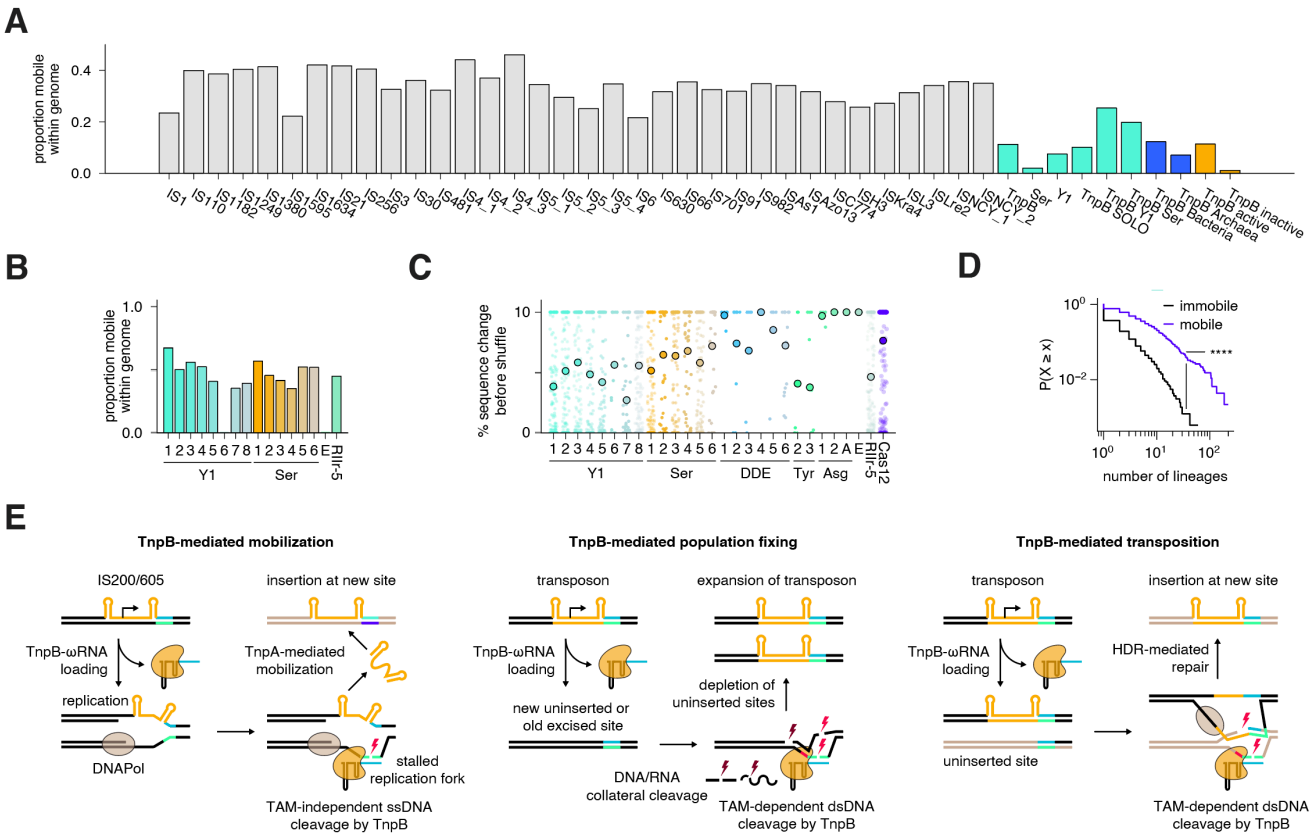

Fig. S5. Mobility analysis of TnpBs

A) Fraction of mobile TnpBs as determined by genomic copy counts in complete genomes. Ser refers to Serine Transposase, while Y1 refers to Y1 transposase. **B)** Proportion of mobile TnpBs as determined by genomic copy counts broken down by TnpB designations. Label E: SpoIIE-TnpB system. **C)** Mobility distribution as determined by percentage sequence divergence (100% - % seq id) before genomic evidence of mobilization. Averages for each designation of TnpB are shown as large, filled circles. **** $p < 1e\text{-}4$. **D)** Number of lineages per branch for mobile vs nonmobile clades of TnpB as determined by genomic copy counts in complete genomes. Kolmogorov-Smirnov test of significance for differences in cumulative density functions. **E)** Putative mechanisms by which TnpB may assist in the transposon life cycle.

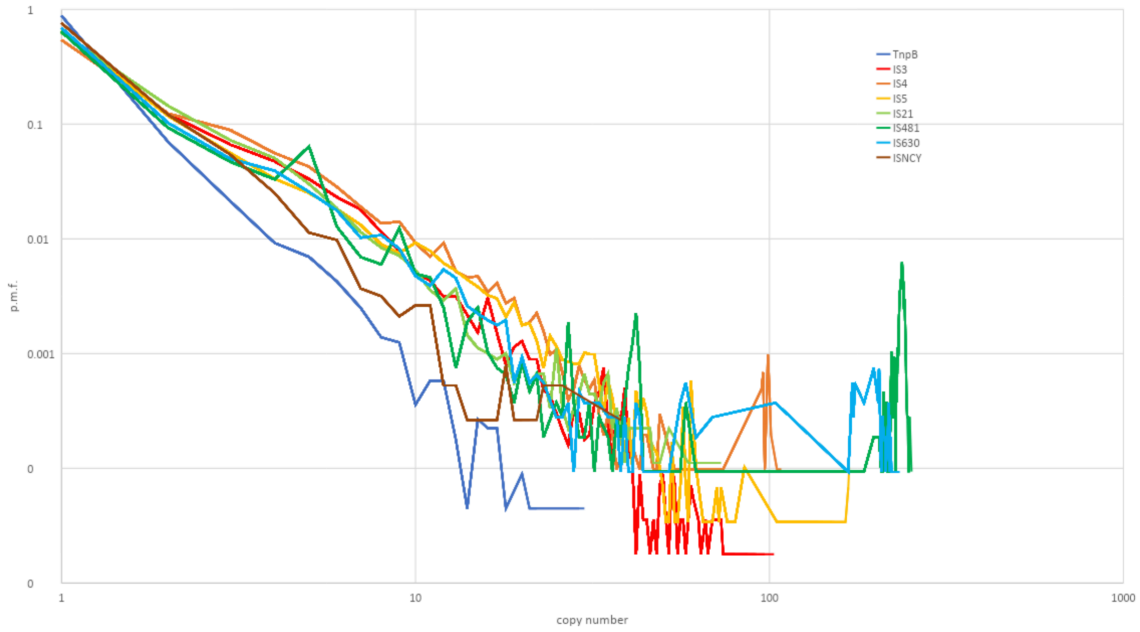

Fig. S6. Copy number distribution of TnpB in complete genomes

Sequences of respective transposases (TnpB, IS3, IS4, IS5) were clustered at 0.98 similarity threshold using MMSEQS2; for complete genome set the number of copies in the same genome assembly was obtained for each 0.98 cluster.

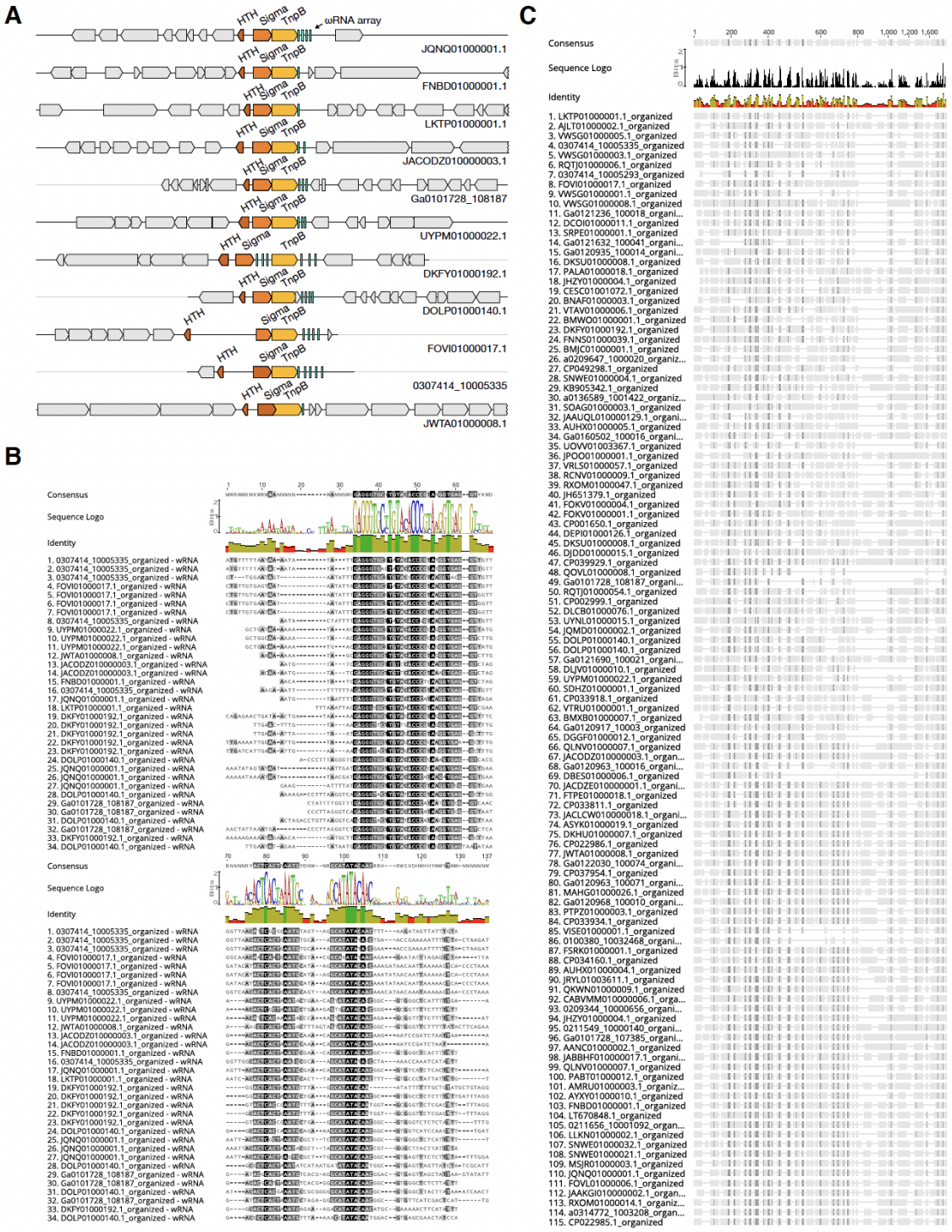

Fig. S7. Sigma Factor-TnpB system (RpoE)

A) Illustration of various wRNA arrays. B) zoomed in MAFFT alignment of representative wRNA array wRNA units (each wRNA is given a row). C) Zoomed out MAFFT alignment of the loci.

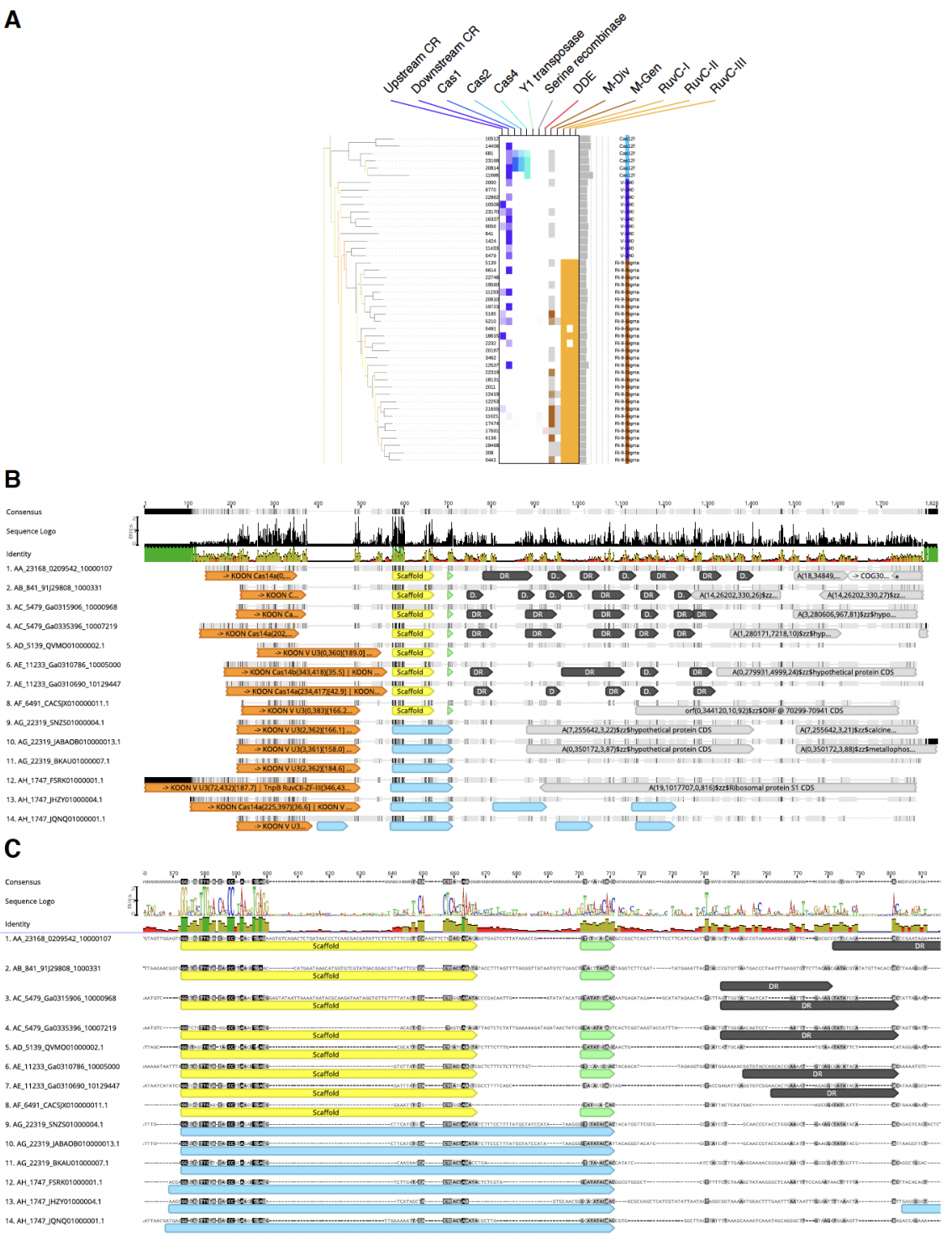

Fig. S8. Sigma Factor/RpoE-TnpB system wRNA evolution

A) Zoomed in visualization of the main tree for the RpoE-TnpB system (Ri-9-Sigma). B). MAFFT alignment of various loci from the system, demonstrating the change in locus architecture through the evolutionary process. C) Zoomed in MAFFT alignment of Ri-9-Sigma loci.

Table S3. Presence and absence of TnpB in complete genomes

Kindom	Phylum*	no. genomes	no. genomes with TnpB	% of genomes coding TnpB
Archaea	Euryarchaeota	274	214	78%
Archaea	Crenarchaeota	97	85	88%
Archaea	Thaumarchaeota	25	9	36%
Archaea	Thermoplasmatota	18	8	44%
Bacteria	Proteobacteria	14096	3601	26%
Bacteria	Firmicutes	5363	1538	29%
Bacteria	Actinobacteria	2424	752	31%
Bacteria	Bacteroidetes	912	90	10%
Bacteria	Tenericutes	442	6	1%
Bacteria	Cyanobacteria	194	117	60%
Bacteria	Chlamydiae	189	8	4%
Bacteria	Spirochaetes	163	11	7%
Bacteria	Verrucomicrobia	120	6	5%
Bacteria	Fusobacteria	77	48	62%
Bacteria	Planctomycetes	66	1	2%
Bacteria	Deinococcus-Thermus	60	49	82%
Bacteria	Chloroflexi	47	6	13%
Bacteria	Thermotogae	41	24	59%
Bacteria	Acidobacteria	26	10	38%
Bacteria	Saccharibacteria	22	0	0%
Bacteria	Aquificae	14	12	86%
Bacteria	Chlorobi	13	0	0%
Bacteria	Nitrospirae	12	3	25%

* NCBI taxonomy as of November 2022

Table S5. Representative non-mobile TnpBs

Accession	Genome	Protein size (aa)	Predicted nuclease activity	Designation in the text	Comment
WP_104011583.1	Streptomyces globisporus TFH56	466	active		Associated with MalK
WP_014177190.1	Streptomyces bingchenggensis BCW-1	676	active		Actinobacteria specific
WP_014484822.1	Bifidobacterium longum	480	active		
WP_024797119.1	Enterococcus faecalis DENG1	356	inactive	DppD	
WP_129256126.1	Enterobacter hormaechei CM18-242-2	310	inactive	FlgL (Flagellin)	
NP_215437.1	Mycobacterium tuberculosis H37Rv	550	active*		Associated with SERINE transposase
WP_012714196.1	Sulfolobus islandicus M.16.4	282	active*	GATase	Associated with glutamine phosphoribosylpyrophosphate amidotransferase and PepP
WP_025566590.1	Deinococcus wulumuqiensis R12	351	active*		
WP_006180995.1	Natrinema pellirubrum DSM 15624	423	active		Halobacteria specific
WP_004216183.1	Natrialba magadii ATCC 43099	397	active		Halobacteria specific
WP_138655450.1	Natrinema pallidum BOL6-1	231	active		Short, RuvC domain only
WP_157823306.1	Bifidobacterium longum NBRC 114370	483	active		Associated with SERINE transposase

WP_076354695.1	Chryseobacterium joostei DSM 16927	364	inactive	RpoE	
QOX64143.1	Clostridiales bacterium MT110	497	inactive	SpoIIE	
WP_218261547.1	Saccharolobus shibatae BEU9	361	active	AbrB/RHH	Potential toxin-antitoxin system
WP_002808807.1	Nitrosococcus oceani ATCC 19707	383	active		
WP_180109962.1	Acinetobacter YH12138	405	active		
WP_179063753.1	Nostoc C052	412	active		
WP_207292034.1	Leclercia 4-9-1-25	396	active		
WP_226845670.1	Bifidobacterium pseudocatenulatum YIT11027	372	active		
WP_208970088.1	Staphylococcus pasteuri FDAARGOS 1152	380	active		
WP_216271562.1	Limosilactobacillus reuteri YLR001	386	active		
WP_000978855.1	Bacillus thuringiensis BMB171	372	active		Associated with SERINE transposase
WP_003400149.1	Clostridium botulinum B1 Okra	375	active		
WP_013785938.1	Alteromonas naphthalenivorans SN2	351	active		
WP_067776435.1	Nostoc NIES-3756	389	active		
WP_083305500.1	Moorea producens PAL-8-15-08-1	287	active		
WP_089369036.1	Pseudoalteromonas nigrifaciens KMM 661	505	active		

WP_212106151.1	Bifidobacterium longum I2-2-3	443	active*		
WP_096396392.1	Halorubrum trapanicum CBA1232	413	active*		
WP_104681718.1	Staphylococcus SB1-57	233	active*		
WP_221881783.1	Mycolicibacterium farcinogenes BKKCU-MFGLA-001	576	active*		
WP_159694767.1	Streptomyces Tu 2975	522	active*		
WP_163745568.1	Mycobacterium lacus JCM 15657	596	active*		
WP_199844217.1	Streptomyces RTd22	544	active*		
WP_148360860.1	Blautia producta DSM 2950	391	inactive		Flanked by phage intergrase and UDP-GlcNAc-inverting 4,6-dehydratase FlaA1
WP_205123683.1	Streptomyces ST1015	546	active		
WP_010867958.1	Pyrococcus abyssi GE5 Orsay	414	active		
WP_013142512.1	Staphylothermus hellenicus DSM 12710	402	active		
WP_014287538.1	Pyrobaculum ferrireducens 1860	404	active		
WP_010902167.1	Halobacterium NRC-34001	384	active*		
WP_115805549.1	Haloferax gibbonsii LR2-5	411	active		
WP_225336361.1	Halosiccatus urmianus IBRC-M 10911	220	active		
WP_128908695.1	Halorubrum BOL3-1	404	active		

Note: The clades were selected manually in the tree for TnpBs from complete genomes based on mobility estimates (see STAR Methods). Active* - TnpB with rearranged RuvC II catalytic site.

Additional file 1. Alignment including all TnpB and Cas12 clusters

Untrimmed, but filtered alignment of all TnpBs and Cas12s used in the main tree in fasta format. Trimmed alignment of all TnpBs and Cas12s used in the tree in pasta format. Alignment generated using Muscle (super5 mode).

Additional file 2. Best IQ-Tree2 tree for TnpB and Cas12 50% clusters

Best tree out of 5 with bootstrap values provided with the --bnni option to reduce bootstrap bias from model misspecification. Tree shown in newick format.

Additional file 3. Downstream alignments of TnpBs

Alignments are performed on the downstream regions from the TreeCluster automatic clusters. Alignments are separated according to whether or not there is a clear RNA-guide boundary.

Additional file 4. Set of all loci used for the main analysis

All grouped loci of TnpBs and Cas12s included in this study. Loci are +/- 10kb window around each TnpB identified by index. The set is organized by cluster id (c_id).

Additional file 5. Expanded view of the tree in pdf format

Contains an additional column, ZF, that is red when an intact zinc finger domain between RuvC-II and RuvC-III is detected. Colormaps are shown on the left, with scaling from 0 to 1000 (maximum value). Maximum value for associations cores corresponds to 100% association. Maximum value for M-div score is 0.1 (10%). For M-gen (mobility metric calculated on complete genomes), white corresponds to an M-gen score of 1 while brown corresponds to an M-Gen score of >1. For RuvC, non-canonical RuvC active site residues are shown as orange lines.

Additional file 6. M-div mobility analysis (sequence divergence until mobilization)

Format: [(c_id, M_div)], where c_id is the 30% cluster id.

Additional file 7. TnpB/Cas12 alignments of all major branches described in the study

Aligned using mafft-einsi

Additional File 8. Tree used for analysis of TnpBs from complete genomes in newick format

Additional File 9. Neighborhood analysis of TnpBs for complete genomes

Additional file 10. Example TnpB loci in genbank format with predicted transposon ends labeled

Additional file 11. DDE-5 examples of loci exhibiting insertion/excision

DDE-5 systems contain an TA motif 8bp upstream from the predicted transposon end.

Table S1. Table of all TnpBs and Cas12s used in the main analysis

Table of all TnpBs and Cas12s used in the main analysis, with a corresponding index to the GenBank files included in the study. The main table columns are the contig name (contig_name), the locus specific index (index), count (which is always one), crispr, cas, Y1 TnpA, DDE, Serine Recombinase, and Phage Integrase counts (1 when found in the vicinity, 0 when not, multiple copies per locus are only counted once), 50% cluster id (c_id), 90% cluster id (mc_id), if the locus is the representative locus for the 90% cluster (mc_rep), the genome accession (various formats for different sources, JGI in particular has _$F_ as a delimiter which separates the two JGI associated ids for the project, the first of which is the portal name, the second of which is the IMG oid). The original protein of interest (TnpB/Cas12) coordinates are provided (orig_POI_coord), as well as the dna sequences and amino acid sequences of the coding sequence of the protein of interest. DR coordinates are also provided when available. The distance to the contig edge (edge_dist) is also provided, as well as distances and relative orientations of associated transposases when available (as calculated up to 10kbp distance, after which no associations are considered). Taxon names and taxon ids are provided according to NCBI for NCBI-linked data. Lastly, the automatic branch identifier is provided (minor_branch_assignment), as well as the formal branch assignment (branch_assignment), and lastly the full clade assignment from the 5 major clades (clade).

Table S2. Table of all representative sequences

Table schema as for Table S1, but only showing representative sequences (one per 50% cluster).

Table S4. Table of all representative sequences from newly identified type V-U*

Table schema as for Table S1, but with reduced columns, as well as information regarding inactivations and catalytic rearrangements (both of which are determined based on an alignment of the proteins in this file using mafft-einsi with BLOSUM62 matrix followed by mafft-linsi on each of the 3 catalytic residue block regions using

BLOSUM30). Only representative sequences from 50% sequence identity clusters are shown.

Table S6. Mobility analysis on complete genomes

SI References

1. M. Steinegger, *et al.*, HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473 (2019).

2. K. Katoh, K. Misawa, K.-I. Kuma, T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).

3. S. R. Eddy, Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).

4. S. A. Shmakov, K. S. Makarova, Y. I. Wolf, K. V. Severinov, E. V. Koonin, Systematic prediction of genes functionally linked to CRISPR-Cas systems by gene neighborhood analysis. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E5307–E5316 (2018).

5. V. Ter-Hovhannisyan, A. Lomsadze, Y. O. Chernoff, M. Borodovsky, Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res.* **18**, 1979–1990 (2008).

6. M. Steinegger, J. Söding, MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).

7. R. C. Edgar, MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).

8. B. Q. Minh, *et al.*, IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

9. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).

10. M. Balaban, N. Moshiri, U. Mai, X. Jia, S. Mirarab, TreeCluster: Clustering biological sequences using phylogenetic trees. *PLoS One* **14**, e0221068 (2019).

11. J. Jumper, *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

12. M. Mirdita, *et al.*, ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).

13. S. F. Altschul, *et al.*, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).

14. A. Marchler-Bauer, *et al.*, CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Res.* **41**, D348-52 (2013).

15. K. S. Makarova, *et al.*, Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. *Nat. Rev. Microbiol.* **18**, 67–83 (2020).

16. S. Shmakov, *et al.*, Diversity and evolution of class 2 CRISPR-Cas systems. *Nat. Rev. Microbiol.* **15**, 169–182 (2017).

17. W. Y. Wu, *et al.*, The miniature CRISPR-Cas12m effector binds DNA to block transcription. *Mol. Cell* **82**, 4487-4502.e7 (2022).

18. L. B. Harrington, *et al.*, Programmed DNA destruction by miniature CRISPR-Cas14 enzymes. *Science* **362**, 839–842 (2018).

19. C. Bland, *et al.*, CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8**, 209 (2007).