Metagenomic discovery of CRISPR-associated transposons

James R. Rybarski[a,1], Kuang Hu[a,b,1], Alexis M. Hill[b,1], Claus O. Wilke[b,2], and Ilya J. Finkelstein[a,c,2]

[a]Department of Molecular Biosciences, The University of Texas at Austin, Austin, TX 78712; [b]Department of Integrative Biology, The University of Texas at Austin, Austin, TX 78712; and [c]Center for Systems and Synthetic Biology, The University of Texas at Austin, Austin, TX 78712

Edited by Jennifer A. Doudna, University of California, Berkeley, CA, and approved October 28, 2021 (received for review July 2, 2021)

CRISPR-associated Tn7 transposons (CASTs) co-opt *cas* genes for RNA-guided transposition. CASTs are exceedingly rare in genomic databases; recent surveys have reported Tn7-like transposons that co-opt Type I-F, I-B, and V-K CRISPR effectors. Here, we expand the diversity of reported CAST systems via a bioinformatic search of metagenomic databases. We discover architectures for all known CASTs, including arrangements of the Cascade effectors, target homing modalities, and minimal V-K systems. We also describe families of CASTs that have co-opted the Type I-C and Type IV CRISPR-Cas systems. Our search for non-Tn7 CASTs identifies putative candidates that include a nuclease dead Cas12. These systems shed light on how CRISPR systems have coevolved with transposases and expand the programmable gene-editing toolkit.

gene editing | bioinformatics | CAST | transposition | CRISPR RNA

CRISPR-associated transposons (CASTs) are transposons that have delegated their insertion site selection to a nuclease-deficient CRISPR-Cas system. All currently known CASTs derive from Tn7-like transposons and retain the core transposition genes *tnsB* and *tnsC* but dispense with *tnsE*, and often *tnsD*, which mediate target selection (1, 2). Tn7 transposons site specifically insert themselves at a single chromosomal locus (the attachment or *att* site) via the TnsD/TniQ family of DNA-binding proteins while TnsE promotes horizontal gene transfer onto mobile genetic elements. In contrast, Class 1 CASTs replace TnsD and TnsE with a CRISPR RNA (crRNA)–guided TniQ-Cascade effector complex (3–6). These CASTs can use the TniQ-Cascade complexes for both vertical and horizontal gene transfer (5). One notable exception is a family of Type I-B CASTs that retains TnsD for vertical transmission but co-opts TniQ-Cascade for horizontal transmission (7). Similarly, Class 2 CASTs use the Cas12k effector to transpose to the *att* sites or to mobile genetic elements (8, 9). CASTs also dispense with the spacer acquisition and DNA interference genes found in traditional CRISPR-Cas operons (2). In short, these systems have merged the core transposition activities with crRNA-guided DNA targeting.

CASTs are exceedingly rare; only three subfamilies of Tn7-associated CASTs have been reported bioinformatically and experimentally (2, 5, 7, 9, 10). These studies have identified that many, but not all, CASTs encode a homing spacer flanked by atypical (privileged) direct repeats (11). However, the prevalence of such atypical repeats, the diversity of homing strategies, and the molecular mechanisms of why CASTs have evolved these repeats remain unresolved. Moreover, all CASTs that have been identified to date have a minimal CRISPR array with as few as two spacers. These systems are also missing the Cas1–Cas2 adaptation machinery, raising the question of how CASTs target other mobile genetic elements for horizontal gene transfer. Another open question is whether non-Tn7 transposons have adapted CRISPR-Cas systems to mobilize their genetic information.

CASTs are also a promising tool for inserting DNA into diverse cells. CASTs have already been used to simultaneously insert large cargos at multiple genomic loci (12–14), build mutant libraries in vivo (15), and edit the genomes of uncultivated members of a bacterial community (16). The characterization of as-yet-undiscovered systems with diverse capabilities may spur additional applications for CASTs in engineering both prokaryotic and eukaryotic cells as had occurred for CRISPR-Cas nucleases. For example, although Cas9 and Cas12a both seemingly catalyze the same reaction—crRNA-guided cleavage of a double-stranded DNA—these enzymes have been harnessed for different biotechnological applications owing to their differing nuclease domain architectures. Cas9 nickases can be readily created by inactivating either the HNH or RuvC nuclease domain, leading to applications such as prime editing (17, 18). Cas12a, in contrast, can cleave nonspecific single-stranded DNA after binding its specific target sequence (19). This activity has been harnessed for a suite of nucleic acid detection technologies (20). We reasoned that an expanded catalog of CASTs may shed light on the many unresolved questions regarding their biological mechanisms and future biotechnological applications.

Here, we have systematically surveyed CASTs across metagenomic databases using a custom-built computational pipeline that identifies both Tn7 and non-Tn7 CASTs. Using this pipeline, we have identified unique architectures for Type I-B, I-F, and V CASTs. Type I-F CASTs show the greatest diversity in *cas* genes, including *tniQ-cas8/5* fusions, split *cas7*s, and even split *cas5* genes. Some I-F CASTs likely assemble a Cascade around a short crRNA for homing from a noncanonical spacer.

Significance

CRISPR-Cas systems confer bacteria and archaea with adaptive immunity against mobile genetic elements. These systems also participate in other cellular processes. For example, CRISPR-associated Tn7 transposons (CASTs) have co-opted nuclease-inactive CRISPR effector proteins to guide their transposition. We bioinformatically survey metagenomic databases to uncover CASTs, including systems with new architectures and ones that use distinct CRISPR subtypes. We also describe a putative non-Tn7 CAST that co-opts Cas12. Our findings propose mechanisms for vertical and horizontal CAST targeting and shed light on how CASTs have coevolved with CRISPR-Cas systems.

Author contributions: J.R.R., K.H., A.M.H., C.O.W., and I.J.F. designed research; J.R.R., K.H., and A.M.H. performed research; J.R.R. and A.M.H. contributed new reagents/analytic tools; J.R.R., K.H., and A.M.H. analyzed data; and J.R.R., K.H., A.M.H., C.O.W., and I.J.F. wrote the paper.

Competing interest statement: The authors are coinventors on patent applications filed based on this work.

This article is a PNAS Direct Submission.

[1]J.R.R., K.H., and A.M.H. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: wilke@austin.utexas.edu or ilya@finkelsteinlab.org.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2112279118/-/DCSupplemental.

Published November 29, 2021.

Type I-B CASTs frequently encode two *tniQ*/*tnsD* homologs, one of which is used for homing via a crRNA-independent mechanism (7). Remarkably, we have also found I-B systems that encode two *tniQ* homologs and a homing crRNA, suggesting additional unexplored targeting mechanisms. In addition, we have observed Type I-C and Type IV family Tn7-like CASTs with unique gene architectures. Both of these subfamilies lack canonical CRISPR arrays, suggesting that CASTs use distal CRISPR arrays, perhaps from active CRISPR-Cas systems, for horizontal gene transfer. We have identified multiple self-insertions and gene loss in Type V systems, indicating that target immunity—a mechanism that prevents transposons from multiple self-insertions at an attachment site—is frequently weakened. Finally, we have found a set of Cas12-associated recombination-promoting nuclease/transposase (Rpn) family transposases that may participate in crRNA-guided horizontal gene transfer. We anticipate that these findings will shed additional light on how CASTs have co-opted CRISPR-Cas systems and further expand the CRISPR gene-editing toolbox.

Results

A Bioinformatic CAST Discovery Pipeline. We developed a bioinformatics pipeline that first searches metagenomic contigs for transposases using protein BLAST (BLASTP) and a curated transposase database with a permissive e-value threshold of 10^{-3} (21) (Fig. 1*A*). All possible open reading frames (ORFs) in a 25-kilobase pair (kbp) neighborhood up- and downstream of each putative transposase are translated and searched with BLASTP using a second curated database of all Cas proteins. Contigs without *cas* genes are not analyzed further. The remaining contigs contain both a transposase and at least one *cas* gene. We identify CRISPR arrays in these contigs using a modified version of PILER-CR that can locate arrays with as few as two repeats (22). A final round of protein annotation searches for accessory transposase subunits (i.e., TnsC/D/E for the Tn7 family) and genetic elements that are near common attachment sites (23, 24). Finally, we filter the contigs by constraints (detailed in later sections) that are designed to isolate novel CASTs.

To test this pipeline, we searched for all previously discovered CASTs in the National Center for Biotechnology Information (NCBI) repository of bacterial genomes (2, 25, 26). We downloaded 951,491 partial and complete bacterial genomes from the NCBI FTP server on May 5, 2021. Using these genomes as input, we identified regions with at least one *cas* gene within an ∼25-kbp neighborhood of a transposase. We filtered for contigs that encoded either a Type I or Type V effector and at least one member of the TniQ/TnsD family of Tn7-associated proteins. As expected, we reidentified previously published systems (2, 25) along with previously unannotated Type V systems. These results confirm that the bioinformatics pipeline is sufficiently sensitive to discover these rare CRISPR-transposon systems in large genomic databases.

Next, we searched the repository of metagenomic sequencing reads from the European Molecular Biology Laboratory (EMBL)–European Bioinformatics Institute for novel CASTs (27). This repository is the largest collection of sequenced DNA from diverse microbiomes, aquatic, manmade, and soil environments. We downloaded >1 petabyte of non-16S ribosomal RNA reads (Fig. 1*B* and see *Materials and Methods*). These reads were assembled into ∼10 billion high-quality contigs (∼30 terabytes). Contigs were annotated for transposases and *cas* genes. Approximately 54 million contigs (∼150 gigabytes) met the twin criteria of having at least one transposase and one *cas* gene. After the deduplication of nearly identical contigs, we searched for putative CAST systems (Fig. 1*B*). Using Class 1 Tn7-associated CASTs as an example, we filtered for contigs that included at least one *cas* gene, *tnsD*/*tniQ*, and had at least one of the core Tn7 genes: *tnsA*, *tnsB*, or *tnsC* (∼1.8M contigs). These systems were additionally filtered by selecting only systems in which two of the Tn7 genes were less than 1,500 bp apart and in which three of the core Class 1 *cas* genes (*cas5*, *cas6*, *cas7*, and *cas8*) were also less than 1,500 bp apart (1,167 systems). Of these, 1,093 systems were identified as being Type I-F CASTs. The remaining high-confidence systems included Type I-B, I-C, IV, and V systems (Fig. 1*C*). To increase the confidence of our annotations and to annotate unknown ORFs, we re-BLASTed all possible reading frames against the UniProtKB/TrEMBL database of high-quality computationally annotated protein variants (28). We assigned high-confidence systems to specific subtypes by re-BLASTing the *cas* genes against a database of subtype-specific effector proteins and by manually reviewing the operon architecture. After removing nearly redundant systems, we found 1,476 high-confidence CRISPR-Tn7 CASTs. Notably, we detected founding members of the Type I-C and IV CASTs in the metagenomic but not NCBI database. All of these systems were missing the interference (*cas3*) and adaptation (*cas1*/*2*) genes in agreement with the “guns for hire” hypothesis for how CRISPR-Cas systems have been co-opted for diverse cellular functions (29).

Diversity of Type I-F CASTs. We identified 1,093 nonredundant I-F subsystems with the prototypical gene arrangement of *tnsA-tnsB-tnsC* separated from *tniQ-cas8*/*cas5-cas7-cas6* by a large cargo region (Fig. 2*A*). Tn7 cargo genes are unrelated to the transposition mechanism and often include antibiotic resistance genes (30). Type I-F3a systems, defined as using the conserved genes *guaC* or *yciA* as their *att* site, comprise ∼61% of all I-F CASTs (Fig. 2*B*). I-F3b systems, which use the *rsmJ* or *ffs att* site, comprise ∼34% of I-F CASTs (11). The remaining 5% of I-F systems form a distinct group, termed I-F3c, with a unique *att* site and homing mechanism.

The most common gene arrangement in our dataset for all three subtypes encodes the *tnsA-C* genes in one operon and the *tniQ* and *cas* genes in a second operon that is adjacent to the CRISPR repeats. A large cargo spanning ∼10 to 20 kbp either separates these operons or is present downstream of the *cas* genes. The stoichiometry of the Cascade effector has been previously reported to be $(\text{Cas6})_{1:}$ $(\text{Cas7})_6$: $(\text{Cas8/Cas5})_1$: $\text{crRNA}_{1:}$ TniQ_2 based on cryogenic electron microscopy of Type I-F CAST complexes (3, 4, 6, 31). In these studies, TniQ interacts with Cas7 and is structurally distant from Cas8/Cas5. However, in two of our I-F3a systems, TniQ is expressed as an N-terminal fusion with Cas8/Cas5. Four distinct I-F3a systems also have a split Cas7, and in one system, both the Cas5 and Cas7 proteins are split into two distinct polypeptides.

All I-F3 systems that we identified appear to use a crRNA-guided homing mechanism that directs Cascade near the *att* site (11). The homing crRNA is either in the leader distal position of the CRISPR array or 80 to 85 nt away from the CRISPR array as reported previously (11). These homing crRNAs are flanked by an atypical direct repeat that has several substitutions relative to the direct repeats within the CRISPR array. I-F3c CASTs attach upstream of a protein of unknown function that encodes seven putative transmembrane regions (see *Materials and Methods*). This *att* site has not been previously reported for any Tn7 family transposon. To determine how Type I-F3c systems use crRNA-guided transposition, we aligned the region around the CRISPR array with the sequence 500 bp upstream of *tnsA*. This identified a short 20-bp sequence immediately after the final canonical CRISPR repeat that matched the region ∼64 bp upstream of the transposon end. This short, atypical spacer is followed by an atypical repeat (Fig. 2*C*) akin to the I-F3a and I-F3b systems.

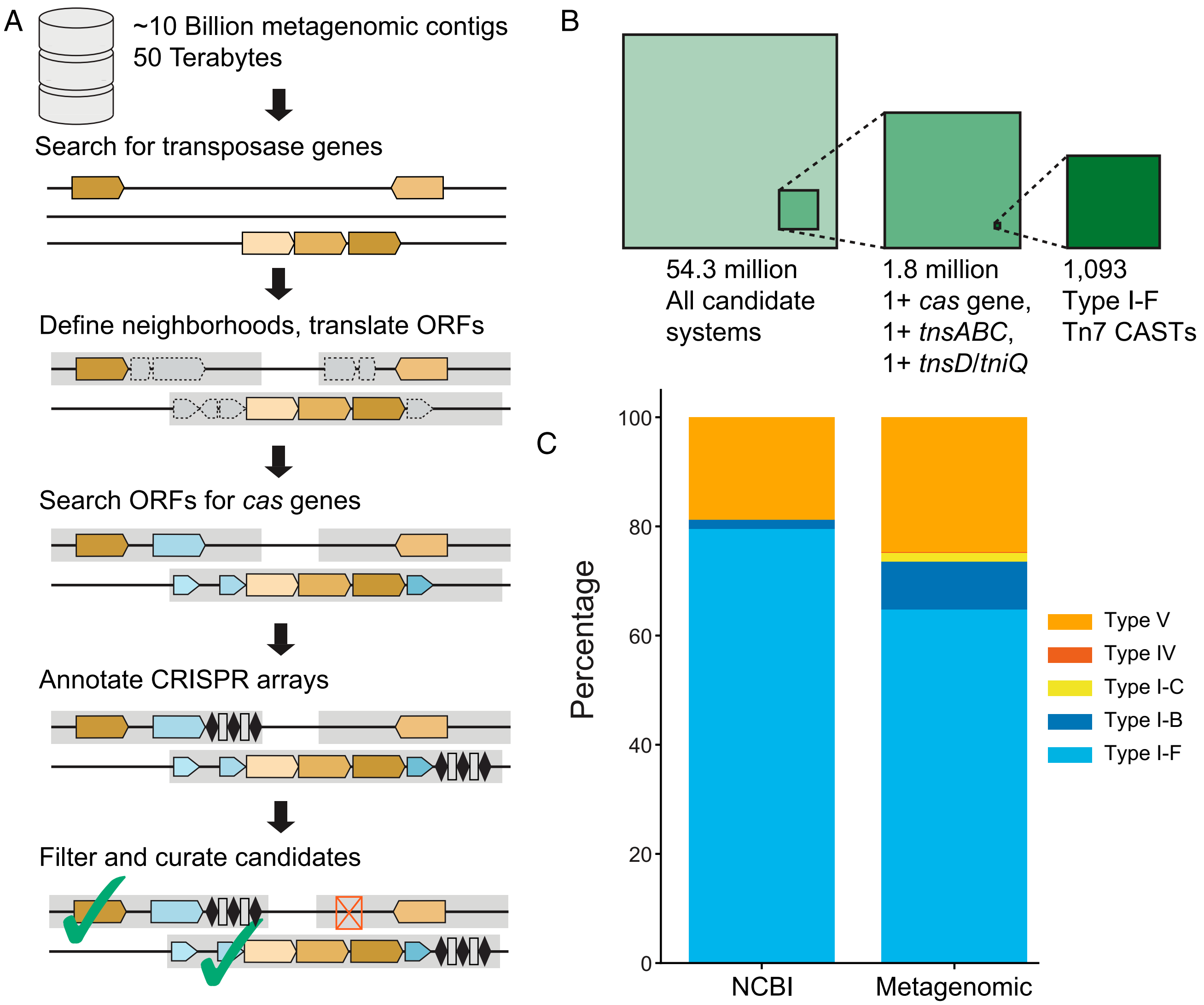

Fig. 1. CAST detection and classification. (*A*) A bioinformatic pipeline for the discovery of CASTs. Brown: transposase genes; blue: *cas* genes; dotted: ORFs; gray: gene neighborhoods. Neighborhoods satisfying initial search criteria are marked with a green check. Red "x" denotes a neighborhood that does not match the initial search criteria (e.g., no detected *cas* genes). (*B*) A summary of the stepwise filtering strategy to identify high-confidence Type I-F Tn7 CASTs. (*C*) The distribution of Tn7-associated CAST subtypes in the NCBI microbial genome and EMBL metagenomic databases.

Type I-F systems recognize a dinucleotide protospacer adjacent motif (PAM) (32). Our analysis of the homing PAMs highlighted that they vary with the *att* site and CAST subfamily (Fig. 2*D*). Next, we analyzed the sequence composition of the inverted repeats that span Tn7. Tn7 transposon ends consist of a short direct repeat, a larger inverted repeat, and several noncontiguous TnsB-binding sites (33). The right inverted repeat starts with a universally conserved 5′TGT that is recognized by the essential TnsB recombinase (34). The rest of this repeat varies but is most similar between CASTs that have the same attachment site (Fig. 2*D*). These results further confirm that I-F3c systems cluster into a distinct CAST subtype.

Type I-B CASTs Encode Multiple Integration Mechanisms. We found four families of Type I-B CASTs that lack interference and adaptation genes. These systems either encode a single *tniQ* or a *tniQ* and *tnsD* (Fig. 3*A*). Systems with dual *tniQ*/*tnsD* genes comprise 79% of all identified systems. The most common Type I-B system (I-B1) encodes a *tniQ* between *tnsC* and a *cas* gene, while *tnsD* is on the distal end of the CRISPR array (Fig. 3 *A*, *Top* and Fig. 3*B*). Systems with a single *tniQ* homolog (I-B2 and I-B3) have two distinct gene architectures and homing modalities. In I-B2 systems, where *tniQ* is sandwiched between *tnsC* and *cas6*, we identified a homing spacer that was complementary to a region downstream of *glmS* (Fig. 3 *A*, *Middle*). However, we did not find any homing spacers in I-B3 systems, where *tnsD* is between the CRISPR array and cargo genes. In addition, this *tnsD* is nearly double the length of the shorter *tniQ*s found between *tnsC* and *cas6*, bears a strong resemblance to the *tnsD* encoded in canonical Tn7 systems, and has recently been shown to recognize the CAST *att* site (third row, Fig. 3*A*) (7).

We identified an atypical Type I-B CAST (I-B4) that had a unique gene architecture and homing mechanism (last row, Fig. 3*A*). This system encodes *tnsB* and *tnsC* but lacks the *tnsA* gene, akin to Type V systems (8, 9). Phylogenetic analysis of TnsB and TnsC shows that this CAST clusters more closely to Tn5053 (which also lacks TnsA) than to Tn7 (*SI Appendix*, Fig. 1). Two *tniQ* homologs of unequal length are immediately adjacent to the inverted repeats but distal from the cas operon. *TniQ$_1$* is sandwiched between the right transposon end and a short CRISPR array; *tniQ$_2$* is only ~450 bp long and is located between *tnsC* and the left transposon end. This short TniQ$_2$ can be aligned against

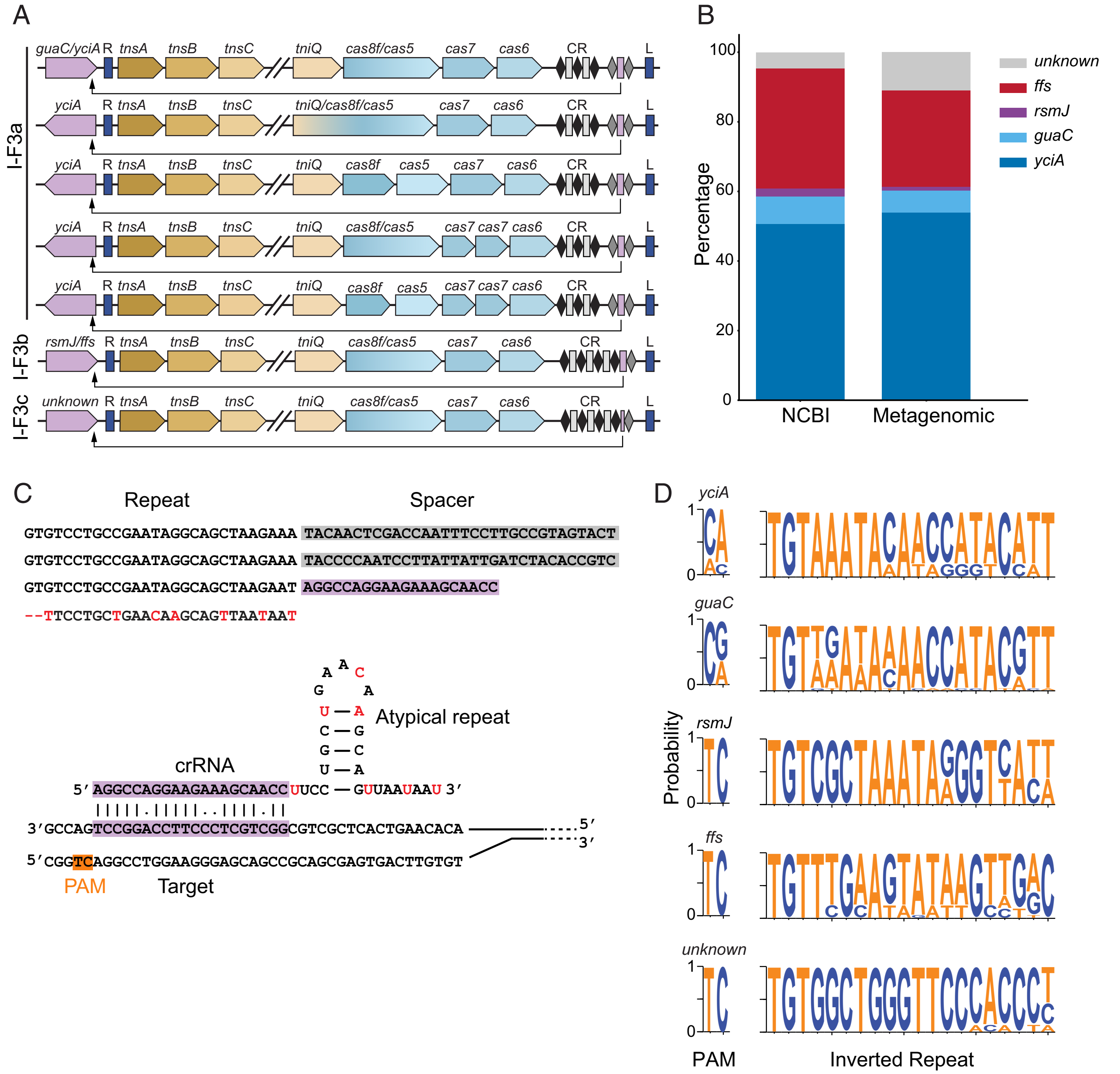
A
I-F3a
I-F3b
I-F3c
guaC/yciA R tnsA tnsB tnsC tniQ cas8f/cas5 cas7 cas6 CR L
yciA R tnsA tnsB tnsC tniQ/cas8f/cas5 cas7 cas6 CR L
yciA R tnsA tnsB tnsC tniQ cas8f cas5 cas7 cas6 CR L
yciA R tnsA tnsB tnsC tniQ cas8f/cas5 cas7 cas7 cas6 CR L
yciA R tnsA tnsB tnsC tniQ cas8f cas5 cas7 cas7 cas6 CR L
rsmJ/ffs R tnsA tnsB tnsC tniQ cas8f/cas5 cas7 cas6 CR L
unknown R tnsA tnsB tnsC tniQ cas8f/cas5 cas7 cas6 CR L
B
Percentage
0
20
40
60
80
100
NCBI
Metagenomic
unknown
ffs
rsmJ
guaC
yciA
C
Repeat
Spacer
GTGTCCTGCCGAATAGGCAGCTAAGAAA TACAACTCGACCAATTTCCTTGCCGTAGTACT
GTGTCCTGCCGAATAGGCAGCTAAGAAA TACCCCAATCCTTATTATTGATCTACACCGTC
GTGTCCTGCCGAATAGGCAGCTAAGAAT AGGCCAGGAAGAAAGCAACC
--TTCCTGCTGAACAAGCAGTTAATAAT
A A C
G A
U — A
C — G
G — C
U — A
Atypical repeat
crRNA
5′ AGGCCAGGAAGAAAGCAACCUUCC — GUUAAUAAU 3′
3′GCCAGTCCGGACCTTCCCTCGTCGGCGTCGCTCACTGAACACA 5′
5′CGGTCAGGCCTGGAAGGGAGCAGCCGCAGCGAGTGACTTGTGT 3′
PAM
Target
D
yciA
guaC
rsmJ
ffs
unknown
Probability
1
0
PAM
Inverted Repeat

Fig. 2. A summary of Type I-F CASTs. (*A*) The gene architectures of Type I-F3a, I-F3b, and I-F3c systems. Unique gene architectures include *tniQ-cas8* fusions, split *cas8* and *cas5*, and dual *cas7* systems. Purple: *att* site; blue: left and right transposon ends. Black diamonds: canonical direct repeats; gray diamonds: atypical direct repeats. Rectangles: spacers; purple rectangle: homing spacer. The arrow indicates the target site. The slanted gapped lines indicate elided cargo regions. (*B*) The distribution of *att* site genes in the NCBI and the metagenomic databases. (*C*, *Top*) The sequence of a CRISPR array with a short, atypical spacer (purple) that may assemble a mini Cascade. The red bases are those that differ from the consensus repeat sequence. (*Bottom*) A schematic of an atypical crRNA and its target DNA sequence. (*D*) Web logos of the PAM and right inverted repeat adjacent to each *att* site. The TnsB-binding site and the homing PAMs are conserved within subsystems.

the N terminus of traditional CAST-associated TniQs (Fig. 3*D*). Notably, this is the only dual *tniQ* CAST that encodes a homing spacer with near-perfect complementarity to a region of DNA just outside the transposon. The *att* site is adjacent to a gene of unknown function near the left transposon end (Fig. 3*A*), akin to the *att* sites in Type V CASTs. The homing spacer is flanked by an atypical direct repeat and is also 6 to 23 bp shorter than the other spacers in the CRISPR array (Fig. 3*C*). Based on these findings and our observation of Type I-F CASTs with short homing spacers, we propose that Type I-B systems can also assemble homing mini Cascades.

To better understand the roles of multiple TniQ/TnsD homologs in Type I-B CASTs, we constructed a phylogenetic tree of Type I-B, I-F3a, and I-F3b TniQs along with TnsD from Tn7 and TniQ from Tn5053 transposons (Fig. 3*B*). Compared to the cluster of TniQ, the short TniQ from Type I-B1 and B2 systems are closer to TniQ from other CASTs, while the Type I-B1 TnsD clusters with canonical Tn7 TnsD. These results are consistent with a recent report that the CAST TnsD serves the same role as Tn7 TnsD, namely that it is a sequence-specific DNA-binding protein that directs transposition downstream of *glmS* (7). We note that both TniQ1 and TniQ2 in Type I-B4

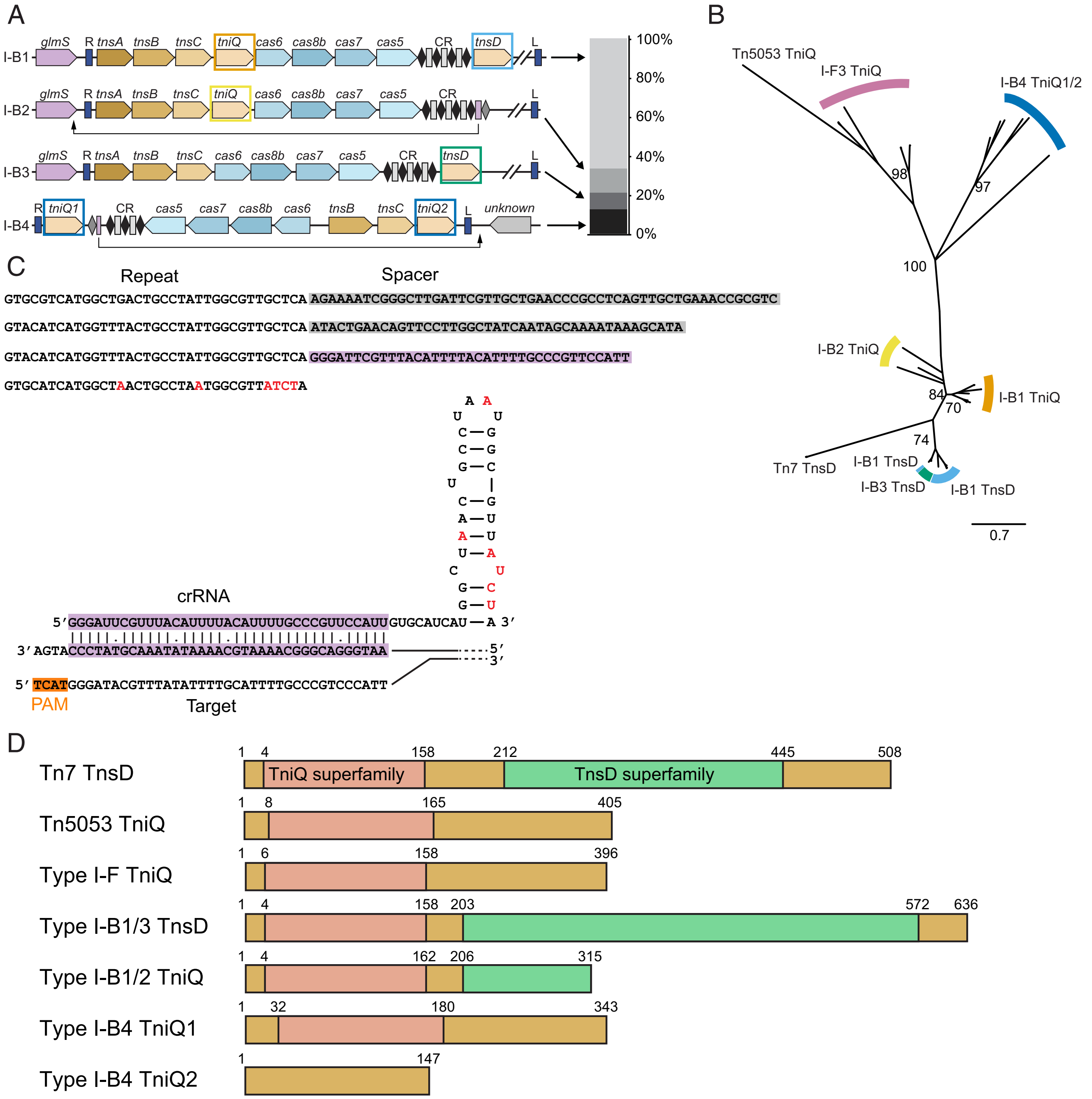

Fig. 3. An analysis of Type I-B CASTs. (*A*, *Left*) The gene architectures of Type I-B systems. Systems can dispense with either the first or the second *tniQ*/*tnsD*, suggesting alternative targeting lifestyles. Type I-B4 systems have a unique architecture that most resembles Type V CASTs. Colored rectangles correspond to phylogenetic groups in *B*. (*Right*) The distribution of Type I-B subsystems in the metagenomic database. (*B*) A phylogenetic tree with TniQ/TnsD variants from Type I-B and I-F CASTs as well as from the canonical Tn7 and Tn5053 transposons. The values at branch points are bootstrap support percentages. (*C*, *Top*) The sequence of a Type I-B4 CRISPR array with a short, atypical spacer. (*Bottom*) A schematic of an atypical crRNA base paired with a target DNA sequence. The red bases are those that differ from the consensus repeat sequence. (*D*) Domain maps of TniQ/TnsD proteins. Regions homologous to the TniQ superfamily and the TnsD superfamily are indicated in pink and light green, respectively. The Type I-B4 system encodes the shortest TniQ variant.

CASTs cluster closely with other TniQ homologs. These CASTs also encode a homing spacer, suggesting that these homologs mediate CRISPR-guided target selection.

Type I-C and IV CASTs from Metagenomic Sources. Our analysis of the metagenomic contigs revealed a Type IV CAST (Fig. 4*A*). Type IV systems are primarily encoded by plasmid-like elements to mediate interplasmid conflicts (35, 36). Phylogenetic trees of Cas6 and Cas7 independently placed this CAST within the Type IV-A3 subfamily (Fig. 4*B*). These systems frequently shed their CRISPR repeats instead of using distal CRISPR arrays (35). Although we did not find any CRISPR repeats in this system using Crass or PILER-CR, we detected a spacer-like DNA segment with strong complementarity to the C terminus of *glmS*, the likely *att* site. This putative homing spacer is adjacent to two hairpins, both of which resemble the direct repeats in other Type IV CRISPR-Cas systems (Fig. 4 *A*, *Bottom*). We conclude that this minimal spacer–repeat motif directs homing by the Type IV system.

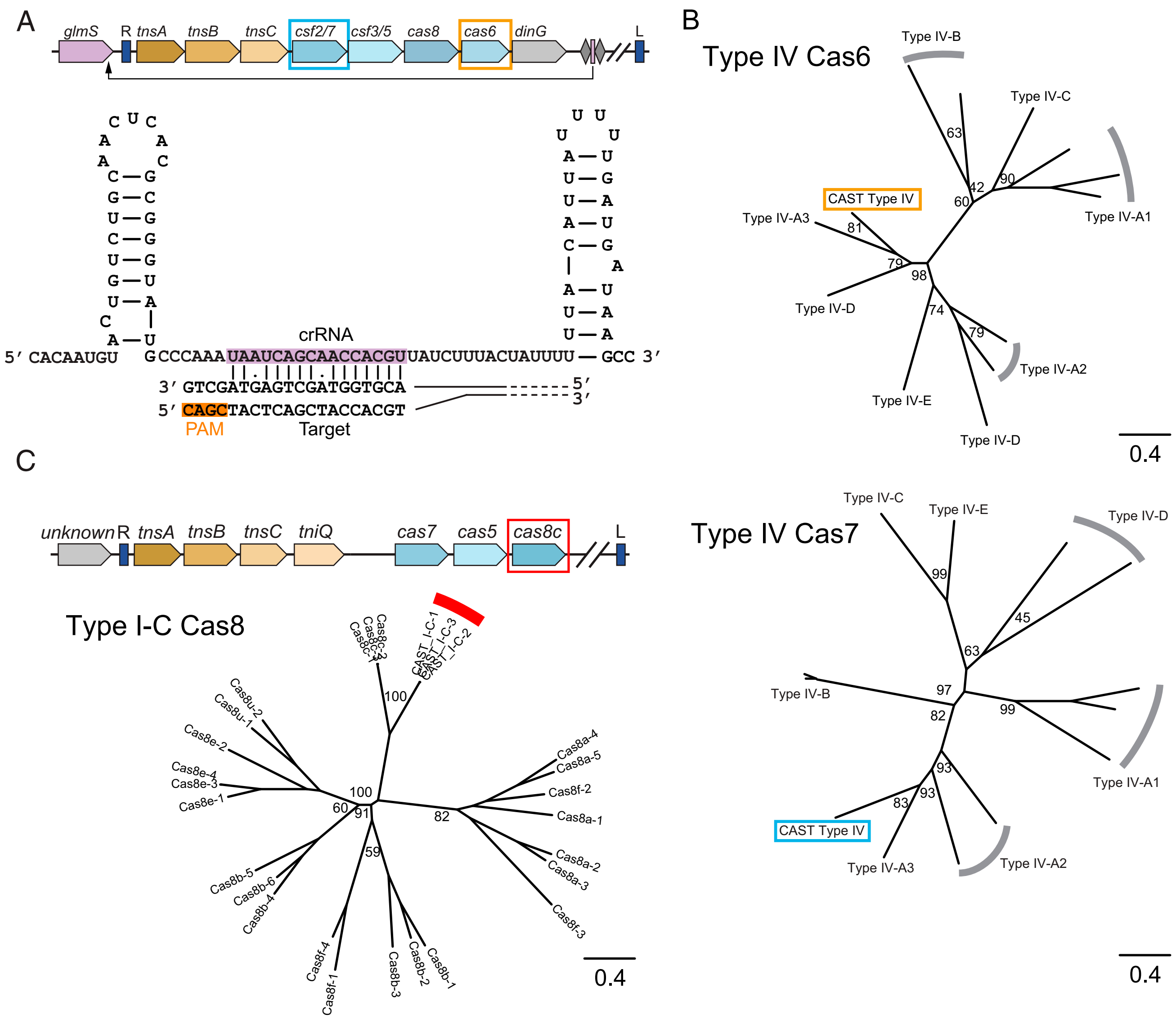

Fig. 4. New Tn7 CASTs from metagenomic databases. (*A*, *Top*) The gene architecture of a Type IV CAST. This system lacks a CRISPR array but encodes a homing spacer. The genes highlighted by colored rectangles correspond to genes in *B*. (*Bottom*) A schematic of a short, homing spacer base paired with its target DNA sequence. (*B*) Phylogenetic trees of Cas6 and Cas7 indicate that the Type IV CAST most closely resembles Type IV-A3 CRISPR-Cas systems. The values at branch points are bootstrap support percentages. (*C*, *Top*) The gene architecture of Type I-C systems. We did not detect any CRISPR arrays or atypical homing spacers. (*Bottom*) A phylogenetic tree of Cas8 confirms that this system is closely related to Type I-C Cascades. The values at branch points are bootstrap support percentages.

Horizontal transfer may still occur via a distal CRISPR array akin to the interference mechanism in other Type IV CRISPR-Cas systems (35).

We found nine nonredundant Type I-C CASTs with *cas5*, *cas7*, and *cas8* downstream of *tnsABC* and *tniQ* (Fig. 4*C*). *TniQ* is immediately adjacent to *tnsABC* rather than the *cas* genes. A phylogenetic analysis of Cas8 showed close similarity to Cas8c (Fig. 4 *C*, *Right*). We did not detect a CRISPR array via CRASS or PILER-CR. We also did not detect any transfer RNA (tRNA) or common Tn7-associated *att* sites near either the left or right transposon ends, precluding a detailed analysis of the homing mechanism. We cannot rule out that these systems use a minimal leader–spacer array that was below the threshold of our detection software. Alternatively, these systems may use a distal CRISPR array in *trans*. Further bioinformatic and experimental analyses will be required to delineate the mechanisms of homing and horizontal transfer in these systems.

Architectural Diversity and Homing in Type V CASTs. Type V CASTs were likely formed when a Tn7-like transposon co-opted a *cas12* gene for RNA-guided DNA targeting (10, 25). Most Type V CASTs contain *tnsB*, *tnsC*, and *tniQ* at one end of the transposon with *cas12k*, a small CRISPR array, and an atypical repeat–spacer on the other end. Cargo genes spanning 2 to 23 kb of additional DNA sequences are sandwiched between *tniQ* and *cas12k* (Fig. 5*A*). In contrast to Class 1 systems, all metagenomic Type V systems lacked *tnsA*, consistent with the proposal that *cas12k* was captured by a Tn5053 family transposon, which contains *tnsB*, *tnsC*, and *tniQ* homologs but also lacks *tnsA* (37, 38). The *trans*-activating CRISPR RNA (tracrRNA) is upstream of the canonical CRISPR array with homology to the atypical repeat. The crRNA usually has good homology to the target DNA, with 98% of systems containing one or zero mismatches in the first 10 bps (Fig. 5*B*). As previously observed (8), atypical spacers generally targeted tRNA genes immediately

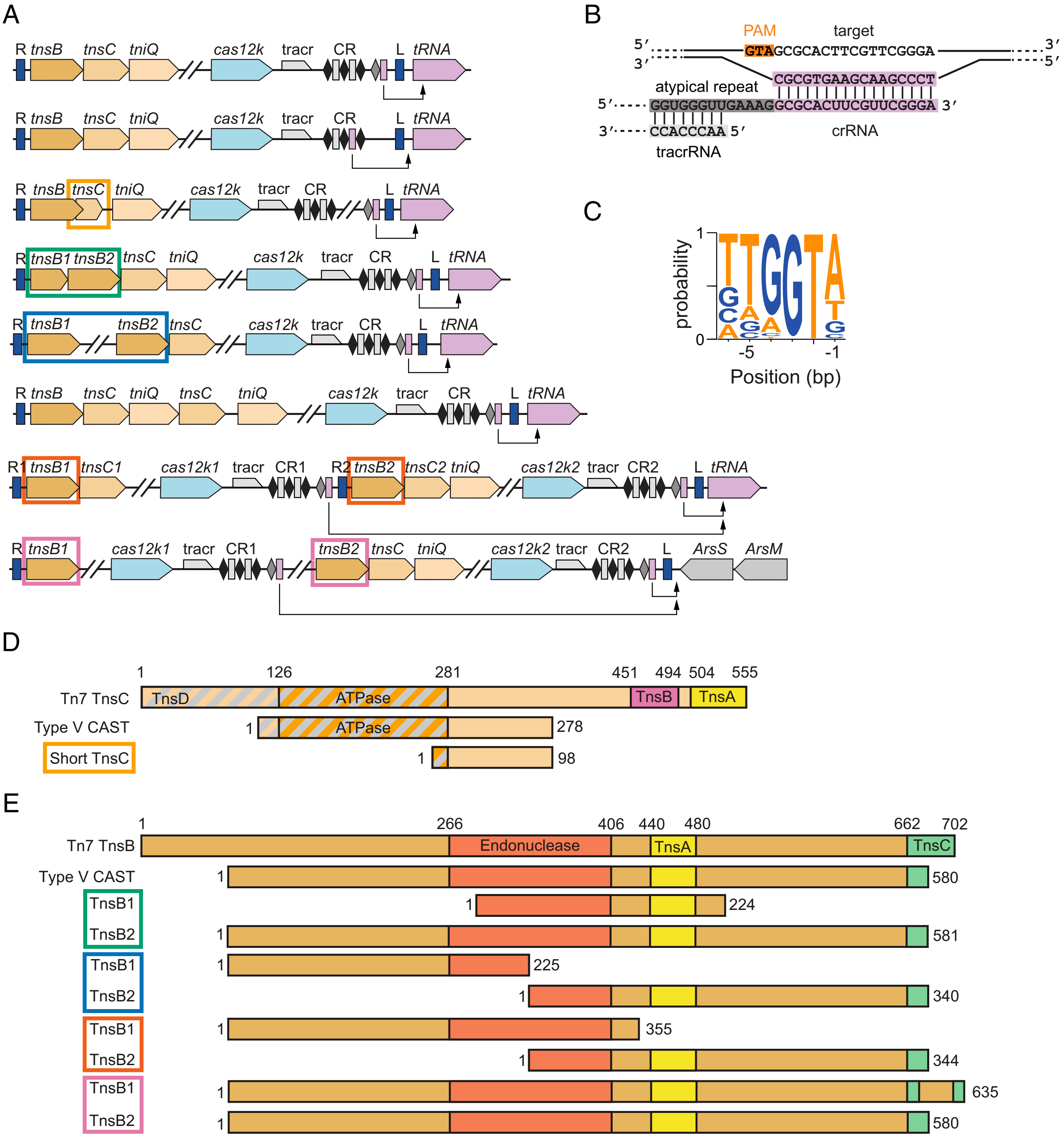
A
R tnsB tnsC tniQ cas12k tracr CR L tRNA
R tnsB tnsC tniQ cas12k tracr CR L tRNA
R tnsB tnsC tniQ cas12k tracr CR L tRNA
R tnsB1 tnsB2 tnsC tniQ cas12k tracr CR L tRNA
R tnsB1 tnsB2 tnsC cas12k tracr CR L tRNA
R tnsB tnsC tniQ tnsC tniQ cas12k tracr CR L tRNA
R1 tnsB1 tnsC1 cas12k1 tracr CR1 R2 tnsB2 tnsC2 tniQ cas12k2 tracr CR2 L tRNA
R tnsB1 cas12k1 tracr CR1 tnsB2 tnsC tniQ cas12k2 tracr CR2 L ArsS ArsM
B
PAM
target
5′
3′
GTAGCGCACTTCGTTCGGGA
3′
5′
atypical repeat
CGCGTGAAGCAAGCCCT
5′ GGUGGGUUGAAAGGCGCACUUCGUUCGGGA 3′
3′ CCACCCAA 5′
crRNA
tracrRNA
C
probability
1
0
-5
-1
Position (bp)
D
1
126
281
451
494
504
555
Tn7 TnsC
TnsD
ATPase
TnsB
TnsA
Type V CAST
1
ATPase
278
Short TnsC
1
98
E
1
266
406
440
480
662
702
Tn7 TnsB
Endonuclease
TnsA
TnsC
Type V CAST
1
580
TnsB1
1
224
TnsB2
1
581
TnsB1
1
225
TnsB2
1
340
TnsB1
1
355
TnsB2
1
344
TnsB1
1
635
TnsB2
1
580

Fig. 5. An analysis of Type V CASTs. (*A*) The gene architectures of Type V CASTs, including dual-insertion systems (*Bottom* two rows). The colored rectangles around genes correspond to alignments in *D* and *E*. (*B*) A schematic of interactions between the target site DNA, a homing crRNA, and a tracrRNA. (*C*) A web logo of PAM sequences found adjacent to spacer targets. (*D*) Aligned domain maps of truncated TnsC variants. Gray diagonal stripes indicate the TnsD-interacting region. Truncated TnsCs lack the TnsA- and TnsB-interacting domains but generally retain the ATPase domain and most of the TnsD-interacting domain. The shortest TnsC has also lost its ATPase domain. (*E*) Aligned domain maps of truncated TnsB variants. Type V CAST TnsB is shorter than Tn7 TnsB but contains the functionally annotated domains. In some dual TnsB systems, the first *tnsB* encodes the N-terminal region, and the second encodes the C-terminal portion.

adjacent to the transposon. We also found one system that attaches 104 bp downstream of *arsS*, which codes for an arseno-sugar biosynthesis radical *S*-adenosylmethionine protein. An analysis of the DNA upstream of the homing spacer revealed 5′-TGGTA as the most common PAM, with some variability in the −5, −4, and −1 positions. (Fig. 5*C*). Experimental evidence for two CASTs showed a preference for a smaller 5′-GTN PAM (9). Whether the broader set of Cas12k proteins have more stringent PAM requirements will require experimental validation. Overall, this architecture and preference for tRNA attachment sites corroborate previous bioinformatic and experimental observations (7–9, 25).

We also found Type V CASTs with unusual *tnsC* and *tnsB* arrangements. Notably, all Type V TnsC proteins lack the canonical TnsA- and TnsB-interacting domains and have partial truncations of the TniQ-interacting domain (39–41). The shortest CAST—only 6.6 kbp in total, including the cargo—encodes a 98-amino acid *tnsC* fragment whose sequence overlaps with *tnsB* by 115 bp (Fig. 5 *A* and *D*). In addition to losing the TniQ-, TnsB-, and TnsA-interacting domains, this TnsC has also lost its ATPase domain (39, 41). We hypothesize that the minimal *tnsC* encodes uncharacterized TnsB- and TniQ-interaction motifs. Because of its compact organization, this CAST is also a prime candidate for gene-editing applications.

Multiple CASTs split *tnsB* into separate ORFs that encode just the *N* or C terminus. Alternatively, a full-length *tnsB* is encoded next to a *tnsB* fragment containing most of the catalytic domain (Fig. 5*E*). Strikingly, two unrelated systems encode the same N-terminal region of *tnsB*. We speculate that these split TnsBs form a heterodimeric $TnsB_1$:$TnsB_2$ transposition complex. These heterodimeric complexes retain the catalytic core while also maintaining the requisite TnsC interaction motifs via the longer TnsB subunit.

We also observed multiple contigs with two CASTs inserted at the same attachment site (*Bottom* rows, Fig. 5*A*). Tn7 transposons can prevent reinsertion at the attachment site by TnsB-mediated dissociation of TnsC from target DNA (42, 43). However, more distant Tn7 family transposases may still insert at a single attachment site, resulting in several transposons that are situated adjacent to each other (41). Consistent with this idea, the dual-insertion Type V CASTs have distinct cargos, unique gene architectures, and divergent *cas12k* sequences. In one contig, the tRNA distal CAST encoded an N-terminal *tnsB* truncation and lost *tniQ* (seventh row, Fig. 5*A*). The tRNA proximal CAST from the same organism encoded the C-terminal *tnsB* fragment and had a complete Type V family *tniQ*. A second dual CAST system had lost both *tnsC* and *tniQ* from the tRNA distal transposon (last row, Fig. 5*A*). Both systems encode a homing spacer and a full *cas12k* gene, suggesting that they are both still active.

Tn7 insertion results in the hallmark duplication of several bps flanking the inverted repeats (44). We identified similar direct repeats in the dual-insertion CASTs. In the first such system (seventh row, Fig. 5*A*), a 5′-GGACA repeat flanks R2 and L, while R1 is not bound by the direct repeat. Such an outcome may have occurred if R1 was the original right transposon end but was degraded over time. Insertion of another transposon bound by R2 and L would then generate the observed repeat. The second system (eighth row, Fig. 5*A*) was bounded by direct repeats of 5′-GTACA. This is consistent with a functional CAST that was formed by dual insertion followed by the loss of the upstream *tnsC* and *tniQ* and the formation of a heterodimeric *tnsB*.

Phylogenetic Analysis of Tn7-CASTs. To clarify the evolutionary relationship between Tn7-CASTs, we built phylogenetic trees of the TnsB (*SI Appendix*, Fig. 1*A*) and TnsC (*SI Appendix*, Fig. 1*B*) proteins from all known CAST subtypes as well as Tn7 family transposons (see *Materials and Methods*). The phylogenetic relationships between subsystems were nearly identical for both TnsB and TnsC, suggesting that these proteins are coevolving as a system. We omitted TnsA from this analysis because Type I-B4 and all Type V Tn7-CASTs lack this gene. We confirmed that all metagenomic Type V CASTs are phylogenetically closer to the Tn5053 transposon than Tn7. In contrast, Type I-B1-3, I-C, and IV CASTs are phylogenetically close to Tn7. Such limited evolutionary drift may suggest a relatively recent co-opting of this CRISPR-Cas system. Type I-B4 CASTs are a notable exception because these systems lack *tnsA* and cluster closer to Tn5053 than to the reference Tn7. Type I-F CASTs are highly divergent from both Tn7 and Tn5053, with a large phylogenetic separation between the I-F3a and I-F3b subtypes. The diversity of Tn7 family transposases and CRISPR subsystems that have coevolved suggests that we will continue to find new CAST families as metagenomic databases expand, and detection sensitivity increases.

Non-Tn7 CASTs that Co-Opt Cas12 and Type I-E Cascade. While transposons other than Tn7 may have co-opted CRISPR-Cas systems for *att* site recognition, this has never been reported. To explore this possibility, we identified contigs that encode 1) at least one non-Tn7 transposase gene (see *Materials and Methods*), 2) a CRISPR array containing at least two spacers, and 3) Class 1 or Class 2 DNA-binding effectors (i.e., *cas9*, *cas12*, or any three of *cas5*/*6*/*7*/*8*). We excluded contigs that encode interference (i.e., *cas3* or *cas10*) or acquisition machinery (i.e., *cas1* or *cas2*). Class 2 nucleases were additionally filtered by size to exclude truncated genes (see *Materials and Methods*). We prioritized Type II, V, and VI systems in which the catalytic nuclease domain residues are mutated or deleted, as these enzymes cannot participate in adaptive immunity. We detected nuclease-inactivating mutations or deletions in 25% of *cas9* genes (in one or both nuclease domains) and 8% of *cas12* genes.

We found 40 nonredundant examples of a nuclease-inactive *cas12* or a Type I-E Cascade near a putative Rpn-like gene (Fig. 6*A*). Rpn family proteins were originally investigated because of their close homology to the catalytic domain of transposase_31 (45). These proteins contain a PDDEXK nuclease domain first discovered in restriction endonucleases but also observed in T7 TnsA and other diverse DNA-processing enzymes (46–48) (Fig. 6*B*). *Escherichia coli* RpnA promotes RecA-independent gene transfer in cells and is a Ca^{2+}-stimulated DNA nuclease in vitro (45). The mechanism of how RpnA promotes horizontal gene transfer is unknown.

The genetic context around *cas12* in our systems is highly enriched with nucleic acid–interacting proteins, including a topoisomerase, a ribonuclease, a DNA polymerase subunit, and one or two helicases. Likewise, the Cascade system encodes a helicase and an HNH endonuclease. Three Cas12 systems encode an HU family DNA-binding protein, and one of those also contains a protein with homology to a phage replisome organizer (49). We detected three systems with putative atypical homing spacers adjacent to a canonical CRISPR array (Fig. 6*A*). In the Cas12 systems, the homing spacer is complementary to two or four nearby targets, all of which are positioned at intergenic sequences. The closest targets of these homing sequences are adjacent to a 5′-CTTA PAM, which is recognized by conventional Cas12a nucleases (Fig. 6*C*) (50, 51). There are 9-bp inverted repeats (with one mismatch) that flank *cas12*, the Rpn family protein, and several other genes.

The Cas12 proteins encoded by these systems cover 90% of the well-characterized AsCas12a sequence (∼24% amino acid identity), including the crRNA-processing, DNA-binding, and DNA nuclease domains (52). A phylogenetic tree of Cas12 family members also places this protein closest to the Cas12a family (*SI Appendix*, Fig. 3). Cas12a can process its own pre-crRNA via a dedicated RNase domain (53). Three residues in this domain have been identified as critical for precrRNA processing; all are conserved in Rpn-associated Cas12 variants (Fig. 6*D*) (53). Cas12a degrades double-stranded DNA by first cleaving the nontarget strand followed by the target strand in its single RuvC nuclease active site (54–57). Phosphate bond scission is catalyzed by two magnesium ions, one of which is coordinated by a critical aspartic residue (position 908 in *Acidaminococcus* [*As*] Cas12a). This residue is mutated to isoleucine in all Rpn-associated systems (Fig. 6*D* and *SI Appendix*, Fig. 2). Similarly, the Rpn-associated Type I-E system encodes all the Cascade subunits but does not have *cas3* or *cas1*–*cas2*. We

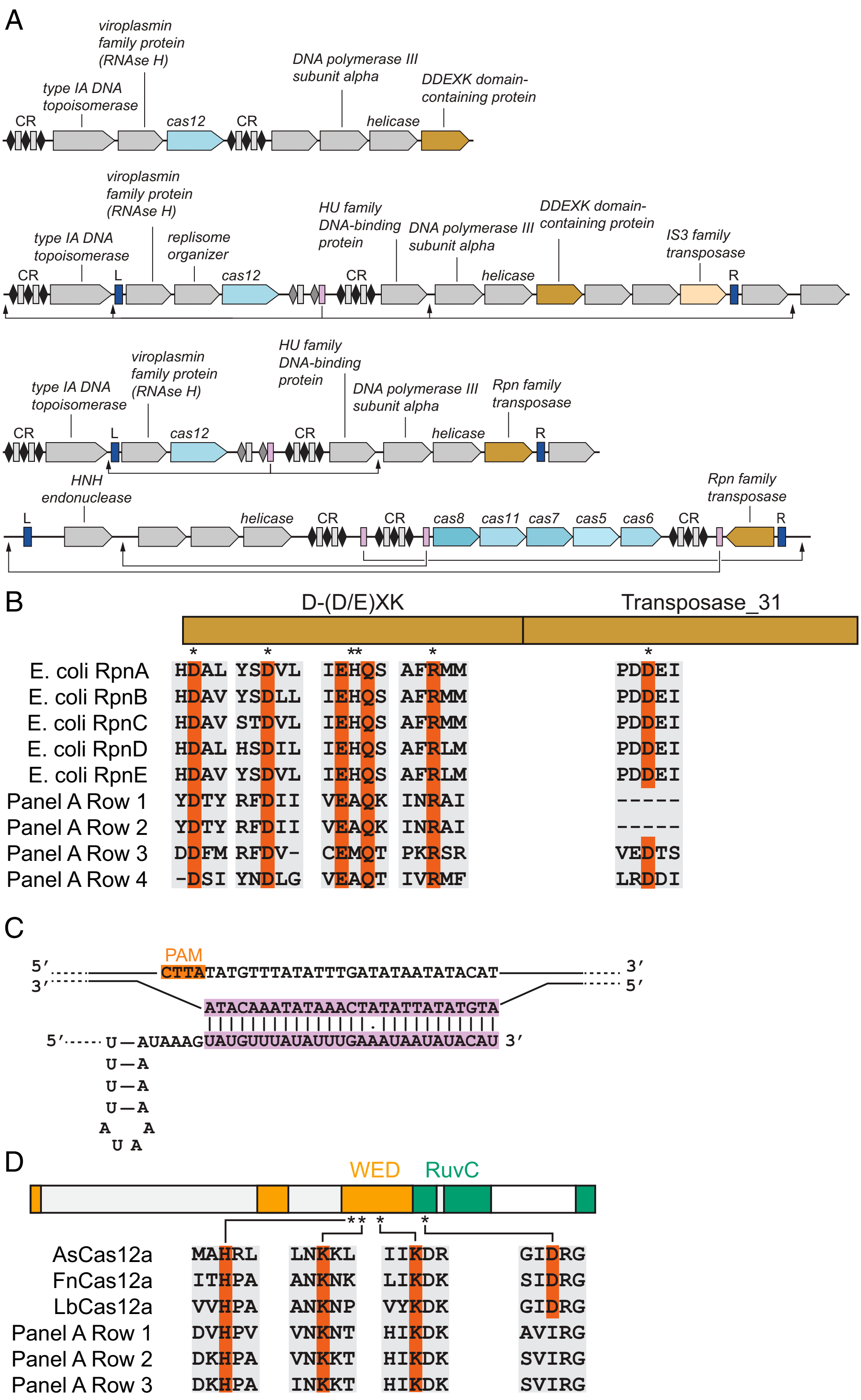
A
viroplasmin family protein (RNAse H)
type IA DNA topoisomerase
DNA polymerase III subunit alpha
DDEXK domain-containing protein
CR
cas12
CR
helicase
viroplasmin family protein (RNAse H)
type IA DNA topoisomerase
replisome organizer
HU family DNA-binding protein
DNA polymerase III subunit alpha
DDEXK domain-containing protein
IS3 family transposase
CR
L
cas12
CR
helicase
R
viroplasmin family protein (RNAse H)
type IA DNA topoisomerase
HU family DNA-binding protein
DNA polymerase III subunit alpha
Rpn family transposase
CR
L
cas12
CR
helicase
R
HNH endonuclease
Rpn family transposase
L
helicase
CR
CR
cas8
cas11
cas7
cas5
cas6
CR
R
B
D-(D/E)XK
Transposase_31
E. coli RpnA HDAL YSDVL IEHQS AFRMM PDDEI
E. coli RpnB HDAV YSDLL IEHQS AFRMM PDDEI
E. coli RpnC HDAV STDVL IEHQS AFRMM PDDEI
E. coli RpnD HDAL HSDIL IEHQS AFRLM PDDEI
E. coli RpnE HDAV YSDVL IEHQS AFRLM PDDEI
Panel A Row 1 YDTY RFDII VEAQK INRAI -----
Panel A Row 2 YDTY RFDII VEAQK INRAI -----
Panel A Row 3 DDFM RFDV- CEMQT PKRSR VEDTS
Panel A Row 4 -DSI YNDLG VEAQT IVRMF LRDDI
C
PAM
5′ CTTATATGTTTATATTTGATATAATATACAT 3′
3′ 5′
ATACAAATATAAACTATATTATATGTA
5′ U—AUAAAGUAUGUUUAUAUUUGAAAUAAUAUACAU 3′
U—A
U—A
U—A
A A
U A
D
WED
RuvC
AsCas12a MAHRL LNKKL IIKDR GIDRG
FnCas12a ITHPA ANKNK LIKDK SIDRG
LbCas12a VVHPA ANKNP VYKDK GIDRG
Panel A Row 1 DVHPV VNKNT HIKDK AVIRG
Panel A Row 2 DKHPA VNKKT HIKDK SVIRG
Panel A Row 3 DKHPA INKKT HIKDK SVIRG

Fig. 6. A family of putative non-Tn7 CASTs. (*A*) The defining features of this family of systems are an Rpn family (PDDEXK domain-containing) nuclease/transposase near a nuclease-dead Cas12 or a Type I-E Cascade complex. The operon is enriched for nucleic acid–processing proteins. We also observed homing spacers (magenta, black arrows) and short inverted repeats (blue) in some systems. (*B*) Multiple sequence alignment of Rpn proteins with the putative transposases from these systems. Residues critical for DNA cleavage in the PDDEXK domain are highlighted in red. The D165A mutant in RpnA more than doubles recombination in vivo; this aspartic acid is highlighted in red below the transposase_31 domain. (*C*) A schematic of an atypical homing spacer and its DNA target. The PAM is highlighted. (*D*) Multiple sequence alignment of nuclease-active Cas12a and putative CAST Cas12 proteins. Putative CAST Cas12 proteins retain the conserved residues in the WED domain that are essential for crRNA processing but lack an aspartic residue in the RuvC domain that is essential for DNA cleavage.

conclude that these systems bear striking resemblance to Tn7-associated CASTs and may also mobilize genomic information for crRNA-guided horizontal transfer.

Discussion

CASTs are rare in fully sequenced prokaryotic genomes and are likely to be missed by traditional CRISPR detection pipelines due to their unusual operon structures and short CRISPR arrays. To address this gap, we have developed a set of Python libraries that allow users to efficiently use BLAST to search for co-occurring genes and to perform subsequent searches for arbitrary gene architectures. We examined ∼50 terabytes of metagenomic contigs to identify ∼1,476 high-confidence CASTs, including Type IV and Type I-C systems as well as a Type I-B4 CRISPR system that coevolved with a Tn5053-like element, a member of the Tn7 family of transposons that lacks *tnsA*. We have also discovered systems that include a putative nuclease-inactive Cas12 and a non-Tn7 transposase-like recombinase. The DNA sequences for all CASTs discovered here are available via our git repository (see *Materials and Methods*). The diversity of CRISPR subtypes that have been co-opted by transposons will likely increase as metagenomic databases and sensitive detection pipelines continue to improve (58, 59). More broadly, the abundance of nuclease-inactive CRISPR-Cas operons may suggest that the scope of cellular processes that have co-opted CRISPR-Cas systems is wider than is currently known.

In the NCBI and metagenomic databases, the most abundant Tn7-associated CASTs are those that co-opt Class 1 CRISPR systems. Of these, Type I-F subsystems are the most structurally diverse. Notably, we found that some I-F CASTs encode *tniQ-cas8*/*cas5* fusions, duplicate *cas5*s, and duplicate *cas7*s. We speculate that gene duplication of the *cas5* and/or *cas7* allowed one of the paralogs to form a protein–protein interface with TniQ. The second paralog may have been subsequently lost. The remaining paralog resulted in the streamlined I-F CASTs that are most frequently found in bacterial genomes. The short atypical precrRNA in some I-F CASTs also suggests that these systems assemble a shortened Cascade for homing. The size of Type I-E and I-F Cascades can be tuned by the length of the crRNA (60–64). Intriguingly, short I-F Cascades cannot recruit Cas3 but are still able to bind the target DNA, making them an ideal system for directed Tn7 transposition (63). Short Cascades, along with the atypical direct repeats, may differentiate homing CASTs from those undergoing horizontal gene transfer in the I-F3c system.

Type I-B CASTs encode one or two *tniQ*/*tnsD* homologs. A recent report has uncovered that homing in some systems proceeds via TnsD, whereas horizontal transfer is crRNA guided (7). We also identified atypical systems that encode two short *tniQ* homologs along with a homing spacer. Both homologs in the atypical dual-*tniQ* systems are related to the I-F CAST *tniQ*. Based on this observation, along with the crRNA-directed homing and the alignment of TniQ_2 to the N terminus of TniQ_1, we propose that this CAST assembles a heterodimeric Cascade consisting of a single repeat of each subunit. Alternatively, this system may assemble TniQ_1- and TniQ_2-only Cascades for homing and horizontal transfer. Additional studies will be required to decipher the role of TniQ_2 in the atypical I-B systems with a homing spacer.

How do CASTs target mobile genetic elements with minimal CRISPR arrays? We did not find any systems that retained the Cas1/Cas2 acquisition machinery, suggesting that strong evolutionary pressure is preventing the CAST-associated CRISPR arrays from expanding. CASTs encode CRISPR arrays that are significantly shorter than the corresponding canonical CRISPR-Cas systems, and these arrays may also be transcriptionally silenced via *xre* elements that are frequently found adjacent to these arrays in CASTs (11). Moreover, we could not identify any CRISPR arrays in Type IV and I-C CASTs. We propose that CASTs use CRISPR arrays that occur elsewhere in the genome—perhaps in functional CRISPR-Cas systems—for horizontal gene transfer. Evidence for such in *trans* CRISPR array usage has already been documented for a large set of canonical CRISPR-Cas systems (65, 66). CRISPR arrays that are associated with active interference machinery serve as an ever-updating record of the most likely mobile genetic elements that the CAST can use for horizontal gene transfer (67). A second possibility is that Cas1/2/4 from an active CRISPR-Cas system can act in *trans* to add spacers to the CAST CRISPR array. This may be an important secondary mechanism when horizontal transfer places the CAST into a host that lacks a compatible CRISPR array.

Our search revealed *cas12* genes associated with Rpn family transposases, two of which appear to have atypical spacers that target two sites up- and downstream of *cas12*. This curious arrangement could be the result of a duplication of the target site that is originally present in only a single copy. We note that the HU family DNA-binding protein is only present in systems with putative homing spacers and that a homolog of this protein is essential in bacteriophage Mu for transpososome assembly (68). Two other genes of viral origin —a replisome organizer and a ribonuclease—are also found near *cas12* in a self-targeting system, hinting at an intriguing evolutionary path for the creation of this putative system. Whether these genes function as bona fide CASTs or other aspects of DNA metabolism will require extensive experimental follow-up. In sum, the CAST identification pipeline and diversity of new systems described herein add to our understanding of CRISPR transposons, expand the gene-editing toolkit, and hints at the possibility that nuclease-inactive *cas* genes may play additional roles in cellular DNA metabolism.

Materials and Methods

We developed a Python library, Opfi (short for Operon Finder), to search genomic or metagenomic sequence data for putative CRISPR transposons. This library consists of two modules, Gene Finder and Operon Analyzer. The Gene Finder module enables the user to use BLAST to identify genomic neighborhoods that contain specific sets of genes such as *cas9* or *tnsA*. It can also identify CRISPR repeats. The Operon Analyzer module further filters the output from Gene Finder by imposing additional user-defined constraints on the initial hits. For example, Operon Analyzer can be used to find all genomic regions that contain a transposase and at least two *cas* genes but no *cas3*.

We used Gene Finder to locate genomic regions of interest using the following logic. First, we located all regions containing at least one transposase gene. Within those regions, we next searched for *cas* genes located no more than 25 kb away from a transposase. Transposase-containing regions without at least one nearby *cas* gene were discarded from further analysis. Finally, the remaining regions were further annotated for Tn7 accessory genes (*tnsC-tnsE* and *tniQ*) and CRISPR arrays.

We further processed and categorized the Gene Finder hits using Operon Analyzer. To identify Tn7-like CRISPR transposons, we required each putative operon to contain *tnsA*, *tnsB*, *tnsC*, and at least one *cas* gene from *cas5-13*; the distance between *tnsA*, *tnsB*, and *tnsC* needed to be less than 500 bp; the *cas* genes needed to be downstream of *tnsA*/*B*/*C*; and the distance between any *cas* gene and *tnsB* needed to be less than 15 kbp. We classify this dataset into putative Class 1 systems and Class 2 systems based on their *cas* signature genes. Class 1 systems were manually reviewed to confirm the loss of adaptation (*cas1* and *cas2*) and interference (*cas3*) genes.

To identify non-Tn7 CRISPR transposons, we required each putative operon to contain a CRISPR array, a transposase, and at least one *cas* gene from *cas5-12*. We also excluded systems containing Tn7 genes, *cas1*, or *cas2*. We then partitioned this dataset into putative Class 1 systems (defined as loci with any three of *cas5*/*6*/*7*/*8*) or Class 2 systems (*cas9* or *cas12*). For Class 1 systems, we further excluded those containing *cas3* or *cas10*. To eliminate systems with fragments of effector proteins or poor matches to unrelated proteins, we required that *cas9* have a size of 2 to 6 kbp and *cas12* a size of 3 to 6 kbp. We eliminated Class 2 systems that were nucleolytically active and finally clustered all systems using mmseqs easy cluster with a minimum sequence identity of 0.95 (69) to simplify manual curation.

SI Appendix, Materials and Methods includes additional details.

Data Availability. All CAST DNA sequences, protein family databases, processed data, and analysis scripts are available on GitHub at https://github.com/wilkelab/Metagenomics_CAST. An archival snapshot of the git repository has been deposited on Zenodo and is publicly available (DOI: 10.5281/zenodo.5498160). The analysis scripts make use of the Gene Finder and Operon Analyzer modules of the Opfi package (70), publicly available on the Python Package Index: https://pypi.org/project/opfi/. The Opfi source code is available on GitHub at https://github.com/wilkelab/Opfi. An archival snapshot of the git repository has been deposited on Zenodo and is publicly available (DOI: 10.5281/zenodo.5497939).

ACKNOWLEDGMENTS. We are grateful to the staff of the Texas Advanced Computing Center for providing computational resources and to members of the I.J.F. and C.O.W. laboratories for helpful discussions. This work was supported by the National Institute of General Medical Sciences Grants R01GM124141 (to I.J.F.) and R01GM088344 (to C.O.W.), the Welch Foundation Grant F-1808 (to I.J.F.), and the College of Natural Sciences Catalyst Award for seed funding.

1. J. E. Peters, Targeted transposition with Tn7 elements: Safe sites, mobile plasmids, CRISPR/Cas and beyond. *Mol. Microbiol.* **112**, 1635–1644 (2019).
2. J. E. Peters, K. S. Makarova, S. Shmakov, E. V. Koonin, Recruitment of CRISPR-Cas systems by Tn7-like transposons. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E7358–E7366 (2017).
3. T. S. Halpin-Healy, S. E. Klompe, S. H. Sternberg, I. S. Fernández, Structural basis of DNA targeting by a transposon-encoded CRISPR-Cas system. *Nature* **577**, 271–274 (2020).
4. N. Jia, W. Xie, M. J. de la Cruz, E. T. Eng, D. J. Patel, Structure-function insights into the initial step of DNA integration by a CRISPR-Cas-Transposon complex. *Cell Res.* **30**, 182–184 (2020).
5. S. E. Klompe, P. L. H. Vo, T. S. Halpin-Healy, S. H. Sternberg, Transposon-encoded CRISPR-Cas systems direct RNA-guided DNA integration. *Nature* **571**, 219–225 (2019).
6. B. Wang, W. Xu, H. Yang, Structural basis of a Tn7-like transposase recruitment and DNA loading to CRISPR-Cas surveillance complex. *Cell Res.* **30**, 185–187 (2020).
7. M. Saito, *et al.*, Dual modes of CRISPR-associated transposon homing. *Cell* **184**, 2441−2453 (2021).
8. S.-C. Hsieh, J. E. Peters, Tn7-CRISPR-Cas12K elements manage pathway choice using truncated repeat-spacer units to target tRNA attachment sites. bioRxiv [Preprint] (2021). https://doi.org/10.1101/2021.02.06.429022 (Accessed 6 February 2021).
9. J. Strecker *et al.*, RNA-guided DNA insertion with CRISPR-associated transposases. *Science* **365**, 48–53 (2019).
10. 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).
11. 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).
12. Y. Zhang *et al.*, Multicopy chromosomal integration using CRISPR-associated transposases. *ACS Synth. Biol.* **9**, 1998–2008 (2020).
13. P. L. H. Vo *et al.*, CRISPR RNA-guided integrases for high-efficiency, multiplexed bacterial genome engineering. *Nat. Biotechnol.* **39**, 480–489 (2021).
14. S. Yang, *et al.*, Orthogonal CRISPR-associated transposases for parallel and multiplexed chromosomal integration. *Nucleic Acids Res.* **49**, 10192−10202 (2021)
15. W. Chen *et al.*, Targeted genetic screening in bacteria with a Cas12k-guided transposase. *Cell Rep.* **36**, 109635 (2021).
16. B. E. Rubin, *et al.*, Targeted genome editing of bacteria within microbial communities. bioRxiv [Preprint] (2020). https://doi.org/10.1101/2020.07.17.209189 (Accessed 21 July 2020).
17. M. Jinek *et al.*, A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
18. A. V. Anzalone *et al.*, Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149–157 (2019).
19. J. S. Chen *et al.*, CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. *Science* **360**, 436–439 (2018).
20. M. I. Mustafa, A. M. Makhawi, SHERLOCK and DETECTR: CRISPR-Cas systems as potential rapid diagnostic tools for emerging infectious diseases. *J. Clin. Microbiol.* **59**, e00745-e20 (2021).
21. C. Camacho *et al.*, BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
22. R. C. Edgar, PILER-CR: Fast and accurate identification of CRISPR repeats. *BMC Bioinformatics* **8**, 18 (2007).
23. V. Burrus, G. Pavlovic, B. Decaris, G. Guédon, Conjugative transposons: The tip of the iceberg. *Mol. Microbiol.* **46**, 601–610 (2002).
24. A. M. Osborn, D. Böltner, When phage, plasmids, and transposons collide: Genomic islands, and conjugative- and mobilizable-transposons as a mosaic continuum. *Plasmid* **48**, 202–212 (2002).
25. G. Faure *et al.*, CRISPR-Cas in mobile genetic elements: Counter-defence and beyond. *Nat. Rev. Microbiol.* **17**, 513–525 (2019).
26. NCBI Resource Coordinators, Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **46**, D8–D13 (2018).
27. A. L. Mitchell *et al.*, MGnify: The microbiome analysis resource in 2020. *Nucleic Acids Res.* **48**, D570–D578 (2020).
28. UniProt Consortium, UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
29. E. V. Koonin, K. S. Makarova, Y. I. Wolf, M. Krupovic, Evolutionary entanglement of mobile genetic elements and host defence systems: Guns for hire. *Nat. Rev. Genet.* **21**, 119–131 (2020).
30. A. R. Parks, J. E. Peters, Tn7 elements: Engendering diversity from chromosomes to episomes. *Plasmid* **61**, 1–14 (2009).
31. Z. Li, H. Zhang, R. Xiao, L. Chang, Cryo-EM structure of a type I-F CRISPR RNA guided surveillance complex bound to transposition protein TniQ. *Cell Res.* **30**, 179–181 (2020).
32. M. F. Rollins, J. T. Schuman, K. Paulus, H. S. T. Bukhari, B. Wiedenheft, Mechanism of foreign DNA recognition by a CRISPR RNA-guided surveillance complex from *Pseudomonas aeruginosa*. *Nucleic Acids Res.* **43**, 2216–2222 (2015).
33. L. K. Arciszewska, N. L. Craig, Interaction of the Tn7-encoded transposition protein TnsB with the ends of the transposon. *Nucleic Acids Res.* **19**, 5021–5029 (1991).
34. K. Y. Choi, Y. Li, R. Sarnovsky, N. L. Craig, Direct interaction between the TnsA and TnsB subunits controls the heteromeric Tn7 transposase. *Proc. Natl. Acad. Sci. U.S.A.* **110**, E2038–E2045 (2013).
35. R. Pinilla-Redondo *et al.*, Type IV CRISPR-Cas systems are highly diverse and involved in competition between plasmids. *Nucleic Acids Res.* **48**, 2000–2012 (2020).
36. H. N. Taylor *et al.*, Positioning diverse type IV structures and functions within class 1 CRISPR-Cas systems. *Front. Microbiol.* **12**, 671522 (2021).
37. J. Strecker, A. Ladha, K. S. Makarova, E. V. Koonin, F. Zhang, Response to comment on “RNA-guided DNA insertion with CRISPR-associated transposases”. *Science* **368**, eabb2920 (2020).
38. G. Y. Kholodii *et al.*, Four genes, two ends, and a res region are involved in transposition of Tn5053: A paradigm for a novel family of transposons carrying either a mer operon or an integron. *Mol. Microbiol.* **17**, 1189–1200 (1995).
39. K. Y. Choi, J. M. Spencer, N. L. Craig, The Tn7 transposition regulator TnsC interacts with the transposase subunit TnsB and target selector TnsD. *Proc. Natl. Acad. Sci. U.S.A.* **111**, E2858–E2865 (2014).
40. J.-U. Park, *et al.*, Structural basis for target site selection in RNA-guided DNA transposition systems. bioRxiv [Preprint] (2021). https://doi.org/10.1101/2021.05.25.445634 (Accessed 25 May 2021).
41. J. E. Peters, “Tn7” in *Mobile DNA III*, N. L. Craig, Ed. (John Wiley & Sons, Ltd, 2015), pp. 647–667.
42. A. E. Stellwagen, N. L. Craig, Avoiding self: Two Tn7-encoded proteins mediate target immunity in Tn7 transposition. *EMBO J.* **16**, 6823–6834 (1997).
43. Z. Skelding, J. Queen-Baker, N. L. Craig, Alternative interactions between the Tn7 transposase and the Tn7 target DNA binding protein regulate target immunity and transposition. *EMBO J.* **22**, 5904–5917 (2003).
44. R. Bainton, P. Gamas, N. L. Craig, Tn7 transposition in vitro proceeds through an excised transposon intermediate generated by staggered breaks in DNA. *Cell* **65**, 805−816 (1991).
45. A. W. Kingston, C. Ponkratz, E. A. Raleigh, Rpn (YhgA-like) proteins of *Escherichia coli* K-12 and their contribution to RecA-independent horizontal transfer. *J. Bacteriol.* **199**, e00787-16 (2017).
46. A. K. Aggarwal, Structure and function of restriction endonucleases. *Curr. Opin. Struct. Biol.* **5**, 11–19 (1995).
47. A. B. Hickman *et al.*, Unexpected structural diversity in DNA recombination: The restriction endonuclease connection. *Mol. Cell* **5**, 1025–1034 (2000).
48. K. Steczkiewicz, A. Muszewska, L. Knizewski, L. Rychlewski, K. Ginalski, Sequence, structure and functional diversity of PD-(D/E)XK phosphodiesterase superfamily. *Nucleic Acids Res.* **40**, 7016–7045 (2012).
49. R. Missich *et al.*, The replisome organizer (G38P) of *Bacillus subtilis* bacteriophage SPP1 forms specialized nucleoprotein complexes with two discrete distant regions of the SPP1 genome. *J. Mol. Biol.* **270**, 50–64 (1997).
50. T. Jacobsen *et al.*, Characterization of Cas12a nucleases reveals diverse PAM profiles between closely-related orthologs. *Nucleic Acids Res.* **48**, 5624–5638 (2020).
51. R. T. Leenay *et al.*, Identifying and visualizing functional PAM diversity across CRISPR-Cas systems. *Mol. Cell* **62**, 137–147 (2016).
52. T. Yamano *et al.*, Crystal structure of Cpf1 in complex with guide RNA and target DNA. *Cell* **165**, 949–962 (2016).
53. I. Fonfara, H. Richter, M. Bratovič, A. Le Rhun, E. Charpentier, The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. *Nature* **532**, 517–521 (2016).
54. Y. Jeon *et al.*, Direct observation of DNA target searching and cleavage by CRISPR-Cas12a. *Nat. Commun.* **9**, 2777 (2018).
55. D. Singh *et al.*, Real-time observation of DNA target interrogation and product release by the RNA-guided endonuclease CRISPR Cpf1 (Cas12a). *Proc. Natl. Acad. Sci. U.S.A.* **115**, 5444–5449 (2018).
56. I. Strohkendl, F. A. Saifuddin, J. R. Rybarski, I. J. Finkelstein, R. Russell, Kinetic basis for DNA target specificity of CRISPR-Cas12a. *Mol. Cell* **71**, 816–824.e3 (2018).
57. D. C. Swarts, M. Jinek, Mechanistic insights into the cis- and trans-acting DNase activities of Cas12a. *Mol. Cell* **73**, 589–600.e4 (2019).
58. S. A. Shmakov *et al.*, Systematic prediction of functionally linked genes in bacterial and archaeal genomes. *Nat. Protoc.* **14**, 3013–3031 (2019).

59. 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).
60. D. Gleditzsch *et al.*, Modulating the Cascade architecture of a minimal Type I-F CRISPR-Cas system. *Nucleic Acids Res.* **44**, 5872–5882 (2016).
61. K. Kuznedelov *et al.*, Altered stoichiometry *Escherichia coli* Cascade complexes with shortened CRISPR RNA spacers are capable of interference and primed adaptation. *Nucleic Acids Res.* **44**, 10849–10861 (2016).
62. I. Songailiene *et al.*, Decision-making in Cascade complexes harboring crRNAs of altered length. *Cell Rep.* **28**, 3157–3166.e4 (2019).
63. D. Tuminauskaite *et al.*, DNA interference is controlled by R-loop length in a type I-F1 CRISPR-Cas system. *BMC Biol.* **18**, 65 (2020).
64. M. L. Luo *et al.*, The CRISPR RNA-guided surveillance complex in Escherichia coli accommodates extended RNA spacers. *Nucleic Acids Res.* **44**, 7385–7394 (2016).
65. V. Hoikkala, Cooperation between different CRISPR-Cas types enables adaptation in an RNA-targeting system. *mBio* **12**, e03338-20 (2021).
66. J. N. Vink, J. H. Baijens, S. J. Brouns, Comprehensive PAM prediction for CRISPR-Cas systems reveals evidence for spacer sharing, preferred strand targeting and conserved links with CRISPR repeats. bioRxiv [Preprint] (2021). https://doi.org/10.1101/2021.05.04.442622 (Accessed 6 May 2021).
67. G. Amitai, R. Sorek, CRISPR-Cas adaptation: Insights into the mechanism of action. *Nat. Rev. Microbiol.* **14**, 67–76 (2016).
68. B. D. Lavoie, G. Chaconas, Site-specific HU binding in the Mu transpososome: Conversion of a sequence-independent DNA-binding protein into a chemical nuclease. *Genes Dev.* **7** , 2510–2519 (1993).
69. M. Steinegger, J. Söding, MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
70. A. Hill, J. Rybarski, K. Hu, I. Finkelstein, C. Wilke, Opfi: A Python package for identifying gene clusters in large genomics and metagenomics data sets. *J. Open Source Softw.* **6**, 3678 (2021).