11670–11681 *Nucleic Acids Research, 2022, Vol. 50, No. 20* *Published online 17 November 2022*
https://doi.org/10.1093/nar/gkac985

Functional characterization of diverse type I-F CRISPR-associated transposons

Avery Roberts [1,2], Matthew A. Nethery[1,2] and Rodolphe Barrangou [1,2,*]

[1]Genomic Sciences Graduate Program, North Carolina State University, Raleigh, NC 27695, USA and [2]Department of Food, Bioprocessing and Nutrition Sciences, North Carolina State University, Raleigh, NC 27695, USA

Received August 25, 2022; Revised October 07, 2022; Editorial Decision October 12, 2022; Accepted October 18, 2022

ABSTRACT

CRISPR-Cas systems generally provide adaptive immunity in prokaryotes through RNA-guided degradation of foreign genetic elements like bacteriophages and plasmids. Recently, however, transposon-encoded and nuclease-deficient CRISPR-Cas systems were characterized and shown to be co-opted by Tn7-like transposons for CRISPR RNA-guided DNA transposition. As a genome engineering tool, these CRISPR-Cas systems and their associated transposon proteins can be deployed for programmable, site-specific integration of sizable cargo DNA, circumventing the need for DNA cleavage and homology-directed repair involving endogenous repair machinery. Here, we selected a diverse set of type I-F3 CRISPR-associated transposon systems derived from *Gammaproteobacteria*, predicted all components essential for transposition activity, and deployed them for functionality testing within *Escherichia coli*. Our results demonstrate that these systems possess a significant range of integration efficiencies with regards to temperature, transposon size, and flexible PAM requirements. Additionally, our findings support the categorization of these systems into functional compatibility groups for efficient and orthogonal RNA-guided DNA integration. This work expands the CRISPR-based toolbox with new CRISPR RNA-guided DNA integrases that can be applied to complex and extensive genome engineering efforts.

INTRODUCTION

CRISPR-Cas systems are often described as DNA-encoded, RNA-mediated, nucleic acid-targeting adaptive immune systems in prokaryotes (1–3). In a deviation from this canonical paradigm, however, nuclease deficient CRISPR-Cas systems were recently identified and shown to be co-opted by Tn7-like transposons for CRISPR RNA-guided DNA transposition (4) (Figure 1A). To date, multiple distinct types of CRISPR-associated transposons (CAST) have been experimentally characterised: type I-F3 (5), type V-K (6) and types I-B1 and I-B2 (7). The transposon-associated type I-F3 and I-B systems utilize the Cascade (CRISPR-associated complex for antiviral defense) effector complex (2) and naturally lack the Cas3 helicase-nuclease typically responsible for DNA cleavage in type I systems (8), while type V-K systems utilize a Cas12k naturally lacking nuclease functionality. All three types of CAST systems use a combination of CRISPR-Cas machinery and transposon proteins for RNA-guided DNA transposition with mechanisms and outcomes similar to those of the relatively well-characterized Tn7.

The prototypical Tn7 transposon possesses core transposition machinery (TnsABC) and target site selection proteins (TnsD and TnsE) that facilitate mobilization via two targeting pathways (9). TnsABC interacts with either TnsD for vertical transmission to a recognized *att*Tn7 site in *glmS*, or with TnsE, which recognizes replicating DNA for transposition to conjugal plasmid-based loci for horizontal transmission between bacteria (10,11). The transposase subunit TnsA and DDE-type integrase TnsB form a heteromeric transposase complex responsible for DNA breakage and joining characteristic of Tn7 transposition, while the ATPase TnsC interacts with TnsD or TnsE and the TnsAB transposase to drive transposition at a target site (12). TnsB of the TnsAB transposase associates with TnsB binding sites present within the transposon ends, resulting in a preferential orientation for integration (13) (Figure 1B). In contrast to prototypical Tn7 transposition, CASTs utilize CRISPR-Cas machinery for RNA-guided target site selection. Type I-F3 CASTs, specifically, possess Cascade and the TnsD homolog TniQ, which complexes with Cascade and facilitates interactions between Cascade and the transposition machinery (14) (Figure 1C). Recent work showed that DNA binding by the type I-F3 CAST TniQ-Cascade complex is relatively promiscuous throughout the genome, and that sequential recruitment and assembly of a heptameric TnsC ring and the TnsAB transposase likely act to regulate highly specific transposition (15). Type I-B CASTs and at least one type I-F3 CAST have been shown

*To whom correspondence should be addressed. Tel: +1 919 513 1644; Email: rbarran@ncsu.edu

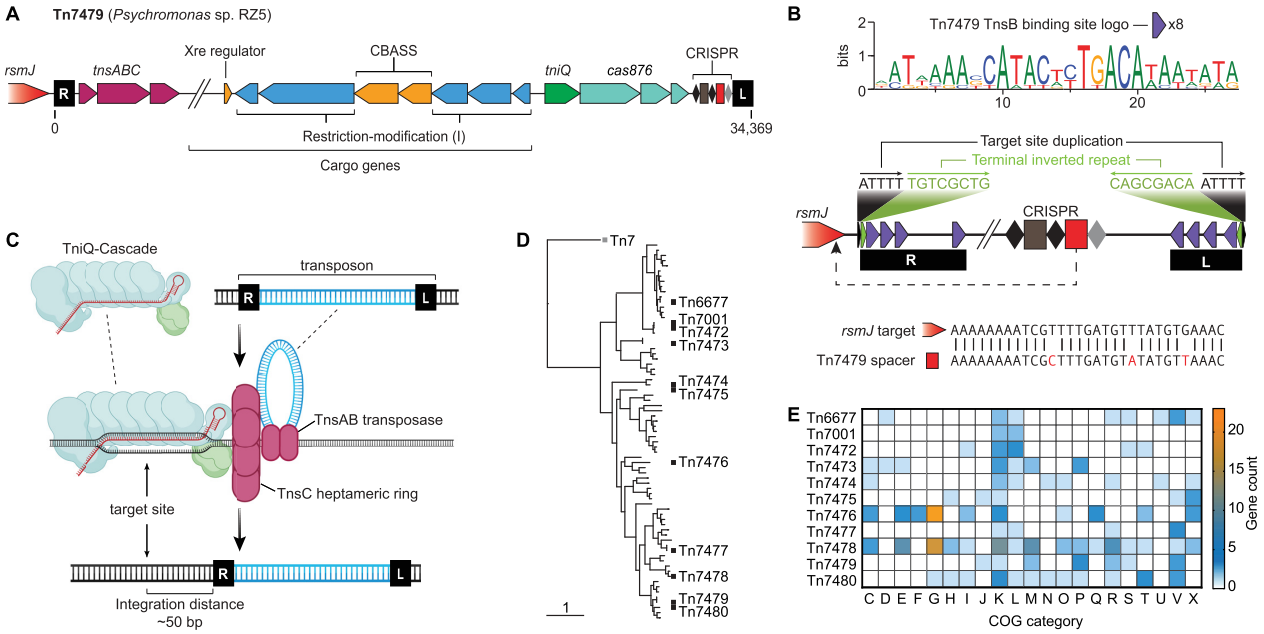

Figure 1. Form and function of Type I-F3 CRISPR-associated transposons. (**A**) Tn7479 is shown as a representative Type I-F3 CAST. Genes essential for transposition, cargo genes not essential for transposition, the *att* site (*rsmJ*), and transposon ends (R and L) are labeled. The atypical repeat of the CRISPR array is colored light grey. (**B**) A detailed view of the transposon ends and CRISPR array of Tn7479 in (A). The eight putative TnsB binding site sequences were used to generate a consensus WebLogo (top). 5-bp target site duplication events (black) indicative of Tn7-like transposition events and the terminal inverted repeats (green) that define the ends of the transposon are shown alongside the TnsB binding sites (purple arrows) within the transposon ends. The CRISPR array contains a self-targeting spacer that is complementary to a region of the *att* site (*rsmJ*) and flanked by an atypical repeat (light grey diamond). The *rsmJ* target site and self-targeting spacer possess mismatches colored in red. (**C**) A mechanistic overview of Type I-F3 CRISPR RNA-guided DNA transposition. The TniQ-Cascade complex is guided to the target site complementary to the spacer sequence of the bound crRNA. TnsB proteins bind to sites present in the transposon ends and transposition, regulated by TnsC activity between the TniQ-Cascade complex and the heteromeric transposase TnsAB, results in integration of the transposon ∼50 bp downstream from the end of the target site. (**D**) A phylogenetic tree of representative TnsB clades is shown with Tn7 as an outgroup and eleven diverse Type I-F3 CASTs selected for characterization. (**E**) The cargo genes of the eleven Type I-F3 CASTs selected for characterization were used to generate a heatmap displaying the gene counts for each Clusters of Orthologous Genes (COG) category.

to utilize both TniQ and TnsD for RNA-dependent and RNA-independent transposition pathways, respectively (7,16).

As genome engineering tools, CASTs offer great promise for site-specific integration of large DNA payloads. Already, the type I-F3 CAST Tn6677 has been streamlined into single-plasmid forms and used for multiplexed genome engineering and for species- and site-specific genome editing in a mixed community context (17,18). Still, there is a need for an expanded genome editing toolbox with highly efficient and orthogonal type I-F3 CASTs to achieve complex genome engineering goals while circumventing issues such as target site immunity and transposon remobilization. Though the type V-K CAST of *Scytonema hofmannii* (ShCAST) was shown to be orthogonal with the type I-F3 Tn6677, type V-K systems suffer from relatively poor specificity and frequent cointegration events (7,17). To support this effort, we selected a diverse set of type I-F3 CAST homologs and characterized their RNA-guided DNA integration activity with regards to different temperatures, transposon sizes, and target sites possessing distinct protospacer adjacent motif (PAM) sequences (19,20). Additionally, we explored the CAST homologs in their native context and identified diverse cargo gene content and instances of type I-F3 CASTs with multiple self-targeting spacers.

MATERIALS AND METHODS

Type I-F3 CRISPR-associated transposon feature prediction and system selection

Bacterial genomes from the RefSeq database were run through CRISPRclassify to predict CRISPR loci associated with type I CRISPR-Cas systems (21,22). Regions 50-kb in size flanking the putative type I CRISPR loci were extracted using custom scripts. The extracted regions were made into a BLAST nucleotide database and the Tn6677 TniQ, Cas8, Cas7, Cas6, TnsA, TnsB and TnsC protein sequences (5) were used as a query for a translated BLAST (tblastn) search against the database using a specified *E*-value 1e–3 parameter (23). The tblastn results were filtered based on the presence of hits for multiple query proteins within a single extracted region, and the associated genomes were manually examined for expected *tnsABC* and *tniQ-cas876* operon configurations and proximity in Geneious Prime (v2020.1). Additional candidate genomes were selected from recent publications (24,25). Candidate systems from all sources were further filtered based on sequence similarity of proteins essential for transposition. Other features used to filter candidate systems were gene completeness, identifiable self-targeting spacers and CRISPR arrays, occasionally encompassing atypical repeat(s), and predicted transposon ends containing terminal inverted repeats and

TnsB binding sites. CRISPR arrays were predicted in candidate genomes using CRISPRclassify and further manually curated in Geneious to confirm the presence of a self-targeting spacer. Using Geneious, self-targeting spacers were aligned to the gene immediately upstream from *tnsA* of the associated system to identify the attachment (*att*) site/protospacer (target site with complementarity to the self-targeting spacer). The presence of a TGT or TGA trinucleotide motif ∼50 bp downstream from the identified protospacer initiated the manual prediction of elements associated with Tn7-like transposon ends (4). The TGT or TGA motif and additional flanking nucleotides were extracted, including 5-bp regions upstream and downstream corresponding, respectively, to the 5-bp target site duplication (TSD) event and remaining 5 bp of the 8-bp terminal inverted repeat (TIR) that begins with the TGT or TGA motif. Genomes were searched for sequences consisting of the reverse complement of the 8-bp TIR immediately followed by the 5-bp TSD. The original *tnsA*-proximal TSD/TIR region and regions containing no or few mismatches from the expected TIR/TSD sequence were then further examined for the presence of ∼3–4 repetitive ∼18-bp TnsB binding site sequences based on known binding site configurations common for the *tnsA*-proximal (right) and non-*tnsA*-proximal (left) ends of Tn7-like transposons (4). Right and left transposon ends were typically defined as ∼150–200-bp regions and used for plasmid design and construction. Eventually, 10 candidate systems were selected based on confidence in the prediction of all essential elements for reconstitution for experimental testing (Table 1, Supplementary Table S1).

Protein comparisons

TnsB protein sequences were extracted from hundreds of previously identified type I-F3 CASTs from recent work (24) and representative TnsB clades were made using CD-HIT with a 90% sequence similarity cut-off (26). TnsB protein sequences from Tn7 and the tested candidate systems were aligned alongside the CD-HIT results using FastTree in Geneious with default parameters. The resulting tree was visualized and annotated using the Interactive Tree Of Life (iTOL) (27) (Figure 1D). Using protein sequences from each tested system, single-protein alignments for TnsA, TnsB, TnsC, TniQ, Cas8, Cas7 and Cas6 were performed using MUSCLE (28) in Geneious with 8 iterations and default parameters and then visualized as a heatmap (Supplementary Figure S2A). A heatmap was generated from an additional alignment performed using TnsB protein sequences from type I-F3 CASTs characterized in a recent study (16) in addition to the TnsB protein sequences derived from systems tested in this study (Supplementary Figure S2B).

Type I-F3 CRISPR-associated transposon genomic analyses

The genomes associated with Tn6677 (*V. cholerae* HE-45) and the ten additional selected candidate type I-F3 CASTs were downloaded from GenBank. Each genome was run through Prokka (v1.14.5) (29) for gene prediction and the resulting output files were used as inputs for further analyses. Bacterial antiviral defence systems were predicted using PADLOC (30) and DefenseFinder (31) and the results were filtered to only include systems within known or predicted transposon boundaries of the tested CAST candidates. To functionally classify the cargo genes, custom scripts were used with the translated CAST cargo genes as queries for RPS-BLAST + against the NCBI Conserved Domain Database (CDD). Clusters of orthologous genes (COG) categories were then assigned to the top BLAST hit per protein and the data was visualized as a heatmap (32) (Supplementary Table S6). Using the associated GenBank files, the predicted CASTs from *Neptunomonas qingdaonensis* CGMCC 1.10971 (Tn7475) and *Neptunomonas japonica* DSM 18939 were merged into single contigs and then aligned and visualized using Clinker with a 30% protein similarity cut off for visually linked genes (33).

Plasmid construction

The base pEffector expression vector was designed to express the CRISPR array and genes essential for transposition from a single T7 promoter and includes a *lac* operator and T7 terminator for gene expression and regulation. A multiple cloning site was included downstream of the *lac* operator for downstream cloning purposes. The base pEffector expression vector was created using the backbone of pSL0284 (Addgene #130635) and a gBlock containing a multiple cloning site. The pSL0284 backbone and multiple cloning site fragments were generated as PCR products using Q5 High-Fidelity 2X Master Mix (NEB #M0492) and assembled using NEBuilder HiFi DNA Assembly Master Mix (NEB #E2621). Further pEffector derivatives were generated by restriction digestion of the pEffector plasmid followed by ligation of annealed oligos (IDT), containing either a CRISPR array or a single spacer sequence, with overhangs complementary to those generated by restriction digestion. pDonor constructs were made by amplifying a backbone from pSL0527 (Addgene #130634) and system-specific gBlocks comprised of a multiple cloning site flanked by the right and left transposon ends. The pSL0527 backbone and system-specific fragments were generated as PCR products using Q5 High-Fidelity 2X Master Mix (NEB #M0492) and assembled using NEBuilder HiFi DNA Assembly Master Mix (NEB #E2621). Further pDonor derivatives were generated by inserting additional DNA fragments within the multiple cloning site flanked by the transposon ends to generate pDonor constructs with transposons of approximately 1 kb or 10 kb in size. Plasmids used in this study are listed in Supplementary Table S2.

Transposition experiments

Transposition experiments were performed using chemically competent (NEB #C2527H) or electrocompetent (Millipore Sigma #CMC0016) *Escherichia coli* BL21(DE3) cells. For transposition experiments involving a pDonor construct with a donor transposon of approximately 1 kb in size, pEffector and pDonor constructs were co-transformed into chemically competent cells by using 50 ng of each plasmid. For transposition experiments involving pDonor constructs with a donor transposon of ∼10 kb in size, 50 ng

Table 1. Data for type I-F3 CRISPR-associated transposons used in this study

Tn#	Type	Genus	Species	Strain	Tn length (bp)	*att* site
Tn6677	I-F3	*Vibrio*	*cholerae*	HE-45	36,168	*guaC*
Tn7001	I-F3	*Photobacterium*	*iliopiscarium*	ATCC 51760	28,940	*guaC*
Tn7472	I-F3	*Photobacterium*	*piscicola*	NCCB 100098	31,741	*guaC*
Tn7473	I-F3	*Aliiglaciecola*	sp.	M165	26,436	Thioesterase family protein
Tn7474	I-F3	*Halomonas*	*titanicae*	BH1	38,808	*ffs*
Tn7475	I-F3	*Neptunomonas*	*qingdaonensis*	CGMCC 1.10971	≥30,165	*ffs*
Tn7476	I-F3	*Klebsiella*	*oxytoca*	67	83,370	tRNA-Ser
Tn7477	I-F3	*Photobacterium*	*aquae*	CGMCC 12159	19,556	*ffs*
Tn7478	I-F3	*Vibrio*	sp.	EJY3	117,187	AraC family protein
Tn7479	I-F3	*Psychromonas*	sp.	RZ5	34,369	*rsmJ*
Tn7480	I-F3	*Colwellia*	*polaris*	MCCC 1C00015	50,364	*rsmJ*

of each plasmid was co-transformed via electroporation into electrocompetent cells using the manufacturer's recommended settings. Cells were plated on two-antibiotics (Spectinomycin 50 μg/ml, Carbenicillin 100 μg/ml) LB-agar plates with 0.1 mM IPTG (Thermo Scientific #R1171) and 40 μg/ml X-gal (Thermo Scientific #R0941). Incubations were performed at 30°C for 30 hours or at 25°C for 48 h. Experiments pertaining to system orthogonality or involving pDonor constructs harboring a donor transposon ∼10 kb in size were incubated at 25°C. Following incubation, sample lysates were generated similarly to previous work (5). For each sample, colonies were scraped and resuspended in PBS, then OD_{600} measurements were taken and approximately 3.2×10^8 cells were transferred to a 1.5 ml microtube and resuspended in PBS for a total volume of 200 μl for downstream processing. The resuspended cells were pelleted by centrifugation, supernatant was decanted, and the pellets were resuspended in 80 μl of nuclease-free H_2O. The cells were then lysed on a dry block incubator at 95°C for 10 min and cooled to room temperature. Cellular debris was pelleted by centrifugation and the lysate, containing genomic DNA, was diluted 20-fold in nuclease-free H_2O in a fresh microtube for further use.

qPCR analysis

Three primer pairs were used for quantification of integration events within a sample. One primer pair served to generate a reference amplicon from the *rssA* gene in the *E. coli* BL21(DE3) genome, while the other two primer pairs were designed to generate amplicons specific to integrated transposons in either the RL or LR orientation near the target site. The integration-specific primers were benchmarked using isolated and serially diluted gDNA from clonal RL or LR integrants by individual systems to ensure adequate PCR efficiency between systems (Supplementary Figure S1). Each qPCR reaction was 10 μl in total volume and consisted of: 5 μl of SsoAdvanced Universal SYBR Green Supermix (Bio-Rad), 2 μl of 2.5 μM primer mix of two primers, 2 μl of a 20-fold dilution of sample lysate, and 1 μl of nuclease-free H_2O. Reactions were loaded in 96-well qPCR plates (Thermo Scientific #AB3396) and run on a CFX Connect Real-Time PCR Detection System (Bio-Rad). The cycling conditions consisted of an initial denaturation step for 2.5 min at 98°C, followed by 40 amplification cycles of 10 sec at 98°C and 1 min at 62°C, followed by melt curve production with 5 s per 0.5°C step for a range of 65–95°C. Samples were from three biological replicates, and three technical replicates were performed per sample, per primer pair. Integration efficiency was calculated using the $2^{\Delta Cq}$ method where, per sample, the Cq of the reference reaction is subtracted from the Cq value of either integration-specific reaction. The two calculated efficiencies are then combined to calculate the total integration efficiency. To determine RL:LR orientation bias, the calculated integration efficiency for the RL orientation was divided by the calculated integration efficiency for the LR orientation. Primers used in this study are listed in Supplementary Table S3.

Plasmid library assay for PAM determination

The plasmid library (pTarget) with a randomized 5N region was created as previously described (34). The pTarget plasmid library was co-transformed into electrocompetent *E. coli* BL21(DE3) cells alongside pEffector and pDonor constructs by using 100 ng each of the pTarget plasmid library, pEffector and pDonor constructs. Cells were plated on triple antibiotic (Spectinomycin 50 μg/ml, Carbenicillin 100 μg/ml, Kanamycin 50 μg/ml) LB-agar plates with 0.1 mM IPTG (Thermo Scientific #R1171) and incubated at 30°C for 30 h. Following incubation, colonies were scraped and plasmid DNA was extracted with a MidiPrep Kit (QIAGEN). PCR products were generated using extracted plasmid DNA as a template with Q5 High-Fidelity 2X Master Mix (NEB #M0492) and 0.5 μM of a primer pair designed to amplify across the 5N region of the pTarget plasmid library with a forward primer specific to the pTarget plasmid library and a reverse primer specific to transposon integration in the RL orientation. PCR thermocycling involved an annealing temperature of 67°C and 25 cycles and PCR products were visualized on 1% agarose gel using SYBR Safe (Thermo Scientific). PCR products were isolated and purified using Monarch DNA Gel Extraction and PCR & DNA Cleanup kits (NEB), and NGS libraries were prepared and sequenced on an Illumina platform for 2 × 250 bp paired-end reads (GENEWIZ Amplicon-EZ). The resulting sequence data was analyzed using SeqKit (35). Reads were filtered based on the presence of no mismatches, when compared to the base pTarget plasmid library, within the 20 bp upstream and 32 bp downstream (spacer sequence) from the 5N region. The 5N region was then extracted from these filtered reads and PAM frequencies were calculated and normalized based on PAM frequencies calculated for the base pTarget plasmid library:

((Count of PAM X)/(Total PAM count))/((Frequency of PAM X in base library)/(Expected frequency of PAM X in base library)). The top 10% of normalized PAM frequencies for each system were used to generate WebLogos (36). The comprehensive normalized PAM data was used to generate PAM wheels in the form of Krona plots (37). PAM wheels represent frequency data for positions –3, –2 and –1 of the variable 5-nucleotide PAM region of the pTarget plasmid library, and PAM wheel segments were colored grey if the corresponding trinucleotide motifs were not representative of >1% of the total data. Library construction oligos and primers are listed in Supplementary Table S3.

Integration distance analyses

For *lacZ* integration experiments, PCR products were generated using diluted sample lysates, Q5 High-Fidelity 2X Master Mix (NEB #M0492), and 0.5 μM of primers specific to RL orientation transposon integration events. Thermocycling conditions involved an annealing temperature of 68°C and 20 cycles and PCR products were visualized on 1% agarose gel using SYBR Safe (Thermo Scientific). PCR products were isolated and purified using Monarch DNA Gel Extraction and PCR & DNA Cleanup kits (NEB), and NGS libraries were prepared and sequenced on an Illumina platform for 2 × 250 bp paired-end reads (GENEWIZ Amplicon-EZ). The resulting sequence data was analyzed using SeqKit (35). Reads were filtered based on 12 perfectly matching bases specific to the right transposon end, followed by extracting the 20 bases immediately upstream of the 5′ end of the integrated right transposon end. After filtering for sequences only 20 bp in length, the remaining sequences were mapped back to *lacZ* using Bowtie (38) with no mismatches allowed. Positional data corresponding to the integration distance of the transposon from the target site was extracted from the resulting .sam files and integration distances comprising at least 1% of the data set were plotted.

Nanopore sequencing and assembly

Individual colonies from the transposition experiments involving 10 kb transposons were restreaked for isolated clones on LB plates supplemented with antibiotics and X-gal. Isolates were scraped, cultured overnight at 30°C, and used for gDNA extraction with the DNeasy PowerLyzer Microbial Kit (QIAGEN). The isolated gDNA was used as a template for confirmatory PCR using Q5 High-Fidelity 2X Master Mix (NEB #M0492) and primer pairs specific to transposon integration in the RL or LR orientation and a primer pair flanking the integration site. Template gDNA that generated an amplicon from the flanking primer pair and an amplicon specific to only one integrated transposon orientation was then used for Nanopore sequencing. Genomic DNA was prepped with a Nanopore Ligation Sequencing kit (SQK-LSK109), barcoded using a Native Barcoding Expansion kit (EXP-NBD104), and sequenced on a FLO-MIN111 flow cell run on a MinION Mk1B device. Basecalling was performed within MinKNOW using the super-accurate basecalling model and a minimum qscore cutoff of 10. Reads at least 10 kb in length were used for *de novo* genome assembly using Flye (v2.9) with the following parameters: –nano-hq, –asm-coverage 100, –genome-size 4.6m, and –iterations 2 for two rounds of polishing (39). Reads used for assembly were mapped to the resulting Flye assemblies to generate coverage graphs in Geneious. Integration verification primers are listed in Supplementary Table S3.

RESULTS

Prediction and genomic characterization of type I-F3 CRISPR-associated transposons

The Tn7-like transposon Tn6677, derived from *Vibrio cholerae* HE-45, was the first characterized type I-F3 CAST and shown to transpose DNA in a CRISPR RNA-guided manner (5). Given the complexity of CRISPR RNA-guided DNA transposition (5,14), we anticipated natural functional variation among diverse CASTs, similar in manner to variation observed among Cas9 orthologs (40). To understand the functional breadth of type I-F3 CASTs, we selected ten diverse and previously uncharacterized systems from *Gammaproteobacteria* for functional characterization (Table 1, Figure 1D). Features necessary for transposition were predicted *in silico*, namely the genes *tnsA*, *tnsB* (or *tnsAB*), *tnsC*, *tniQ*, *cas8*, *cas7* and *cas6*, the CRISPR array and repeats flanking a self-targeting spacer, and the transposon ends that define the transposon boundaries due to the presence of a terminal inverted repeat (TIR) and TnsB binding sites. This set of type I-F3 CASTs represents significant diversity with regards to protein identity, origin, taxonomy, native transposon size, and attachment site. Tn7479 of *Klebsiella oxytoca* 67, for example, possesses a natural TnsAB fusion and a tRNA-Ser attachment site, a notable feature given tRNA attachment sites are common for type V-K CASTs, but relatively rare for type I-F3 systems (7,24). We initially explored the CASTs within their native context by analyzing the genes non-essential for transposition activity, otherwise referred to as cargo genes, present within the boundaries of each CAST. Similar analyses were performed using PADLOC (30) and DefenseFinder (31) to identify putative antiviral defense systems present within the cargo genes (Supplementary Tables S4 and S5). The results from both platforms were congruent and, interestingly, several recently defined antiviral systems, including Gabija (41) and qatABCD (42), were present among this small subset of CASTs. Using the cargo-derived protein sequences, we also identified functional domains and assigned COG categories to generally assess their biological roles (32) (Figure 1E, Supplementary Table S6). We found a striking diversity in functional categories for the cargo genes and many systems possessed proteins related to transcription (COG category K) and as previously noted, defense mechanisms (V). Noteworthy, our two largest CASTs, Tn7476 and Tn7478, each possessed high counts of genes associated with energy production and conversion (C) and carbohydrate transport and metabolism (G). These results highlight the role of CASTs as mobile genetic elements that harbor relatively rare antiviral defense systems and confer other beneficial features, such as expanded metabolic capabilities and resistances, to the host (16,43).

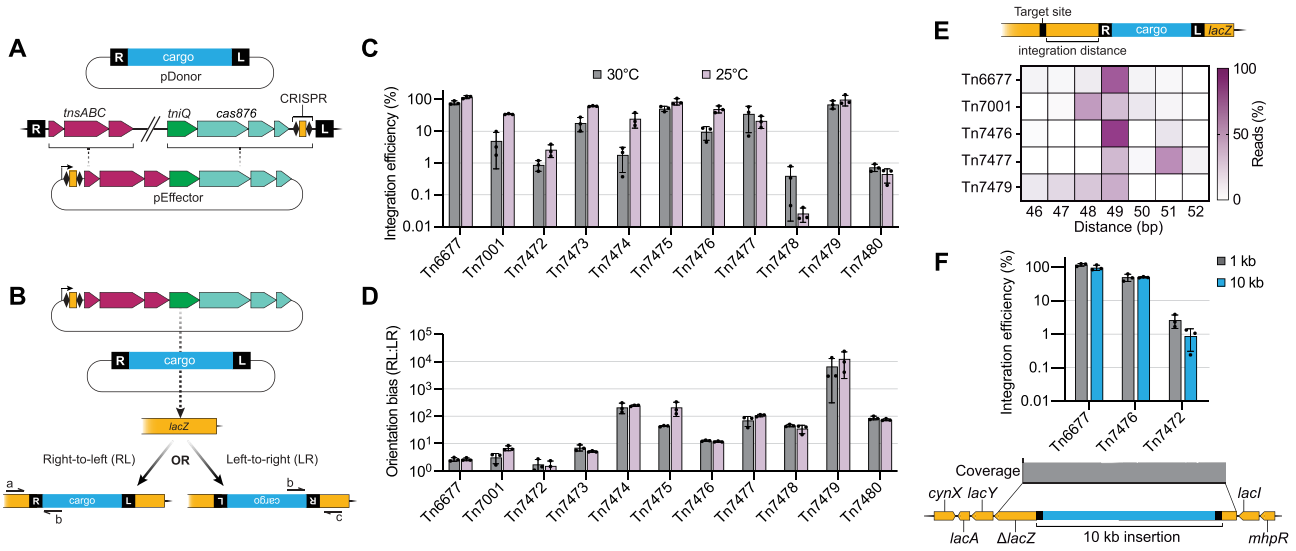

Figure 2. Reconstitution and characterization of RNA-guided DNA integration using a diverse set of Type I-F3 CRISPR-associated transposons. (**A**) Diagram of the reconstitution of each CAST into a two-plasmid system. A pEffector plasmid contains a CRISPR array, *tnsABC*, *tniQ*, and *cas876* from the associated transposon. A pDonor plasmid contains cargo DNA flanked by the predicted right and left ends (R and L) of the associated transposon. (**B**) A *lacZ*-targeting spacer present within the pEffector CRISPR array drives RNA-guided DNA integration into *lacZ*. The integration event results in a transposon integrated in either the right-to-left (RL) or left-to-right (LR) orientation. Two primer pairs (ab, bc) were used to quantify integration events by qPCR. (**C**) RNA-guided DNA integration efficiency was measured by qPCR for eleven CASTs at two different temperatures, 30°C and 25°C, at a single *lacZ* target site with a 5′-TCC-3′ PAM. (**D**) Orientation bias data related to (C). (**E**) Integration distance data, related to the target site from (C), is shown. Integration distance is the number of base pairs between the end of the target site and the start of the integrated transposon. (**F**) Integration efficiency involving transposons approximately 1 kb and 10 kb in size. Coverage is shown from Nanopore sequencing of genomic DNA derived from an isolate containing the integrated 10 kb transposon. The expanded genomic context is illustrated below. For parts C, D and F, shown are the mean, SD, and individual data points from $n = 3$ biological replicates.

Validation of ten type I-F3 CRISPR-associated transposons for RNA-guided DNA integration

Each CAST was reconstituted as a two-plasmid system for testing in *E. coli* BL21(DE3). A pEffector plasmid that contains the CRISPR array and all genes essential for transposition and a pDonor plasmid that contains a donor transposon of approximately 1 kb or 10 kb in length, including transposon ends (referred to as 'right' and 'left' ends) (Figure 2A). Integration activity was measured using qPCR for transposition assays in *E. coli* BL21(DE3), capturing transposition products integrated in a right-to-left (RL) or left-to-right (LR) orientation at a given target site (Figure 2B). Initial CAST characterization involved targeting a *lacZ* locus previously tested extensively with Tn6677 (5), which possesses a 5′-CC-3′ PAM canonical for type I-F3 CRISPR-Cas systems. Given that this target site and PAM were previously shown to enable efficient RNA-guided DNA transposition with Tn6677, we tested our ten candidate CASTs at this target site at both 30°C and 25°C incubation temperatures (Figure 2C). Interestingly, in these conditions we found that integration efficiencies for these CASTs had unexpected variability, with integration efficiencies ranging from approximately 0.1% to 100%. Overall, transposition efficiencies generally increased, if not remained comparable, when the incubation temperature was reduced from 30°C to 25°C. We note that an elongated window of opportunity for integration to occur and potentially alleviated cytotoxicity due to slower growth rates at a lower incubation temperature may contribute to increased integration efficiencies at 25°C. However, some increases in efficiency at 25°C may be due to enhanced kinetics of the CRISPR and transposition molecular machinery at lower temperatures, as many type I-F3 CASTs are found in isolates derived from aquatic environments with naturally lower ambient temperatures (i.e. <30°C). As the total integration efficiency is the result of combining the quantification of right-to-left and left-to-right orientation transposition events, we can compare the quantification of each orientation to determine transposon orientation bias (Figure 2D). We found that, at a single target site with a 5′-CC-3′ PAM, these systems possess orientation biases universally skewed toward the RL orientation, with biases ranging from slightly greater than 1:1 (RL:LR orientation bias) to greater than 1000:1, and that these biases remained generally consistent regardless of incubation temperature. In addition, we tested randomized, non-targeting spacers for each system and did not detect any integration events at the *lacZ* locus using our methods. Overall, we show that our ten previously uncharacterized CASTs are capable of RNA-guided DNA transposition, with integration orientation preference universally biased toward the RL orientation, at a single target site with a 5′-CC-3′ PAM.

For select CASTs, we generated integration distance data, where integration distance is defined as the number of nucleotides between the end of the target site and the beginning of the transposon end sequence (Figure 2E). Tn6677 was previously shown to integrate DNA with an integration distance most frequently centered around 49 bp across multiple target sites (16). For this target site with a 5′-CC-3′ PAM, we found that integration distance for these di-

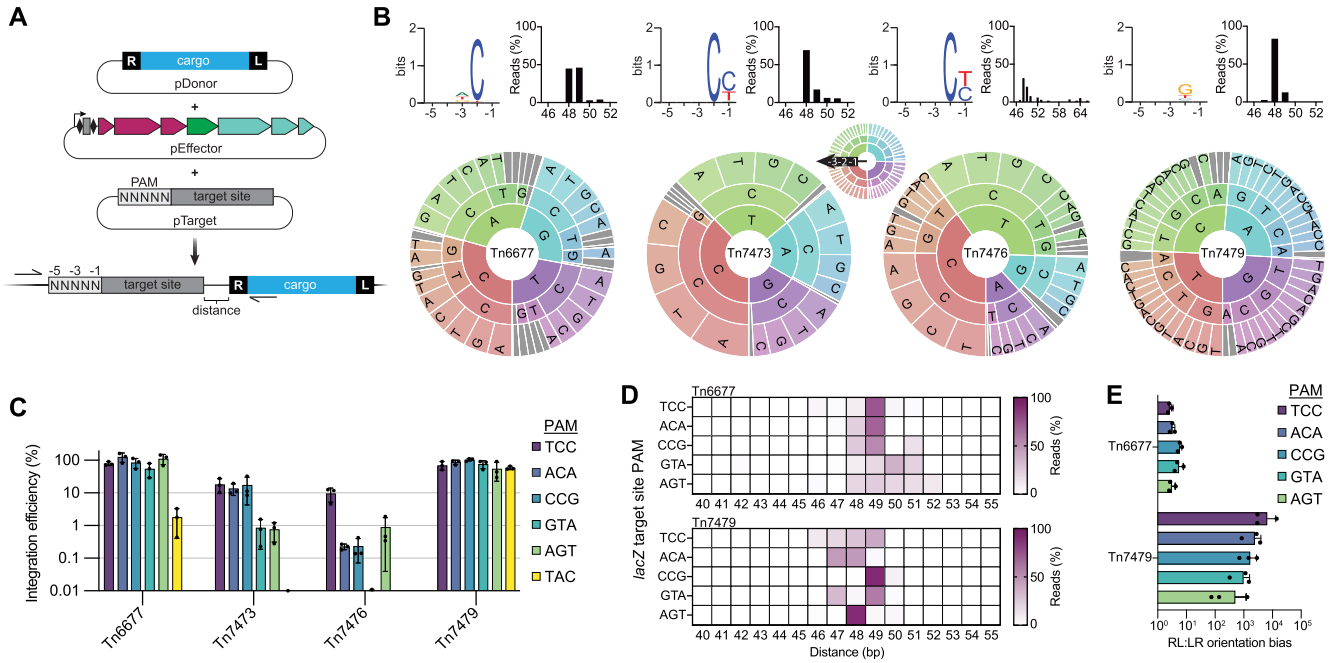

Figure 3. Assessment of PAM preferences and RNA-guided DNA integration activity at multiple target sites. (**A**) Illustration of the plasmid library-based approach to PAM determination. A pDonor plasmid is co-transformed with pEffector and the pTarget library. The pEffector plasmid contains a spacer complementary to a target site of pTarget, which possesses a randomized 5-nucleotide region immediately upstream of the target site, serving as the variable PAM region. Primers shown were used to capture integration events and amplicons were deep sequenced and analyzed for PAM frequency and integration distance data. (**B**) For each shown CAST, the top 10% most frequent PAMs are represented as WebLogos. Integration distance in base pairs is shown. PAM wheels were generated using the comprehensive PAM data, with the outer and inner rings corresponding to the –3 and –1 positions of the PAM, respectively. Grey color is applied to any PAM that did not represent >1% of the dataset. (**C**) RNA-guided DNA integration measured by qPCR at multiple target sites with distinct PAM sequences. (**D**) For Tn6677 and Tn7479, integration distance data is shown for target sites in (C). (**E**) For Tn6677 and Tn7479, orientation bias is shown for target sites in (C). For parts C and E, shown are the mean, SD and individual data points from $n = 3$ biological replicates.

verse CASTs is centered around 49 bp, but the integration distance patterns vary among systems. While Tn6677 and Tn7476 largely integrated 49 bp from the target site, Tn7477 integrated DNA almost evenly between the two distances of 49 and 51 bp and Tn7479 integrated with greater frequency as the distance increased in the 46–49 bp range. Our data shows that integration distance for type I-F3 CASTs is variable between systems at a single target site using a consistent spacer length (32 bp) for all systems.

Next, we aimed to determine whether transposon size has a significant effect on integration efficiency for select systems. CAST Tn7476, which possesses a natural TnsAB fusion, and Tn7472, which had relatively low integration efficiencies in our initial tests, were chosen for testing with pDonor plasmids containing a transposon ∼10 kb in size. The same *lacZ* target site possessing a 5′-CC-3′ PAM was used as in our initial tests, and incubation occurred at 25°C to alleviate potential cytotoxicity and maximize integration efficiency. We found that integration efficiencies were similar when comparing integration involving transposons of approximately 1 kb and 10 kb in size (Figure 2F), highlighting the potential for efficient integration of large payloads regardless of transposon size. Indeed, the upper limit for CRISPR RNA-guided DNA transposition is likely context-dependent and may be limited more so by construct generation and delivery methods than transposition mechanisms, especially given that type I-F3 CASTs naturally occur >100 kb in length, suggesting cargo size is not material.

PAM determination and DNA integration activity at multiple target sites

As we saw a wide range of integration efficiency across our CASTs for a single target site with a 5′-CC-3′ PAM, we wanted to determine whether PAM preference varies between our CASTs and if those variations in preference have a significant impact on integration activity. To determine PAM preference for each CAST, we conducted a plasmid library-based assay in which the pEffector plasmid for a given system bears a spacer to direct integration within the pTarget plasmid library at a target site with a randomized five-nucleotide (5N) region upstream (5′ end) of the target site (Figure 3A). For each CAST, we used PCR to generate an amplicon specific to RL orientation integration events within the plasmid library, deep sequenced the amplicon, and analyzed the nucleotide frequencies within the variable region to determine PAM preference. Interestingly, we discovered that the PAM profiles for these systems are generally consistent (Figure 3B; Supplementary Figure S3). For most systems, we found a strong preference for a C nucleotide in the –2 position, and we note that the –2 position is repeatedly the position displaying the most significant degree of preference within these PAM profiles. We also found that the canonical 5′-CC-3′ PAM (for positions –2 and –1) preference is only present within a few systems, and that the preference for a C in the –1 position is lesser than the preference for a C in the –2 position. The most strik-

ing results from this assay was the discovery that Tn7479 possesses near PAM-less functionality, showing only a very slight preference in the –2 position for a G nucleotide, which may be attributed to the motif frequency within the base plasmid library. Integration distance data generated from the plasmid library assay revealed similar findings to our initial data from the *lacZ* target site with a 5′-CC-3′ PAM, as integration was largely centered around ∼49 bp. Tn7476, however, had a wider integration distance range as integration events were detected >60 bp from the target site. This deviation from a tighter integration distance window may be influenced by the natural TnsAB fusion of Tn7476, but further investigation is required.

To validate our findings within the plasmid library assay, we next selected five additional target sites within *lacZ*, each possessing a distinct PAM sequence (Figure 3C; Supplementary Figure S4A). Two of the additional targets retain a C in the –2 position of the PAM, while the remaining three targets instead have a T, G or A at the –2 position. This set of target sites allowed us to evaluate the extent of preferences toward a canonical 5′-CC-3′ PAM or the less restrictive 5′-CN-3′ PAM that we found to be more applicable to most of our CASTs. We discovered that the differences in PAM flexibility initially seen in our plasmid library assay do translate to differences in integration efficiency when targeting different genomic loci. We found that only one system, Tn7476, relies heavily on a C in the –1 position, as it integrates most effectively at the target site with a 5′-CC-3′ PAM and shows a severe drop in efficiency when the –2 position C is retained but the –1 position C is replaced with an A or a G. Additionally, integration with Tn7476 was nearly or entirely undetectable when the –2 position in the PAM contained a T or A. Other CASTs, such as Tn7473 and Tn7474, have similar integration activity at all three 5′-CN-3′ PAM targets, but efficiency noticeably declines when the –2 C of the PAM is replaced with a T, G or A. Interestingly, most systems show near or entirely undetectable integration activity for the target site possessing a 5′-AC-3′ PAM. This bias against a 5′-AC-3′ PAM is likely to mitigate self-targeting activity within the CRISPR array due to the 5′-AN-3′ motif often present at the 3′ end of the CRISPR repeats associated with these systems (Supplementary Figure S5A). Excitingly, we validated the plasmid library assay results that highlighted Tn7479 as an effectively PAM-free system, exhibiting PAM flexibility similar to the type I-F3 CASTs of a previous study (44). Regardless of the target site, Tn7479 exhibited highly efficient integration activity and, notably, Tn7479 integrated DNA with significantly higher efficiency than Tn6677 at the target site with a 5′-TAC-3′ PAM.

Though we previously showed that integration distance varies between CASTs at a single target site, we wanted to explore the variation in integration distance between two systems at the same set of target sites to determine whether individual CASTs have identifiable preferences regarding integration distance across target sites (Figure 3D). We generated integration distance data for Tn6677 and Tn7479 for five of the target sites used for integration assays. We found that the integration distance pattern remains strongly centered around ∼49 bp for both systems, however the integration distance is spread unpredictably when looking at multiple targets for a single system or when comparing the two systems at a single target. However, we do note that the integration distance for Tn6677 tends to skew toward >49 bp for these sites, while the Tn7479 integration distance tends to be <49 bp. Overall, we show that integration distance is often largely present in a ∼5 bp window for a single target site, but no system has a single, defined integration distance. Additionally, based on current knowledge, there is no way to predict the integration distance pattern for a system apart from a ∼49 bp approximation. These results align with recent work that revealed TnsC assembly at a target site explains the strict ∼50 bp integration distance for Tn6677, though structural and functional variations between homologous proteins may explain deviations from the defined integration distance for Tn6677 (15) (Supplementary Figure S2A). We note that the PCR-based method for detection means we may have missed distant integration events that occurred outside of the range defined by the primers used and that PCR-free methods may result in more clearly defined integration patterns. As mentioned above, we can parse out transposition orientation bias data from the sum integration efficiency data. For the additional integration at different targets within *lacZ*, we found that orientation bias is generally consistent between targets for any given system (Figure 3E; Supplementary Figure S4B). Our original bias data is supported here, as we found a range of bias from ∼1:1 to >1000:1 for RL:LR orientation events. Indeed, the highly skewed bias of Tn7479 was present for DNA integration at all target sites.

Type I-F3 CRISPR-associated transposon orthogonality

Recent work revealed that type I-F3 CASTs can be generally categorized into compatibility groups based on the orthogonality resulting from the TnsB binding sites present within the transposon ends (16). Significantly diverged type I-F3 CASTs are unable to mobilize each other's associated transposon based on the inability for one system's TnsB protein to recognize the TnsB binding sites within the transposon ends of the other systems. We used this recent data to predict compatibility groups for two of our novel CASTs and performed orthogonality testing by using all nine pEffector-pDonor combinations for the three selected systems, including Tn6677 (Figure 4A). The CASTs tested were selected based on the associated TnsB binding sites (Figure 4B, Supplementary Figure S5B) that putatively classified the systems into three distinct compatibility groups. Though full orthogonality was not always observed between members of different compatibility groups (16), we only detected integration activity with pEffector-pDonor combinations where both plasmids were associated with the same CAST (Figure 4C). This data supports the use of Tn6677, Tn7477 and Tn7479 as orthogonal RNA-guided DNA integrases and additionally supports the general classification scheme of type I-F3 CASTs into functional compatibility groups based on TnsB binding site divergence (Supplementary Figure S2B). These orthogonal systems can be used for sequential integration events at the same or distance target sites to, respectively, circumvent issues arising from target site immunity and transposon remobilization.

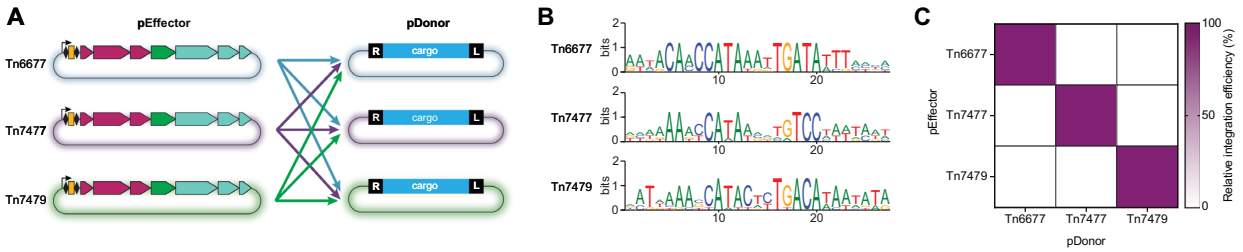

Figure 4. Orthogonality testing of three Type I-F3 CRISPR-associated transposons. (**A**) The orthogonality testing scheme is illustrated. The pEffector plasmid from each CAST was cotransformed with each of the three possible pDonor plasmids for a total of nine pEffector-pDonor combinations. (**B**) WebLogos were generated using 27-bp regions predicted to contain the ∼20-bp TnsB binding sites for each system. (**C**) Relative RNA-guided DNA integration efficiency, as measured by qPCR, for the testing scheme shown in (A).

Type I-F3 CRISPR-associated transposons with multiple self-targeting spacers

During the CAST candidate curation process, we discovered a unique instance of two self-targeting spacers among a selected candidate system, Tn7475, derived from a *Neptunomonas qingdaonensis* strain. We examined additional *Neptunomonas* genomes and identified a second instance of a CAST with two self-targeting spacers in a *N. japonica* strain (Figure 5A). In both instances, two self-targeting spacers were identified with complementarity to the same region of the *ffs* attachment site previously identified as one of four major attachment sites for type I-F3 CASTs (24). Mismatches between these self-targeting spacers and their target site are generally present at 6-nt intervals within the spacer sequence, following a previously identified prevalent pattern of mismatch tolerance for type I-F3 CASTs (24) (Figure 5A). We found no significant similarity between the non-self-targeting spacers of the two systems and were unable to confidently identify any protospacers corresponding to these non-self-targeting spacers. We also failed to detect any candidate Xre transcription factor binding sites adjacent to the CRISPR arrays in either system, in agreement with regulation patterns seen in other systems of the type I-F3b branch (24). Alignment of the CRISPR repeats from Tn7475 revealed that the atypical repeats associated with the self-targeting spacers are distinct from one another in terms of sequence and putative crRNA structure. For instance, one atypical repeat (R4) from Tn7475 contains a shifted stem loop structure with a shorter 4-nt loop, while the other atypical repeats (R5, R6) have highly disrupted putative stem-loop structures (Figure 5B). Given the presence of multiple self-targeting spacers and their associated atypical repeats in Tn7475, we wanted to determine if one set of repeats enables more efficient RNA-guided DNA integration, as previous work has shown that privatized crRNAs incorporating atypical repeats generally result in more efficient DNA integration (16,24). Intriguingly, we discovered that Tn7475 utilizing the first two repeats of the CRISPR array, which flank a non-self-targeting spacer, resulted in the most efficient integration activity at a *lacZ* target site (Figure 5C). Finally, we performed a genomic comparison using the predicted *Neptunomonas* type I-F3 CASTs, highlighting the split *tnsB* and *cas8* genes of the *N. japonica* CAST, and found no significant overlap between cargo genes of the two CASTs aside from the transposition-essential genes and an expected Xre family transcription factor (Figure 5D).

DISCUSSION

We identified and characterized diverse type I-F3 CASTs and provide evidence that all of our selected systems are active for RNA-guided DNA transposition. In their native context, these systems may span over 100 kb in size and possess diverse cargo genes that likely confer a multitude of benefits to the recipient host, including an expanded antiviral defense system arsenal and metabolic potential. Our selected systems are capable of DNA integration at a wide range of maximum efficiencies across multiple target sites, and at least one system (Tn7479) possesses remarkably flexible PAM requirements for efficient integration. To further test the PAM flexibility of Tn7479, additional target sites with less frequent PAMs from the PAM determination assay could be validated to identify relative rates of integration. Additional testing is required to confirm the specificity of a system with such relaxed PAM requirements, as well as to confirm the expected high frequency of simple insertions events rather than cointegrate products. Type I-F3 CASTs with highly flexible PAM requirements, such as PtrCAST, Tn7007 and Tn7016, have been shown to retain high specificity within *E. coli*, though stringent target selection criteria may need to be applied when expanding the applications of these flexible systems to avoid off-target integration events (16,44).

We also explored integration distance patterns for these systems at multiple target sites and found that, currently, integration distance is largely confined to a small ∼5-bp range based on a central distance of 49 bp. Given the variation in integration distance, these genome editing tools cannot currently be used for integration events requiring precise, nucleotide-level insertions at a fixed distance from the target site. Still, many genome engineering goals, especially those incorporating large DNA insertions, are not hindered by such limitations, and a recent study applied ShCAST to ultra-long DNA insertions for metabolic engineering (45). We note that our PCR-based detection method may have failed to capture distant integration events, and further studies are required to elucidate the mechanisms behind integration distance variation with regards to system, target site, transposon size, TnsB binding site identity and configuration, and biological context.

Recent work categorized type I-F3 CASTs into functional compatibility groups based on TnsB binding site sequence divergence (16), and our orthogonality testing results support this classification scheme by providing two systems that are fully orthogonal with Tn6677. Though

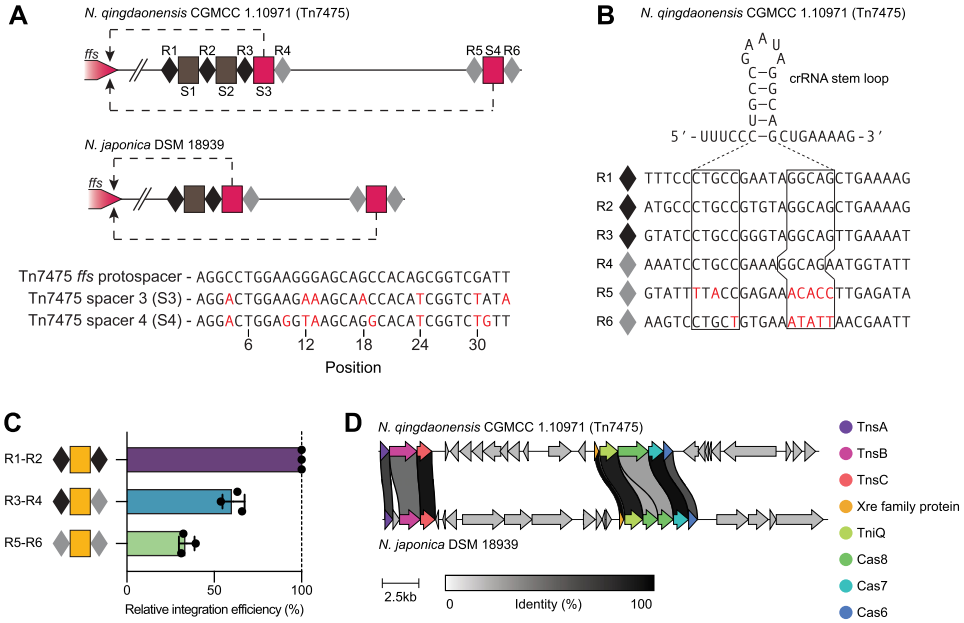

Figure 5. Instances and characterization of Type I-F3 CRISPR-associated transposons with two self-targeting spacers. (**A**) Two *Neptunomonas* strains each possess a type I-F3 CAST with two self-targeting spacers with complementarity to the *ffs att* site. Atypical repeats are colored light grey, the *N. qingdaonensis* repeats and spacers are numbered (R#, S#), and the mismatches between the two self-targeting spacers and the *ffs* target site are colored red. Non-self-targeting spacers and self-targeting spacers are shown as brown and pink rectangles, respectively. (**B**) The Type I-F3 crRNA stem loop structure typically forms from CUGCC/GGCAG RNA pairing. These nucleotides are boxed for each repeat and deviations from the expected CTGCC/GGCAG repeat DNA sequence are colored red. (**C**) Comparison of RNA-guided DNA integration efficiency, as measured by qPCR, using a *lacZ*-targeting spacer and repeat pairs comprised of typical and atypical repeats derived from *N. qingdaonensis* (Tn7475). (**D**) Genomic comparison between the two *Neptunomonas* CASTs shown in (A). Protein sequences with at least 30% identity are linked between the two CASTs with identity (%) shown. For part C, shown are the mean, SD and individual data points from $n = 3$ biological replicates.

these systems can be used for iterative DNA integration events without the issue of target site immunity or transposon remobilization, CRISPR repeat recognition orthogonality between systems must be assessed before any simultaneous integration events can be performed to ensure target sites remain specific to the intended system. Recent work regarding the structural basis of TnsB binding site recognition may also allow for refinement of the predicted compatibility groups (46). Even with orthogonal and multiplexed approaches, holistic consideration of these features will likely enable precise genome editing events with controlled integration copy number.

To our knowledge, we described the first instances of multiple self-targeting spacers in two distinct type I-F3 CASTs. Unexpectedly, both sets of repeats associated with the two self-targeting spacers of Tn7475 resulted in less efficient integration relative to integration involving the typical repeats of the CRISPR array, in contrast to previous work that has shown atypical repeats of type I-F3 systems privatize crRNAs and, in many cases, increase integration efficiency (16,47). Type I-F3 CASTs Tn6677 and Tn6900 were thoroughly tested with typical and atypical repeats and alterations thereof, however, none of the repeats tested had crRNA stem-loop structure disruptions as significant as those found in the atypical repeats of Tn7475 (Figure 5B) (47). Though the Tn7475 atypical repeat-derived crRNAs are likely privatized from canonical type I-F CRISPR-Cas systems, they may result in less efficient integration due to decreased rates of crRNA binding and processing by Cascade, caused by the severely altered stem-loop structure and divergence from the 5′ and 3′ handle sequences of the typical repeats. It is tempting to speculate about the role and source of these potentially redundant self-targeting spacers, but comprehensive bioinformatic analyses are required to further explore these systems and potentially uncover multiple self-targeting spacers within other types of CASTs and the evolutionary pressures likely to have influenced such features.

Overall, we validated ten previously uncharacterized type I-F3 CASTs alongside Tn6677, and further expand the recently established CAST toolbox (16). Our findings expand the repertoire of available efficient, flexible, and orthogonal CASTs for complex genome engineering efforts encompassing large cargo insertions, including multiplexed editing strategies, complex mixed bacterial communities, and, potentially, eukaryotic systems.

DATA AVAILABILITY

NGS data from this work has been deposited in the NCBI Sequence Read Archive (SRA) under BioProject ID PRJNA862512.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Samuel H. Sternberg for plasmids pSL0284 (Addgene #130635) and pSL0527 (Addgene #130634). The authors would also like to acknowledge all members of the CRISPR lab at North Carolina State University for helpful discussion related to the manuscript. Some figure elements were generated using BioRender.

FUNDING

Syngenta AG. Funding for open access charge: Syngenta AG.

Conflict of interest statement. R.B. is a cofounder of Intellia Therapeutics, Locus Biosciences, TreeCo, CRISPR Biotechnologies and Ancilia Biosciences, and a shareholder of Caribou Biosciences, Inari Ag, Felix Biotechnologies and Provaxus. A.R. and R. B. are co-inventors on a related patent application.

REFERENCES

1. Barrangou,R., Fremaux,C., Deveau,H., Richards,M., Boyaval,P., Moineau,S., Romero,D.A. and Horvath,P. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, **315**, 1709–1712.
2. Brouns,S.J.J., Jore,M.M., Lundgren,M., Westra,E.R., Slijkhuis,R.J.H., Snijders,A.P.L., Dickman,M.J., Makarova,K.S., Koonin,E.V. and van der Oost,J. (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science*, **321**, 960–964.
3. Marraffini,L.A. and Sontheimer,E.J. (2008) CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science*, **322**, 1843–1845.
4. Peters,J.E., Makarova,K.S., Shmakov,S. and Koonin,E.V. (2017) Recruitment of CRISPR-Cas systems by Tn7-like transposons. *Proc. Natl Acad. Sci. U.S.A.*, **114**, E7358–E7366.
5. Klompe,S.E., Vo,P.L.H., Halpin-Healy,T.S. and Sternberg,S.H. (2019) Transposon-encoded CRISPR–Cas systems direct RNA-guided DNA integration. *Nature*, **571**, 219–225.
6. Strecker,J., Ladha,A., Gardner,Z., Schmid-Burgk,J.L., Makarova,K.S., Koonin,E.V. and Zhang,F. (2019) RNA-guided DNA insertion with CRISPR-associated transposases. *Science*, **365**, 48–53.
7. Saito,M., Ladha,A., Strecker,J., Faure,G., Neumann,E., Altae-Tran,H., Macrae,R.K. and Zhang,F. (2021) Dual modes of CRISPR-associated transposon homing. *Cell*, **184**, 2441–2453.
8. Sinkunas,T., Gasiunas,G., Fremaux,C., Barrangou,R., Horvath,P. and Siksnys,V. (2011) Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. *EMBO J*, **30**, 1335–1342.
9. Peters,J.E. and Craig,N.L. (2001) Tn7 recognizes transposition target structures associated with DNA replication using the DNA-binding protein tnsE. *Genes Dev.*, **15**, 737–747.
10. Waddell,C.S. and Craig,N.L. (1988) Tn7 transposition: two transposition pathways directed by five Tn7-encoded genes. *Genes Dev.*, **2**, 137–149.
11. Wolkow,C.A., DeBoy,R.T. and Craig,N.L. (1996) Conjugating plasmids are preferred targets for tn7. *Genes Dev.*, **10**, 2145–2157.
12. Peters,J.E. and Craig,N.L. (2001) Tn7: smarter than we thought. *Nat. Rev. Mol. Cell Biol.*, **2**, 806–814.
13. Choi,K.Y., Li,Y., Sarnovsky,R. and Craig,N.L. (2013) Direct interaction between the TnsA and TnsB subunits controls the heteromeric Tn7 transposase. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, E2038–E2045.
14. Halpin-Healy,T.S., Klompe,S.E., Sternberg,S.H. and Fernández,I.S. (2020) Structural basis of DNA targeting by a transposon-encoded CRISPR–Cas system. *Nature*, **577**, 271–274.
15. Hoffmann,F.T., Kim,M., Beh,L.Y., Wang,J., Vo,P.L.H., Gelsinger,D.R., George,J.T., Acree,C., Mohabir,J.T., Fernández,I.S. *et al.* (2022) Selective TnsC recruitment enhances the fidelity of RNA-guided transposition. *Nature*, **609**, 384–393.
16. Klompe,S.E., Jaber,N., Beh,L.Y., Mohabir,J.T., Bernheim,A. and Sternberg,S.H. (2022) Evolutionary and mechanistic diversity of type I-F CRISPR-associated transposons. *Mol. Cell*, **82**, 616–628.
17. Vo,P.L.H., Ronda,C., Klompe,S.E., Chen,E.E., Acree,C., Wang,H.H. and Sternberg,S.H. (2020) CRISPR RNA-guided integrases for high-efficiency, multiplexed bacterial genome engineering. *Nat. Biotechnol.*, **39**, 480–489.
18. Rubin,B.E., Diamond,S., Cress,B.F., Crits-Christoph,A., Lou,Y.C., Borges,A.L., Shivram,H., He,C., Xu,M., Zhou,Z. *et al.* (2022) Species- and site-specific genome editing in complex bacterial communities. *Nat. Microbiol.*, **7**, 34–47.
19. Deveau,H., Barrangou,R., Garneau,J.E., Labonté,J., Fremaux,C., Boyaval,P., Romero,D.A., Horvath,P. and Moineau,S. (2008) Phage response to CRISPR-Encoded resistance in streptococcus thermophilus. *J. Bacteriol.*, **190**, 1390–1400.
20. Mojica,F.J.M., Díez-Villaseñor,C., García-Martínez,J. and Almendros,C. (2009) Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology*, **155**, 733–740.
21. Nethery,M.A., Korvink,M., Makarova,K.S., Wolf,Y.I., Koonin,E.V. and Barrangou,R. (2021) CRISPRclassify: repeat-based classification of CRISPR loci. *CRISPR J.*, **4**, 558–574.
22. Makarova,K.S., Wolf,Y.I., Iranzo,J., Shmakov,S.A., Alkhnbashi,O.S., Brouns,S.J.J., Charpentier,E., Cheng,D., Haft,D.H., Horvath,P. *et al.* (2020) Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nat. Rev. Microbiol.*, **18**, 67–83.
23. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinf.*, **10**, 421.
24. Petassi,M.T., Hsieh,S.-C. and Peters,J.E. (2020) Guide RNA categorization enables target site choice in Tn7-CRISPR-Cas transposons. *Cell*, **183**, 1757–1771.
25. Rybarski,J.R., Hu,K., Hill,A.M., Wilke,C.O. and Finkelstein,I.J. (2021) Metagenomic discovery of CRISPR-associated transposons. *Proc. Natl. Acad. Sci. U.S.A.*, **118**, e2112279118.
26. Fu,L., Niu,B., Zhu,Z., Wu,S. and Li,W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
27. Letunic,I. and Bork,P. (2021) Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.*, **49**, W293–W296.
28. Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinf.*, **5**, 113.
29. Seemann,T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.
30. Payne,L.J., Todeschini,T.C., Wu,Y., Perry,B.J., Ronson,C.W., Fineran,P.C., Nobrega,F.L. and Jackson,S.A. (2021) Identification and classification of antiviral defence systems in bacteria and archaea with PADLOC reveals new system types. *Nucleic Acids Res.*, **49**, 10868–10878.
31. Tesson,F., Hervé,A., Mordret,E., Touchon,M., d'Humières,C., Cury,J. and Bernheim,A. (2022) Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat Commun*, **13**, 2561.
32. Galperin,M.Y., Wolf,Y.I., Makarova,K.S., Vera Alvarez,R., Landsman,D. and Koonin,E.V. (2021) COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res.*, **49**, D274–D281.
33. Gilchrist,C.L.M. and Chooi,Y.-H. (2021) clinker & clustermap.js: automatic generation of gene cluster comparison figures. *Bioinformatics*, **37**, 2473–2475.
34. Maxwell,C.S., Jacobsen,T., Marshall,R., Noireaux,V. and Beisel,C.L. (2018) A detailed cell-free transcription-translation-based assay to decipher CRISPR protospacer-adjacent motifs. *Methods*, **143**, 48–57.
35. Shen,W., Le,S., Li,Y. and Hu,F. (2016) SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One*, **11**, e0163962.
36. Crooks,G.E., Hon,G., Chandonia,J.-M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
37. Leenay,R.T., Maksimchuk,K.R., Slotkowski,R.A., Agrawal,R.N., Gomaa,A.A., Briner,A.E., Barrangou,R. and Beisel,C.L. (2016) Identifying and visualizing functional PAM diversity across CRISPR-Cas systems. *Mol. Cell*, **62**, 137–147.
38. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

39. Kolmogorov,M., Yuan,J., Lin,Y. and Pevzner,P.A. (2019) Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol.*, **37**, 540–546.
40. Gasiunas,G., Young,J.K., Karvelis,T., Kazlauskas,D., Urbaitis,T., Jasnauskaite,M., Grusyte,M.M., Paulraj,S., Wang,P.-H., Hou,Z. *et al.* (2020) A catalogue of biochemically diverse CRISPR-Cas9 orthologs. *Nat. Commun.*, **11**, 5512.
41. Doron,S., Melamed,S., Ofir,G., Leavitt,A., Lopatina,A., Keren,M., Amitai,G. and Sorek,R. (2018) Systematic discovery of antiphage defense systems in the microbial pangenome. *Science*, **359**, eaar4120.
42. Gao,L., Altae-Tran,H., Böhning,F., Makarova,K.S., Segel,M., Schmid-Burgk,J.L., Koob,J., Wolf,Y.I., Koonin,E.V. and Zhang,F. (2020) Diverse enzymatic activities mediate antiviral immunity in prokaryotes. *Science*, **369**, 1077–1084.
43. Benler,S., Faure,G., Altae-Tran,H., Shmakov,S., Zheng,F. and Koonin,E. (2021) Cargo genes of Tn7-Like transposons comprise an enormous diversity of defense systems, mobile genetic elements, and antibiotic resistance genes. *Mbio*, **12**, e0293821.
44. Yang,S., Zhang,Y., Xu,J., Zhang,J., Zhang,J., Yang,J., Jiang,Y. and Yang,S. (2021) Orthogonal CRISPR-associated transposases for parallel and multiplexed chromosomal integration. *Nucleic Acids Res.*, **49**, 10192–10202.
45. Cheng,Z.-H., Wu,J., Liu,J.-Q., Min,D., Liu,D.-F., Li,W.-W. and Yu,H.-Q. (2022) Repurposing CRISPR RNA-guided integrases system for one-step, efficient genomic integration of ultra-long DNA sequences. *Nucleic Acids Res.*, **50**, 7739–7750.
46. Kaczmarska,Z., Czarnocki-Cieciura,M., Górecka-Minakowska,K.M., Wingo,R.J., Jackiewicz,J., Zajko,W., Poznański,J.T., Rawski,M., Grant,T., Peters,J.E. *et al.* (2022) Structural basis of transposon end recognition explains central features of Tn7 transposition systems. *Mol. Cell*, **82**, 2618–2632.
47. Petassi,M.T., Hsieh,S.-C. and Peters,J.E. (2020) Guide RNA categorization enables target site choice in Tn7-CRISPR-Cas transposons. *Cell*, **183**, 1757–1771.