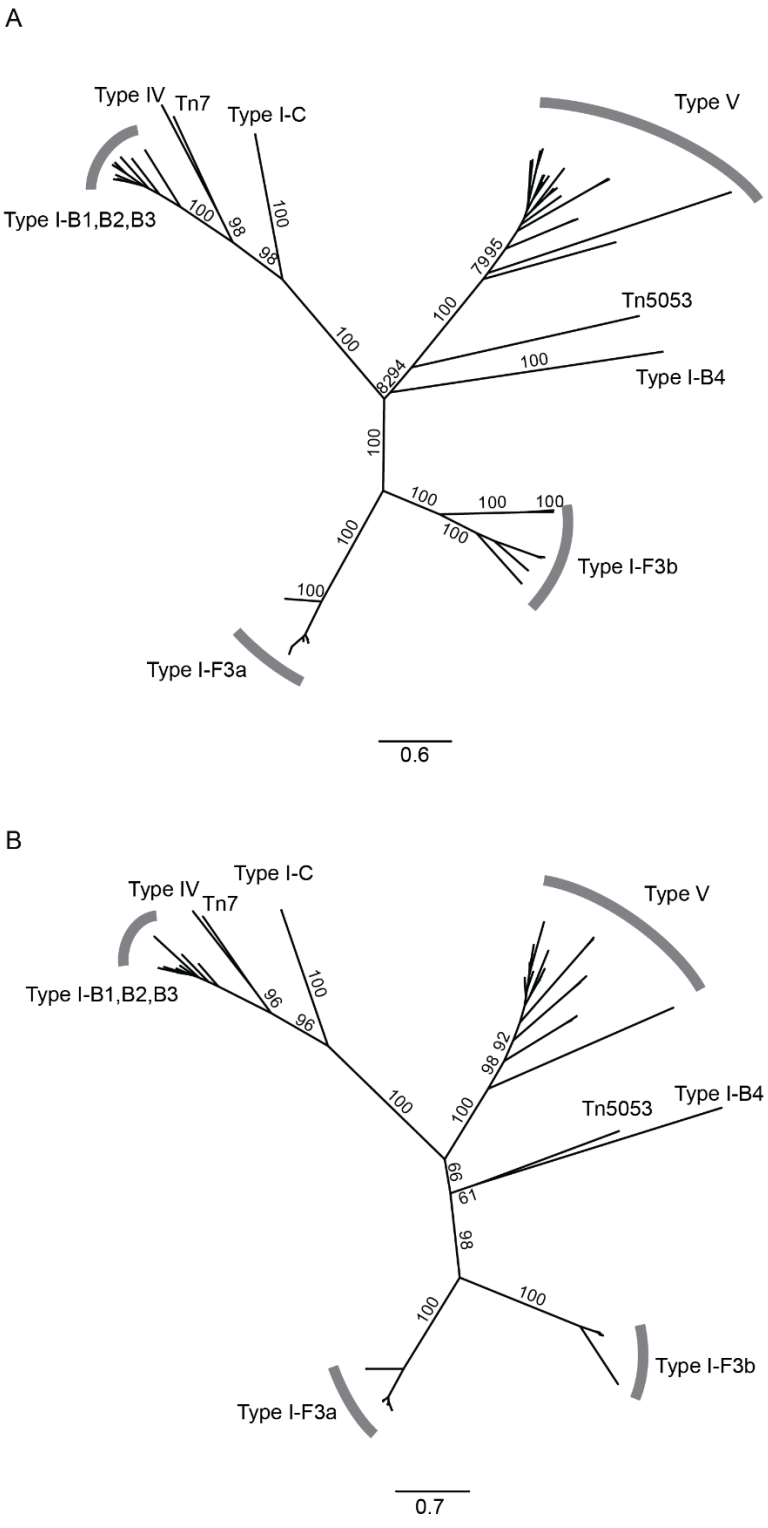
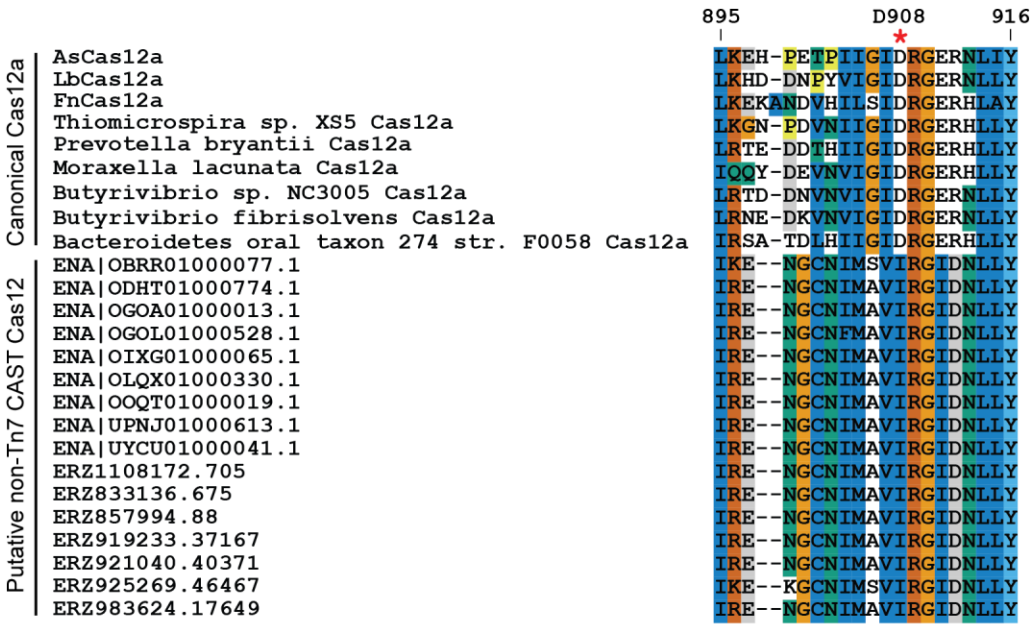
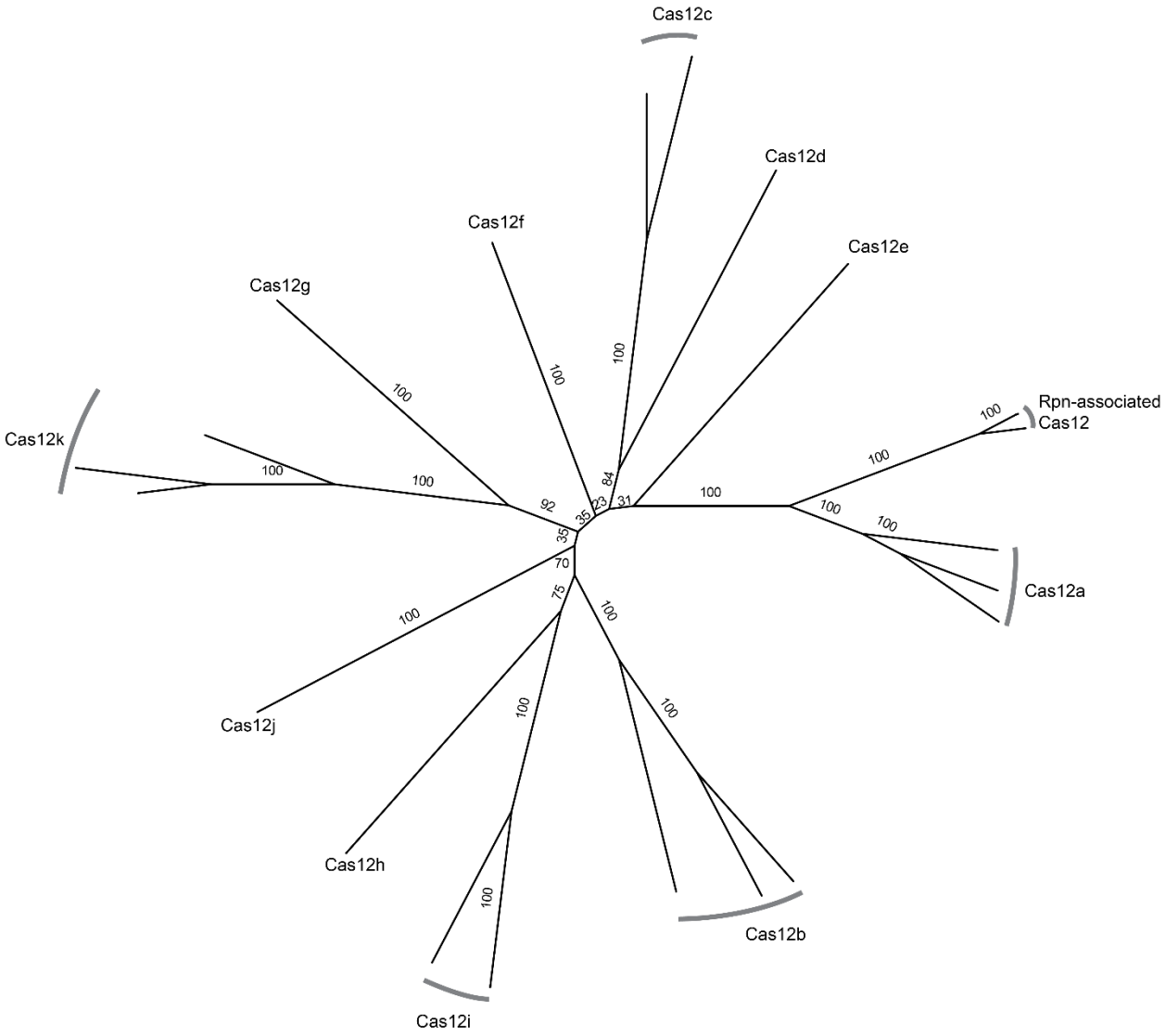


Supplemental Figure 1. Phylogenetic tree of (**A**) TnsB and (**B**) TnsC proteins from each subtype of Tn7 CAST investigated in this work, as well as from Tn7 and Tn5053. Values at branch points are bootstrap support percentages.



Supplemental Figure 2. Multiple sequence alignment of nine canonical Cas12a proteins with putative non-Tn7 CAST Cas12. The position that aligns to D908 in AsCas12a is an isoleucine in all variants. Mutation of this residue alone is sufficient to abolish DNA cleavage in AsCas12a, FnCas12a, and LbCas12a.



Supplemental Figure 3. Phylogenetic analysis of all Cas12 subtypes. The Rpn-associated Cas12 proteins from this work most closely resemble Cas12a but may comprise a distinct subtype. Values at branch points are bootstrap support percentages.

Supplementary Information Text

EXTENDED MATERIALS AND METHODS

BLAST database construction

To find as many systems as possible, we assembled separate databases for Cas proteins, Tn7-family proteins, and non-Tn7 transposases. We also developed databases for common Tn7 attachment sites following a separate effort (1).

We downloaded all available bacterial and archaeal transposase sequences from UniRef50, excluding partial sequences and sequences annotated with the word "zinc" (which tended to be false positives) as well as Tn7-related proteins. All transposases associated with transposons listed in the Transposon Registry (2) were downloaded from NCBI. Finally, 100 transposases associated with each of the major families of insertion sequences were downloaded from NCBI, again excluding partial sequences, and using the 'relevance' sort parameter.

Amino acid sequences for Cas1–Cas12 and Tn7 family proteins (TnsA–TnsE, TniQ) were downloaded from UniRef50 (https://www.uniprot.org/uniref/). Additional Cas12 sequences, representing recently identified variants (e.g., Cas12k), were downloaded from the NCBI protein database (https://www.ncbi.nlm.nih.gov/protein/) and primary literature sources (3–5).

To eliminate redundant sequences, each database was clustered using CD-HIT (6) with a 50% sequence identity threshold and 80% alignment overlap. The clustered datasets were converted to the BLAST database format using makeblastdb (version 2.6.0 of NCBI BLAST+) with the following arguments:

```
makeblastdb
 -in <sequence fasta file>
 -title <database name>
 -out <database name>
 -dbtype prot
 –hash_index
```

The full-length sequences of GuaC (PF00478), RsmJ (PF04378), YciA (PF03061) were downloaded from (http://pfam.xfam.org/). The attachment site SRP-RNA gene (ffs) (RF00169) was downloaded from RFAM (https://rfam.xfam.org/).

To assign putative Cas5-Cas8 proteins to specific Type I CRISPR-Cas subtypes, we manually collected Cas proteins and their assignments from reviews by Koonin and colleagues (4, 7, 8). All Cas protein sequences were converted into BLAST databases using makeblastdb (version 2.6.0) with default parameters.

Database acquisition and contig assembly

NCBI genomes were downloaded using NCBI Genome Downloading Scripts (https://github.com/kblin/ncbi-genome-download) on May 5, 2021, with the command:

```
ncbi-genome-download --formats fasta bacteria
ncbi-genome-download --formats fasta archaea
```

Raw FASTQ files were downloaded from the EMBL-EBI repository (9) of metagenomic sequencing at ftp://ftp.ebi.ac.uk/vol1/ between January and February 2020. For each sample, the quality of the raw data was assessed with FastQC (10) using the command:

```
fastqc tara_reads_*.fastq.gz
```

Low quality reads were trimmed with Sickle (11) using the command:

```
sickle pe
 -f name_reads_R1.fastq.gz
 -r name_reads_R2.fastq.gz
 -t sanger
 -o name_trimmed_R1.fastq
 -p name_trimmed_R2.fastq
 -s /dev/null
```

We then used Megahit (12) to assemble the trimmed data with:

```
megahit
 -1 name_trimmed_R1.fastq
 -2 name_trimmed_R2.fastq
 -o name_assembly
```

Identification of inverted repeats and target site duplications

To identify inverted repeats, we used Generic Repeat Finder (commit hash: 35b1c4d6b3f6182df02315b98851cd2a30bd6201) (13) with default parameters except as follows:

-c: 0

-s: 15

--min_tr: 15

--min_space <operon length>

--max_space <buffered operon length>

where <operon length> is the length of the putative operon and <buffered operon length> is the length of the putative operon, plus up to 1000 bp to allow a 500 bp buffer on either side of the operon. This detected inverted repeats that were at least 15 bp long. In cases where one inverted repeat fell within the bounds of the putative operon, it was discarded.

De-duplication of putative operons

Approximately 57% of the metagenomic systems that passed our initial filter were nearly identical at the nucleotide sequence level. However, exact nucleotide comparisons were too slow to de-duplicate this large dataset. Instead, we considered two systems to be identical if they met the following properties: (1) they had the same protein-coding genes and CRISPR arrays in the same order; (2) the genes had the same relative distances to each other; and (3) the translated sequences of all proteins were identical. This de-duplication was applied to all systems before the downstream analysis.

Homing spacer identification

Spacer sequences that were identified with PILER-CR were pairwise aligned with the contig sequence that contained them, using the Smith-Waterman local alignment function from the parasail library (14), with gap open and gap extension penalties of 8, and using the NUC44 substitution matrix. Spacers with at least 80% homology to a location in the contig were classified as homing.

For Type V systems, we augmented the CRISPR array search with minCED 0.4.2 (15) after noticing transposons that were otherwise intact but seemingly lacked CRISPR arrays. The region between *cas12k* and 200 bp after the end of the nearest CRISPR array was used to search for spacers (both atypical and canonical). Targets were searched for in the 500 bp region immediately downstream of the spacer search region, using the method described in the previous paragraph. For Type V systems with multiple *cas12k* genes, each spacer region was aligned to each target region to discover systems where multiple transposons had inserted at the same attachment site.

Phylogenetic analysis

Alignments of protein sequences were constructed with MAFFT, version v7.310 (16). Phylogenetic analysis was performed on the aligned sequences using the IQ-TREE, version 1.6.12 (17), with automatic model selection. Models used were as follows: Figure 3B: JTT+F+R3, Figure 4B Cas6: PMB+G4, Figure 4B Cas7: PMB+G4, Figure 4C: PMB+G4, Supplemental Figure 1 TnsB: LG+R5, Supplemental Figure 1 TnsC: LG+G4. Trees were visualized using the Figtree program version 1.4.4.

Classification of nuclease-dead systems

To identify catalytically inactive Class 2 nucleases, we aligned each nuclease to a reference protein with MAFFT (version v7.310, with the FFT-NS-2 strategy for Cas9 and Cas12). Cas9 homologs were aligned to SpCas9 (UniProtKB Q99ZW2.1, residues D10 and H840) and Cas12 homologs to AsCas12a (UniProtKB U2UMQ6, residues D908 and E993). Mutations of D/E to anything other than D/E, or H/R to anything other than H/R/K were considered nuclease-dead. To test this strategy, we aligned 279 Cas12k proteins from NCBI as well as LbCas12a and FnCas12a—two nuclease-active Cas12a proteins. All proteins in this test case were correctly categorized via this approach.

SI References

1. M. T. Petassi, S.-C. Hsieh, J. E. Peters, Guide RNA Categorization Enables Target Site Choice in Tn7-CRISPR-Cas Transposons. *Cell* **183**, 1757-1771.e18 (2020).

2. S. Tansirichaiya, Md. A. Rahman, A. P. Roberts, The Transposon Registry. *Mob. DNA* **10**, 40 (2019).

3. P. Pausch, *et al.*, CRISPR-CasΦ from huge phages is a hypercompact genome editor. *Science* **369**, 333–337 (2020).

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

5. W. X. Yan, *et al.*, Functionally diverse type V CRISPR-Cas systems. *Science* **363**, 88–91 (2019).

6. L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

7. D. Burstein, *et al.*, New CRISPR-Cas systems from uncultivated microbes. *Nature* **542**, 237–241 (2017).

8. K. S. Makarova, *et al.*, An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol.* **13**, 722–736 (2015).

9. A. L. Mitchell, *et al.*, MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.* **48**, D570–D578 (2020).

10. FastQC, FastQC: A quality control tool for high throughput sequence data (2015).

11. N. A. Joshi, J. N. Fass, *Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33)[Software]* (2011).

12. D. Li, C.-M. Liu, R. Luo, K. Sadakane, T.-W. Lam, MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinforma. Oxf. Engl.* **31**, 1674–1676 (2015).

13. J. Shi, C. Liang, Generic Repeat Finder: A High-Sensitivity Tool for Genome-Wide De Novo Repeat Detection1. *Plant Physiol.* **180**, 1803–1815 (2019).

14. J. Daily, Parasail: SIMD C library for global, semi-global, and local pairwise sequence alignments. *BMC Bioinformatics* **17**, 81 (2016).

15. C. Skennerton, *MinCED* (2019).

16. K. Katoh, D. M. Standley, MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

17. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).