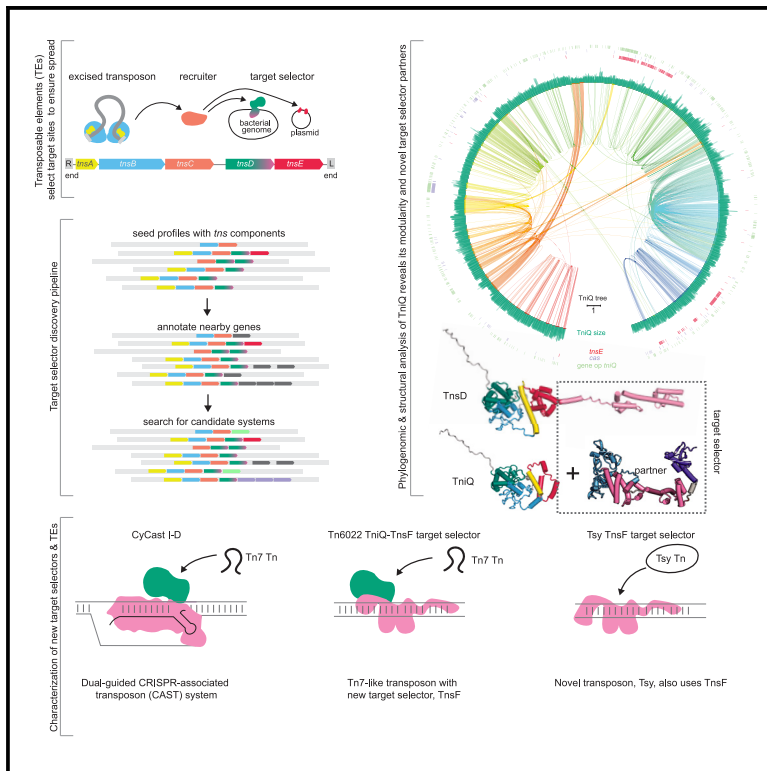


# Modularity and diversity of target selectors in Tn7 transposons

## Graphical abstract



## Authors

Guilhem Faure, Makoto Saito, Sean Benler, ..., Rhiannon K. Macrae, Eugene V. Koonin, Feng Zhang

## Correspondence

koonin@ncbi.nlm.nih.gov (E.V.K.), zhang@broadinstitute.org (F.Z.)

## In brief

Faure et al. use phylogenomic and structural approaches to systematically survey transposon target selectors, revealing a modular architecture that promotes transposon spread. Through this work, they identified a new target selector, TnsF, and a previously uncharacterized transposable element, Tsy.

## Highlights

- A broad survey of Tn7-like transposons clarifies their target selection strategies
- Type I-D CASTs transpose via both RNA-guided and protein-mediated mechanisms
- Transposon-encoded TnsF is a novel target selector that targets the *comM* gene
- Tsy is a tyrosine recombinase transposon that shares TnsF with Tn7-like Tn6022



Article

# Modularity and diversity of target selectors in Tn7 transposons

Guilhem Faure,<sup>1,2,3,4,5,7</sup> Makoto Saito,<sup>1,2,3,4,5,7</sup> Sean Benler,<sup>6,7</sup> Iris Peng,<sup>1,2,3,4,5,7</sup> Yuri I. Wolf,<sup>6</sup> Jonathan Strecker,<sup>1,2,3,4,5</sup> Han Altae-Tran,<sup>1,2,3,4,5</sup> Edwin Neumann,<sup>1,2,3,4,5</sup> David Li,<sup>1,2,3,4,5</sup> Kira S. Makarova,<sup>6</sup> Rhiannon K. Macrae,<sup>1,2,3,4,5</sup> Eugene V. Koonin,<sup>6,\*</sup> and Feng Zhang<sup>1,2,3,4,5,8,\*</sup>

<sup>1</sup>Howard Hughes Medical Institute, Cambridge, MA 02139, USA

<sup>2</sup>Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

<sup>3</sup>McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

<sup>4</sup>Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

<sup>5</sup>Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

<sup>6</sup>National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD 20894, USA

<sup>7</sup>These authors contributed equally

<sup>8</sup>Lead contact

\*Correspondence: [koonin@ncbi.nlm.nih.gov](mailto:koonin@ncbi.nlm.nih.gov) (E.V.K.), [zhang@broadinstitute.org](mailto:zhang@broadinstitute.org) (F.Z.)

<https://doi.org/10.1016/j.molcel.2023.05.013>

## SUMMARY

To spread, transposons must integrate into target sites without disruption of essential genes while avoiding host defense systems. Tn7-like transposons employ multiple mechanisms for target-site selection, including protein-guided targeting and, in CRISPR-associated transposons (CASTs), RNA-guided targeting. Combining phylogenomic and structural analyses, we conducted a broad survey of target selectors, revealing diverse mechanisms used by Tn7 to recognize target sites, including previously uncharacterized target-selector proteins found in newly discovered transposable elements (TEs). We experimentally characterized a CAST I-D system and a Tn6022-like transposon that uses TnsF, which contains an inactivated tyrosine recombinase domain, to target the *comM* gene. Additionally, we identified a non-Tn7 transposon, Tsy, encoding a homolog of TnsF with an active tyrosine recombinase domain, which we show also inserts into *comM*. Our findings show that Tn7 transposons employ modular architecture and co-opt target selectors from various sources to optimize target selection and drive transposon spread.

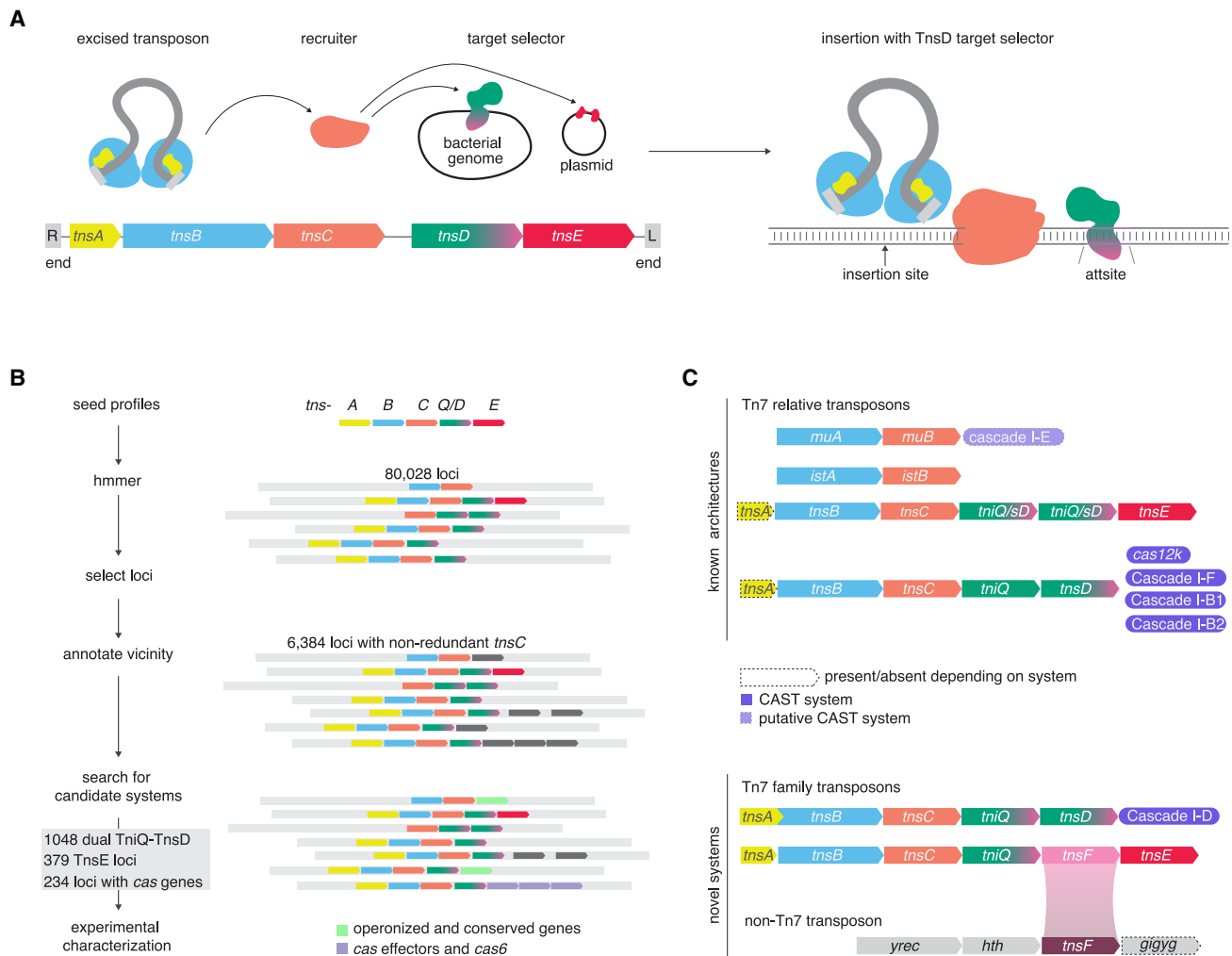
## INTRODUCTION

Transposable elements (TEs) are DNA sequences that can move around and across genomes, employ diverse molecular mechanisms to achieve mobility, and exhibit a broad range of targeting specificities.<sup>1</sup> Where a TE integrates is critical for its survival, and various strategies have evolved to select target sites, both for homing and for jumping to a mobile genetic element. For example, members of the Tn7 group of prokaryotic DNA transposons recognize (1) a highly conserved sequence of an essential gene to guide integration to a safe locus for homing and (2) a particular DNA conformation (which is agnostic to sequence) to guide integration to mobile elements at replication forks.<sup>2–4</sup> These two modes of target recognition are carried out by two dedicated proteins, TnsD (sequence-specific) and TnsE (structure-specific) (Figure 1A). In addition to these two target selectors, Tn7 encodes the heterocomplex TnsA/TnsB, which recognizes the ends of the transposons (TnsB) and excises the transposon (TnsA and TnsB), and TnsC, the central hub

component that coordinates the transpososome assembly with target-site selection.<sup>5</sup> TnsC recognizes TnsD bound to the attachment site, recruits the transpososome, and directs its integration.<sup>6</sup>

In addition to these canonical modes of target-site selection, several groups of Tn7-like transposons have co-opted CRISPR-Cas systems, enabling RNA-guided transposition. These CRISPR-associated transposons (CASTs) target mobile genetic elements (MGEs) using matching spacers encoded in the CRISPR array and Cas-effector components coupled with the small protein TniQ, a homolog of TnsD.<sup>7–12</sup> The CASTs target homing sites in two alternative modes, either by RNA-guided transposition or through TnsD, similarly to the canonical Tn7-like transposons.<sup>12</sup> The CASTs appear to have evolved as a result of the recruitment of CRISPR-Cas effector modules by Tn7-like transposons on multiple independent occasions.<sup>7,13</sup> Specifically, different groups of Tn7-like transposons acquired CRISPR subtype I-B (at least twice, independently) and subtype I-F and subtype V-K effectors. In each of these cases, the





**Figure 1. Prediction of novel target selectors in Tn7-like transposons**

(A) Schematic of Tn7 transposition. TnsB (cyan) recognizes both ends (R, right and L, left) and excises the transposon with the help of TnsA (yellow). TnsD is a sequence-specific target selector and binds an attachment site in the bacterial genome to recruit TnsC (orange) and the transposon for insertion.

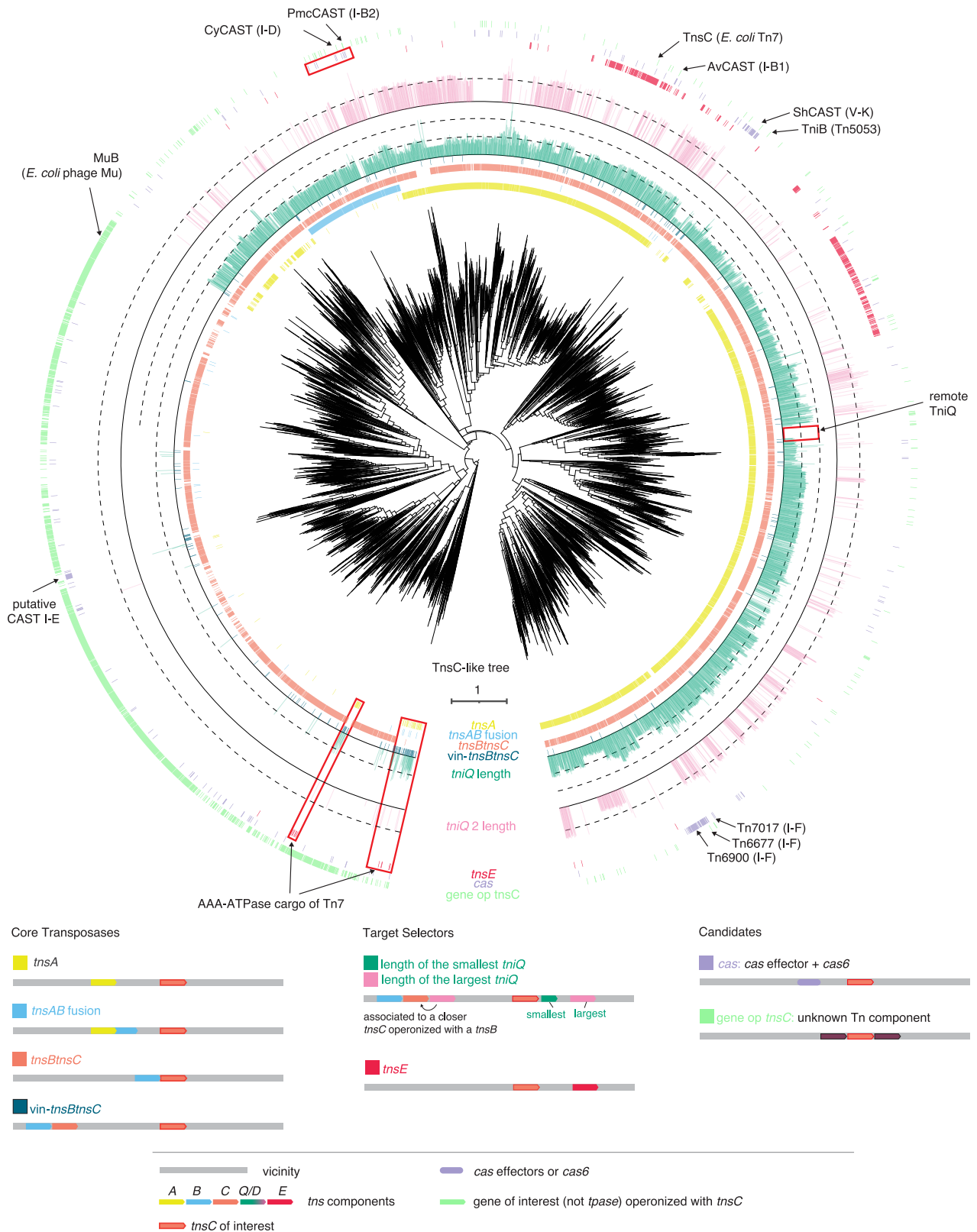
(B) Pipeline for discovery of novel target selectors. Sequence databases were mined for Tn7 component seeds and searched for genomic co-localization of these seeds. The genomic neighborhoods of the detected loci were annotated, with the focus on *cas* effectors and genes that appear to be operonized with *tnsC* and *tniQ/tnsD*.

(C) Locus architectures of known systems and novel systems identified in this study. Mu (*muA* and *muB*) and IS21 (*istA* and *istB*) encode relatives of TnsB and TnsC. IS21 has not been reported to be associated with a target selector. Tn7 encodes various target selectors including TniQ/TnsD, TnsE, and Cas effectors (Cas12k, Cascade I-F, and Cascade I-B, which all partner with TniQ), the latter of which constitute CAST systems. We found a novel CAST system containing Cascade I-D and a novel target selector we named TnsF. We also found a TnsF-like target selector in a distinct non-Tn7 transposon.

CRISPR effector module acquired by the TE retained the ability to recognize and bind the target DNA but lost the capacity of typical CRISPR systems to cleave the target DNA. In the case of type I CRISPR effectors, the elimination of the cleavage activity results from the loss of the Cas3 helicase-nuclease, whereas in the case of subtype V-K, it is due to the mutation of the catalytic amino acids in the active site of the RuvC-like nuclease.<sup>8,9</sup>

The existence of these distinct modes of target-site selection by Tn7-like transposons suggests a high degree of flexibility that maximizes their spread and highlights the utility of multiple, functionally orthogonal target-site selectors. Given this flexibility, comprehensive identification of target selectors is challenging.

We used a combination of phylogenomic and structural analyses to discover target selectors. Among these candidate target selectors, we identified and characterized three TE systems: a distinct CAST subtype I-D, a Tn7-like transposon that uses a protein we denoted TnsF as a target selector, and a previously unreported TE we name Tsy. Our results expand the understanding of target selection by Tn7-like transposons, reveal the structural features linked to RNA-guided or protein-guided modes of transposition in CAST systems, characterize the modular architecture of Tn7 target selectors, and discover a distinct target-selector partner co-opted from a previously undescribed non-Tn7 TE.



(legend on next page)

## RESULTS

**TnsC phylogeny reveals the diversity of target selectors in Tn7-like transposons**

Known target-selector proteins exhibit a wide range of diversity, but they all use TnsC to bridge target selection with transposition. We therefore used phylogenomic analysis of TnsC, which is the most prominently conserved protein among the Tn7-encoded proteins, as the framework to investigate the diversity of target selectors and search for new ones (Figures 1B and 1C). We selected 80,028 Tn7-like loci identified in publicly available prokaryotic datasets from NCBI (National Center for Biotechnology Information), JGI (Joint Genome Institute), and MG-RAST (Metagenomic Rapid Annotation using Subsystem Technology), that together included about  $1.6 \times 10^6$  bacterial and archaeal genomes (STAR Methods) and from these, extracted a representative set of TnsC for phylogenetic analysis. To build the phylogenetic tree, 6,384 TnsC homologs were selected. The tree included two main clades, one consisting of *MuB*—the TnsC homolog from the transposable phage Mu (MuTn)—and the other one of TnsC from Tn7-like transposons (Figure 2).

To explore the diversity of target selectors, we first mapped the known ones (*tniQ/tnsD*, *tnsE*, and *cas* genes from CAST systems) on the tree; *tniQ/tnsD* is ubiquitous in the Tn7 branch and is represented either by a single gene (2,905 loci) or as tandem genes (1,048 loci). These tandems consist of either two *tniQ* genes, or *tniQ* and *tnsD*, or two *tnsD* genes, which we collectively refer to as dual *tniQ-tnsD*. In contrast to the ubiquity of *tniQ/tnsD*, *tnsE* is more restricted in its spread and is present in transposons closely related to the canonical *E. coli* Tn7 and in the more distantly related group of Tn6022 transposons (379 loci total) (Figure 2). CAST systems are spread around the tree and generally grouped according to their subtypes. However, as noted previously, CAST I-B is represented in two distinct clades (1 and 2), suggesting independent capture by two distinct transposons.<sup>12</sup> We made similar observations for CAST I-F: Tn7017—a CAST I-F variant that harbors a dual *tniQ-tnsD* and uses TnsD for protein-mediated homing<sup>14</sup>—belongs to a branch distant from other CAST I-Fs, which use a dedicated spacer for RNA-guided transposition.<sup>10</sup> This branch consists of transposons encoding *tnsC* and dual *tniQ-tnsD*, but mostly lacking Cascade I-F, suggesting multiple gains or losses of Cascade I-F.

**Identification of CAST I-D**

To identify potential distinct CAST systems, we searched for *cas* genes encoding CRISPR-effector components (see STAR Methods) in the vicinity of *tnsC* and mapped the detected *cas*

genes onto the tree (Figures 1 and 2). We identified 234 groups of loci harboring at least one of these *cas* genes. Manual examination of *tnsC* tree branches bearing *cas* genes showed that several of these genes are part of the cargo and are unlikely to be involved in transposition<sup>15</sup> or belong to already reported CAST systems. However, we identified one group of loci in the Tn7 clade in branches closely related to CAST I-B2 that encodes transposase components closely similar to those of I-B2 PmcCAST, with ~50% sequence identity between TnsABs and TnsCs and ~30% sequence identity between the dual TniQ-TnsDs (Figure S1A). However, these loci encoded Cascade I-D, rather than Cascade I-B, and thus comprise a distinct CAST variety.

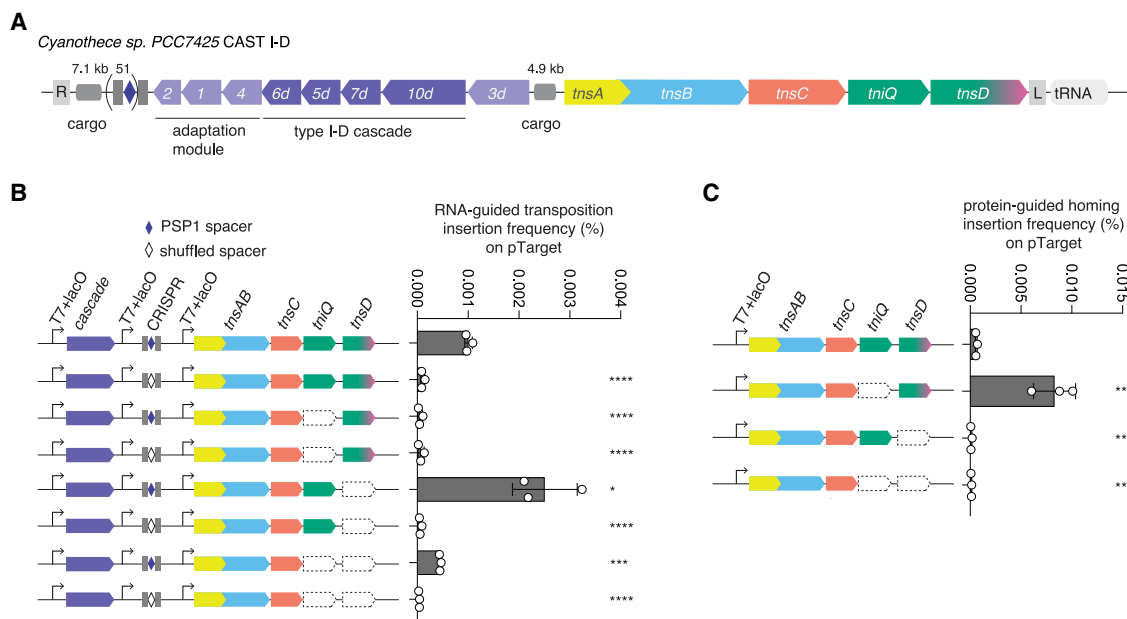
To experimentally characterize CAST I-D, we chose a locus from the cyanobacterium *Cyanothece* sp. PCC 7425, CyCAST, which encodes a complete subtype I-D CRISPR-Cas system encompassing both the adaptation module (Cas1, Cas2, and Cas4), the Cas6 processing nuclease, and the Cascade complex, along with TniQ and TnsD (Figure 3A). Unlike other known CASTs, the CRISPR-Cas system of CyCAST appears to be fully functional—that is, competent for both adaptation and interference—based on the conservation of catalytic residues in the HD-nuclease domain of Cas10d, suggesting that this is a recent acquisition of a CRISPR system not yet fully domesticated by Tn7 (Figure S1B). Similar Cas10d proteins are also found encoded in loci where Tn7 components are absent in the vicinity (Figure S1B). Manual identification of the CyCAST boundaries indicated potential attachment sites in a tRNA gene, similar to the attachment site of CAST I-B2 systems (Figure S1A).

We expressed CyCAST heterologously in *E. coli* and tested for activity. Using previously established assays,<sup>8,9,12</sup> we determined that CyCAST exhibits a GTT protospacer-adjacent motif (PAM) preference, as observed in subtype I-D CRISPR-Cas systems<sup>16</sup> (Figure S1C). We detected mainly unidirectional left-end (LE) and cargo right-end (RE) insertions within a 70 to 80-bp window downstream of the protospacer on a target plasmid (pTarget) (Figure S1D).

To clarify the roles of TniQ and TnsD in RNA-guided insertion, we deleted either TniQ, TnsD, or both and checked for activity. The full CyCAST system with both TniQ and TnsD showed RNA-guided transposition at an insertion frequency of 0.001% (Figure 3B). Deletion of TnsD boosted RNA-guided transposition about 2.5-fold, to 0.0025%, suggesting that TnsD partially inhibits this activity. Elimination of TniQ abolished RNA-guided insertion activity, but perhaps unexpectedly, in the absence of both TniQ and TnsD, a low level of RNA-guided transposition was detected (0.0004%). Thus, CAST I-D retains some basal RNA-guided transposition activity in the absence of a

**Figure 2. Phylogenetic tree of TnsC homologs**

Rings around the tree show the presence of a particular gene or a feature in the vicinity of *tnsC* within the genomic contig. From inner to outer ring: *tnsA* is shown in yellow, *tnsAB* fusion is shown in light blue, presence of *tnsB* operonized with the central *tnsC* (representative of the leaf) is shown in orange, presence of an additional distinct *tnsC* operonized with *tnsB* in the vicinity is shown in dark blue, *tniQ/tnsD* is shown in dark green and the presence of a second *tniQ/tnsD* (*tniQ2*) in pink where both their protein size are proportional to size of the ring bar, *tnsE* is shown in dark red, *cas* effectors and *cas6* genes are shown in purple, the presence of a gene operonized with the central *tnsC* is shown in green. Various known transposons are annotated around the tree including known CAST systems. Red boxes highlight areas of interest. The Mu clade corresponds to the left branch harboring a conserved gene operonized with MuB (homologous to *tnsC*). This gene is part of the Mu phage genome. By subtraction, the Tn7 clade corresponds to the remaining clade and is characterized by the presence of TniQ/TnsD.



**Figure 3. Characterization of CAST I-D**

(A) Schematic of *Cyanotherce* sp. PCC 7425 CAST I-D (CyCAST) locus architecture.

(B) RNA-guided insertion frequency of CyCAST into pTarget with PSP1, with or without TniQ and TnsD.

(C) Protein-mediated insertion frequency of CyCAST into pTarget with tRNA-leu, with or without TniQ and TnsD. ddPCR experiments were performed with three biological replicates.

All data points are shown with an error bar showing standard deviation, and statistical significance was assessed by t test. \*p < 0.05; \*\*p < 0.01; \*\*\*p < 0.001; \*\*\*\*p < 0.0001.

See also Figure S1.

protein-target selector, in contrast to the previously characterized CAST systems, which seems to suggest that TnsC can recognize Cascade at the insertion site (Figure S1E).

We also tested CRISPR-independent transposition (homing) of CyCAST. To this end, we cloned into the pTarget plasmid a leucine tRNA gene from *Cyanotherce* sp. PCC 7425 to serve as a homing site. Homing transposition was observed when TnsAB, TnsC, and TnsD were expressed in the absence of TniQ (0.008%); the presence of TniQ drastically diminished this activity, but transposition was still detectable (0.0006%) (Figure 3C). Homing transposition occurs around 30–33 nt downstream of the end of the tRNA homing site, resembling the insertion site of CyCAST in the genome of *Cyanotherce* sp. PCC 7425 (Figure S1F). Thus, CyCAST exhibits dual modes of transposition that rely on different target selectors, namely, the small TniQ protein for RNA-guided transposition and the larger TnsD for protein-mediated homing.<sup>12,14</sup>

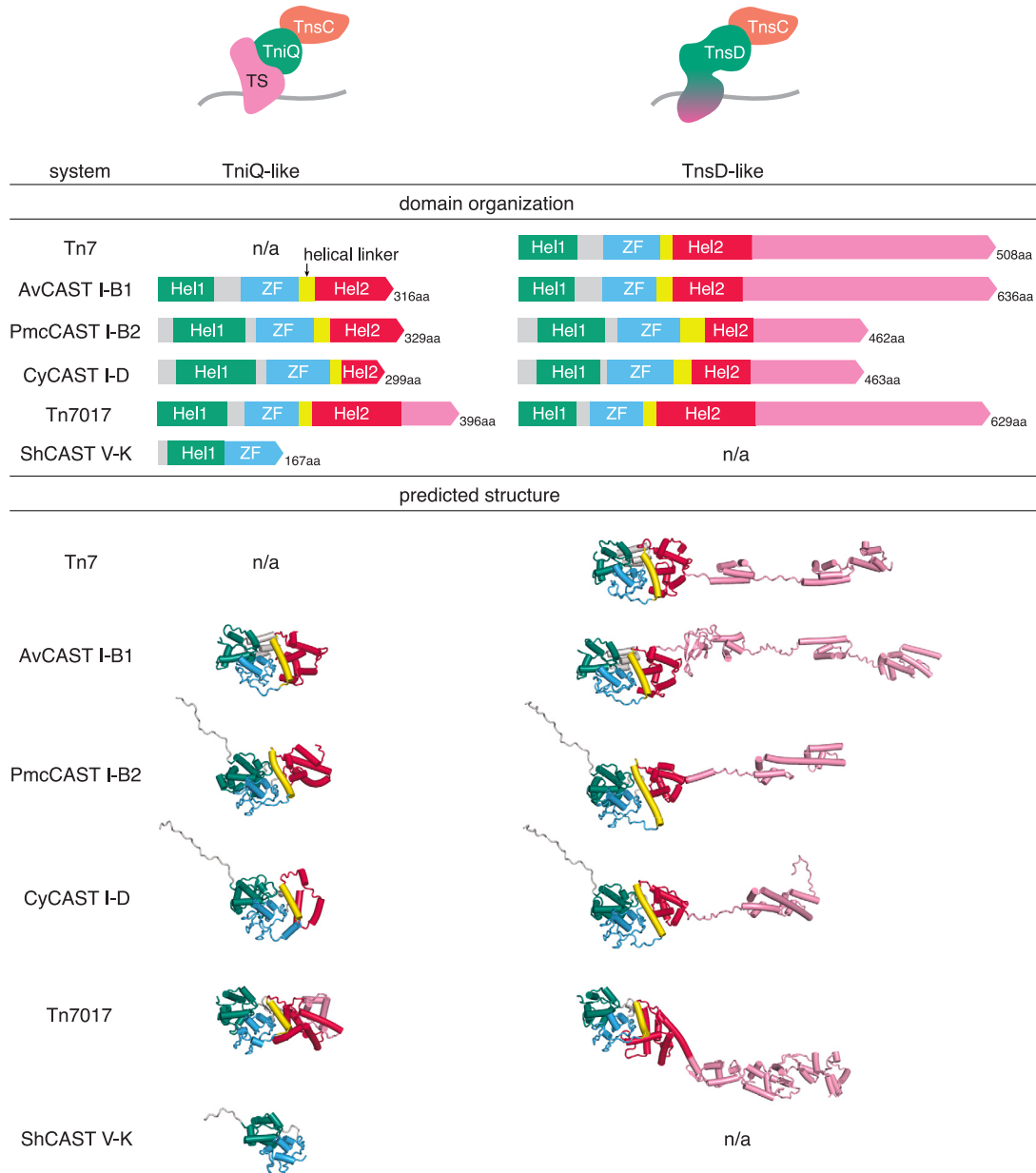
### Comparison of TniQ and TnsD reveals a modular architecture of target selectors

The presence of dual *tniQ-tnsD* genes in a variety of Tn7 loci motivated us to examine these proteins in greater detail to gain additional insight into their roles in target selection. Mechanistic studies of CAST systems have shown that TniQ, which is smaller than TnsD, partners with Cascade to mediate RNA-guided transposition, whereas TnsD mediates protein-guided transposition (similar to its role in *E. coli* Tn7).<sup>12</sup> We therefore first focused on

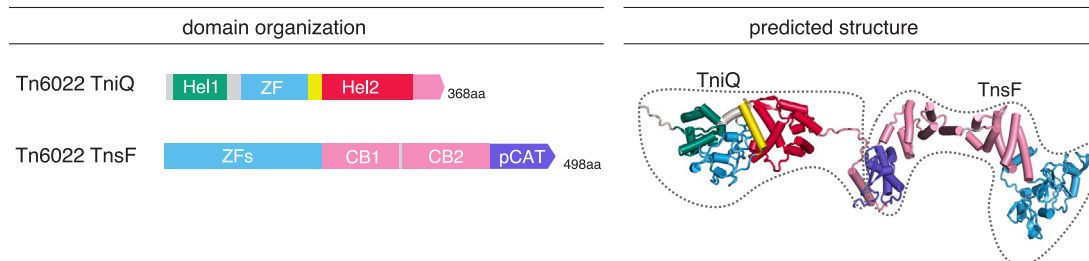
TniQ and TnsD from CASTs and non-CAST Tn7-like transposons to determine how they function in these two capacities. We employed structural prediction using AlphaFold2 (AF2)<sup>17–19</sup> to compare the domain organizations of TniQ and TnsD variants encoded by *E. coli* Tn7, CAST I-B, CAST I-D, and Tn7017 (the dual TniQ-TnsD CAST I-F) (Figures 4A and S2A). From the structural models, we identified a common core of about 300 amino acids (aa) that consists of a helical domain (Hel1), a zinc finger (ZF), a connector helix, and another helical domain (Hel2). Even within this core region, however, there are notable differences among the TniQ and TnsD proteins from different transposons (Figure S2A). TniQ from Tn6677 CAST I-F interacts with Cas6 as well as Cas7 and the guide RNA through a loop in the Hel2 domain,<sup>20</sup> suggesting that Hel2 provides a bridge to Cascade and that the structural diversity of Hel2 translates into compatibility with distinct Cascades. By contrast, TniQ from CAST V-K, which associates with the Cas12k effector rather than Cascade, contains only the Hel1 and ZF domains,<sup>21,22</sup> indicating a different mode of interaction between the target-selector components and highlighting the flexibility of TniQ as an adaptor between target selection and transposase machineries.

The CAST TniQ proteins consist largely of the core region, whereas the CAST and Tn7 TnsD proteins have diverse, long C-terminal extensions, which might confer target-DNA recognition to enable protein-guided transposition (Figure 4A). However, some of these regions share similarities that might reflect overlapping functions. For example, the TnsDs of Tn7 and

A



B



**Figure 4. Comparison of dual TniQ-TnsD in CAST systems**

(A) Domain architecture comparison of TniQ/TnsD. Left: CAST TniQ-like proteins involved in RNA-guided transposition. Right: TnsD-like proteins involved in protein-guided transposition compared with *E. coli* canonical Tn7 TnsD. Except for CAST V-K TniQ, all TniQ/TnsD share a common core region composed of a

(legend continued on next page)

CAST I-B, both of which home to the same site (the *glmS* gene),<sup>2,12,23</sup> share common C-terminal domain architectures, which might indicate that this C-terminal region is involved in the recognition of the attachment site. We also detected similar structures and domain architectures between the C-terminal regions of TnsDs from CAST I-B2 and CAST I-D, suggesting that they also recognize similar attachment sites.

### Identification of additional target selectors

The diversity of the interactions between TniQ and Cas effectors raises the possibility that the TniQ core acquired the ability to bind other target-selector proteins as well. Furthermore, these findings suggest that the presence of a small TniQ lacking a large C-terminal extension is a general hallmark of Tn7-like transposons that employ additional target selectors. We therefore expanded our analysis of TniQ and TnsD beyond the CAST systems to systematically analyze their diversity and identify potential partners of TniQ. We extracted the sequences of all TniQ and TnsD proteins encoded in the vicinity of the *tnsC* homologs in the Tn7 clade (5,072 altogether). We found 2,905 loci encoding a single TniQ or TnsD, 998 loci encoding dual TniQ-TnsD proteins, and 50 loci encoding more than 2 TniQ-TnsDs (33 loci with 3, 14 loci with 4, 2 loci with 5, and 1 locus with 6) (Figures 1B and S2B). Loci with more than a single TniQ or TnsD are from Tn7 co-occurring or TniQ-TnsD split into partial genes. TniQ and TnsD protein size distribution falls into four bins, suggesting selection driven by particular size restraints (Figure S2C), but TniQ and TnsD seem to represent two ends of a continuum, spanned by proteins containing extensions of variable length (Figure S2C). This variation makes it difficult to predict, for some of these proteins, whether they bind directly to attachment sites or require a target selector partner. We refer to such proteins as TniQ/TnsD, reflecting this uncertainty.

In dual TniQ-TnsD loci, one of these proteins is usually larger than 400 aa, suggesting direct target selection, whereas the second one is substantially smaller (Figure S2C). Such an architecture could provide transposons with two distinct options for target selection: direct, via TnsD, and indirect, via TniQ interacting with an additional target selector. Analysis of the phylogenetic tree of TniQ/TnsD built using the multiple alignment of the core regions (made from 4,916 proteins passing alignment filters; see STAR Methods) indicates that the dual TniQ-TnsD arrangement is polyphyletic (i.e., it emerged independently on multiple occasions via duplication of a single *tniQ/tnsD* [Figure S2B]), with the monophyly of the dual TniQ-TnsD (i.e., evolution via a single duplication of an ancestral TniQ/TnsD) compellingly ruled out ( $p$  value =  $2.4e-236$ ). The ancestral Tn7-like transposons likely encoded a single TniQ/TnsD protein, and the dual TniQ-TnsD configurations apparently evolved by *in situ* duplication of a single *tniQ/tnsD* (Figure S2B), followed

by neofunctionalization. Such duplication might maintain the compatibility with the other transposase components, while opening the possibility of evolving different modes of target selection. Consistent with the independent duplication scenario, we identified only a few examples of dual TniQ-TnsD loci containing distantly related proteins, as would be expected under an alternative scenario including the exchange of *tniQ/tnsD* genes between different transposons (Figures S2B and S2D).

Apart from the dual TniQ-TnsD loci, numerous Tn7-like transposons encompass TniQ together with another, unrelated target selector, such as TnsE, the plasmid-target selector (Figure S2B). We identified two distinct branches encoding TniQ and TnsE (Figures S3A and S3B). Although these TnsEs are highly divergent in sequence (less than 10% of sequence identity), they are predicted to form closely similar structures (Figure S3C). The N-terminal domain of TnsE binds dsRNA via a unique fold,<sup>24</sup> but the function of this domain has not been explored. Using structural prediction and structural mining<sup>25</sup> of the N-terminal region, we found that it folds into two single-strand binding (SSB) domains related to PriB (Figure S3D), a component of the bacterial primosome, which can bind both ssDNA and ssRNA and is involved in restarting replication at the fork.<sup>26-28</sup> Such structural similarity with PriB suggests that TnsE might have been co-opted from a system that functions at the replication fork. The domain architecture of TnsE has features including a dsDNA-binding domain and a domain predicted to bind ssRNA or DNA, suggesting that it targets the lagging strand of replication. Thus, TnsE is likely to specifically target replication forks of conjugative plasmids, further highlighting the remarkable diversity of target selectors co-opted by Tn7-like transposons.

To search for additional target selectors, we focused on genes operonized with TniQ/TnsD and conserved in multiple nodes of the tree (see STAR Methods and Figure S2D). The most common candidate gene encodes an uncharacterized protein of 498 aa and forms a putative operon with a gene encoding a short TniQ (369 aa) in the Tn6022 family of Tn7-like transposons (121 groups of loci). This gene was previously annotated as *orf3* or *tniE*,<sup>29,30,31</sup> but we denote it TnsF (Figure S4A). AF2 prediction of the TnsF structure revealed a distinct domain architecture including an N-terminal region containing multiple ZFs in the first 199 aa, whereas the remaining ~300 aa exhibit significant structural similarity to the N- and C-terminal domains of the tyrosine recombinase superfamily member XerH (PDB: 5jk0<sup>32</sup>) (Dali score 3.9) (Figures 4B and S4B). However, the chamber holding the tyrosine catalytic site is missing in TnsF (Figure S4B). Tyrosine recombinases typically contain an N-terminal DNA-binding domain (CB domain<sup>32</sup>) and a C-terminal catalytic domain (CAT) and dimerize or tetramerize on DNA during site-specific recombination. Both the N- and C-terminal domains of XerH interact with DNA in the crystal structure (Figure S4B). TnsF is predicted

N-terminal helical domain (Hel1), a zinc finger (ZF), a helical linker, and a C-terminal helical domain (Hel2) (CAST I-D TniQ contains only a partial Hel2 domain). TnsD-like proteins performing protein-guided insertion in CASTs harbor long and diverse C-terminal regions folding into multiple HTH domains similar to Tn7 TnsD.

(B) Domain architecture of TniQ and TnsF in Tn6022 (left). Docking prediction of TniQ and TnsF (right). Pink in TniQ indicates C-terminal extension predicted to interact with TnsF. Pink and purple in TnsF indicate the predicted tandem-core binding domain (CB1 and CB2) and the partial catalytic domain (pCAT), respectively.

See also Figures S2 and S3.



to contain tandem domains (designated CB1 and CB2), which are structurally similar and, by inference, homologous to the CB of XerH (Figure S4B), suggesting that these domains may impart the ability to interact with DNA. A  $\beta$  sheet at the C-terminal region of XerH maps to the last 70 aa of TnsF (designated partial CAT or pCAT) (Figure 4B) and corresponds to the DNA-binding region of the recombinase domain of XerH (Figure S4B) but lacks the helix K.<sup>33</sup> In XerH, the helix K contributes substantially to its interaction with DNA, suggesting that the pCAT domain of TnsF lost the DNA-binding capacity and might instead interact with TniQ. Indeed, based on an AF2 multimer model, TnsF is predicted to interact with TniQ through its C-terminal region (Figure 4B). Additionally, we found one case where the N-terminal region of TnsF is fused to the C-terminal region of TniQ (GenBank: SCZ64694). This fusion protein lacks the entire CAT but contains an additional CB-like domain within the linker between TniQ and TnsF.

The Tn6022 transposons also encode *tnsE* on the opposite strand, suggesting they can jump to conjugative plasmids, whereas TniQ and TnsF are likely involved in target selection within the bacterial chromosome. Together, these data suggest that TnsF is a distinct target selector and that TniQ of Tn6022 serves as a hub that bridges TnsF, which binds directly to the attachment site, with the transposition machinery.

### TnsF is essential for Tn6022 transposition and interacts with TniQ

To experimentally test the predicted target selector function of TnsF, we focused on the Tn6022 transposon. Tn6022 encodes TnsA, TnsB, TnsC, TniQ, TnsF, and TnsE and is inserted in the *comM* gene, which encodes a protein containing a AAA+ ATPase domain and a Mg chelatase domain<sup>34,35</sup> (Figure 5A; Data S1). We reconstituted *Acinetobacter johnsonii* Tn6022 (hereafter, AjTn6022) in *E. coli*. We determined the ends of the transposon (see STAR Methods) and cloned the left and right ends into a pDonor plasmid with a kanamycin-resistance gene as a cargo. We also cloned a 100-bp fragment (50 bp upstream and 50 bp downstream of the insertion site) of the *AjcomM* gene into a pTarget plasmid. These plasmids were co-electroporated into *E. coli* with a pHelper plasmid (bearing *tnsA*, *tnsB*, *tnsC*, *tniQ*, and *tnsF*). To determine the structure of the insertion, we performed long-read, amplification-free nanopore sequencing. We found simple insertions (60.9% of insertions) and co-integrate insertions (39.1% of insertions) (Figure S4C), and we confirmed the presence of target-site duplications (TSDs), a signature of Tn7-like transposition (Figure S4D). To determine whether all Tns proteins including TnsF are essential for transposition into the *comM* gene, we generated pHelper variants lacking each of the Tns proteins and repeated the transposition assay. AjTn6022 achieves transposition at 2.4% efficiency, and removal of any Tns protein (A, B, C, Q, or F) impaired transposition, as quantified by droplet digital PCR (ddPCR) (Figure 5B).

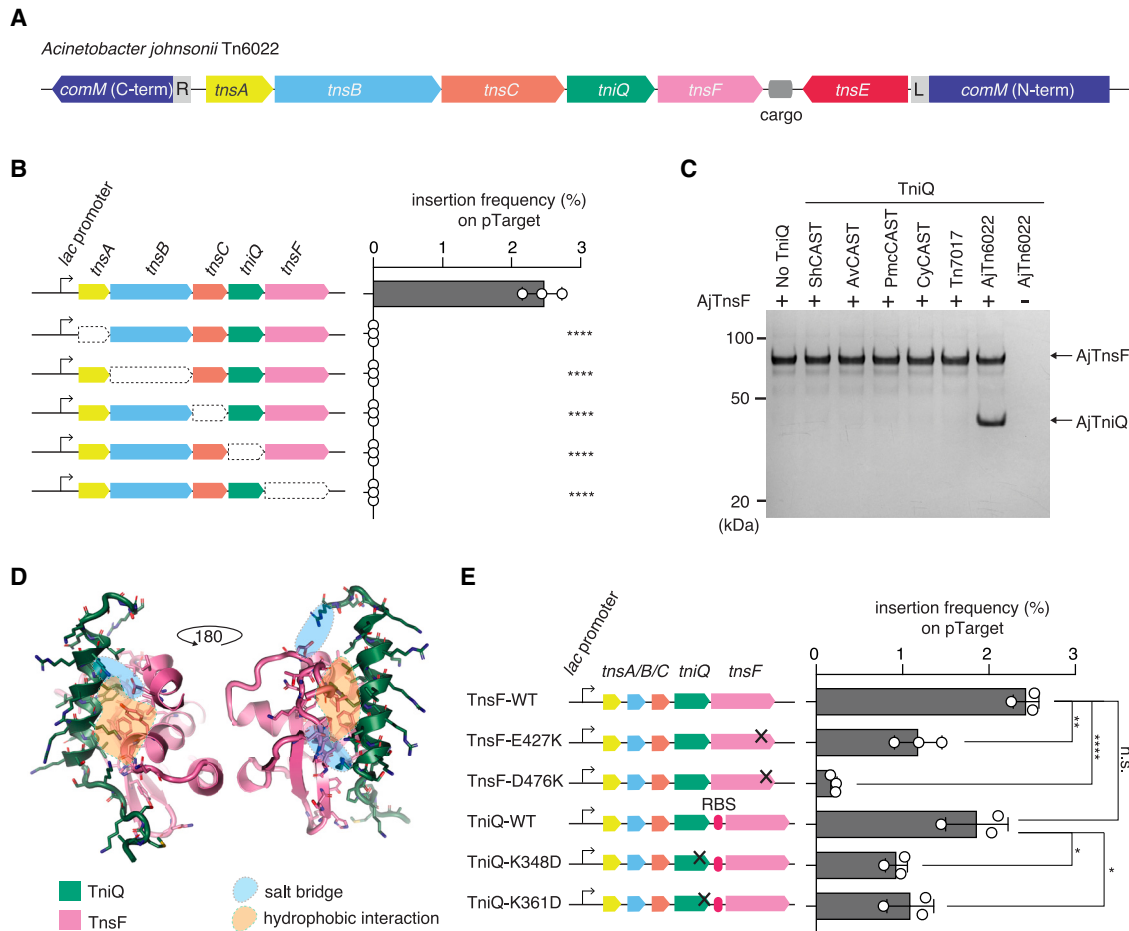
To test our prediction that TnsF and TniQ interact, we purified TnsF and TniQ proteins from AjTn6022, and performed pull-down assays. We showed that *in vitro* TnsF interacts specifically with AjTniQ, but not with TniQ proteins from ShCAST, AvCAST, PmcCAST, CyCAST, or Tn7017 (Figure 5C; Data S2). The docking model of AjTnsF:AjTniQ predicts an interaction between the

C-terminal helix of TniQ and the pCAT region of TnsF via hydrophobic contacts and salt bridges (Figure 5D). TniQ from other systems lack this helical region, which may explain why they do not interact with AjTnsF. Mutants designed to disrupt the predicted salt bridges (TnsF E427K/D476K and TniQ K348D/K361D) abrogated or substantially reduced transposition (Figure 5D), supporting the hypothesis that this Aj-specific region of TniQ is important for the interaction with TnsF.

### Identification of Tsy, a non-Tn7 transposon that uses TnsF for target selection

We next searched for TnsF homologs in genomic databases and identified 1,099 nonredundant TnsF homologs (STAR Methods; Data S1), including Tn6022 TnsF and a homolog of TnsF (referred to as TnsF-like protein) containing a predicted active CAT (Figures S4A and S4B). Although we also detected more distant structural homologs of TnsF containing adjacent CB and CAT domains, these proteins lacked the ZF-containing N-terminal regions, so we did not include them in the further analysis (Figure S4E). We built a phylogenetic tree of the TnsF and TnsF-like proteins (Figure S4F) and mapped on it the conserved genes located in the vicinity of *tnsF* and *tnsF*-like genes. Tn7 TnsF forms a distinct clade and apparently evolved from TnsF-like proteins with an active tyrosine recombinase CAT. The genomic neighborhood of these TnsF-like proteins lacks Tn7 components but instead includes upstream genes encoding a tyrosine recombinase (YRec) and a small helix turn helix (HTH)-domain protein as well as a downstream gene encoding a GIY-YIG nuclease (present only in the branch close to Tn7) (Figures 6A and S4F). Although we could not detect inverted repeats or any canonical ends in these loci, we noticed the presence of *comM* fragments, namely, the 5' terminal portion located upstream of YRec and the 3' terminal portion located downstream of the putative transposon (Figure 6A). The downstream portion of the *comM* gene is in some cases located after several additional genes, which likely represent transposon cargo. These features suggest that this locus is a distinct transposon and that the TnsF-like protein recognizes *comM*, similarly to TnsF of Tn6022.

To determine the function of these potentially enzymatically active TnsF-like proteins, we experimentally characterized this mobile element, which we designate Tsy (target selector based on tyrosine recombinase). We reconstituted the system from *Zoogloea* sp. LCSB751 (hereafter, ZooTsy) in *E. coli*. To assess ZooTsy transposition, we cloned 135 and 39 bp of each transposon end1 (*comM* 5' terminal portion) and end2 (*comM* 3' terminal portion) with 12 bp of homology arm extensions into an R6K origin pDonor plasmid with a kanamycin-resistance gene as cargo (Figures S5A and S5B). We also cloned a Tsy attachment site (a 100-bp fragment with 50 bp upstream and 50 bp downstream of the *Zoogloea* sp. *comM* insertion site) into a pTarget plasmid (Figure S5A and S5B). These plasmids were co-electroporated into *E. coli* with a pHelper plasmid (bearing YRec, HTH, *tnsF*, and the GIY-YIG nuclease, *nuc*). We detected transposition into pTarget and observed circular intermediates (CIs) derived from pDonor by PCR in a YRec-dependent manner (Figure S5A), as previously demonstrated for various transposons encoding tyrosine recombinases.<sup>34</sup> To confirm these findings, we established an assay to isolate and confirm the structure of the CI



**Figure 5. Characterization of TnsF-containing Tn6022**

(A) Schematic of *Acinetobacter johnsonii* Tn6022 (AjTn6022) locus architecture.

(B) ddPCR experiments showing the insertion frequency of AjTn6022 into pTarget with a 200 bp-fragment of *comM* in the absence of indicated AjTn6022 component.

See also Figures S4C and S4D.

(C) Protein gel showing TnsF-TniQ interaction. TwinStrep-bdSUMO-TnsF was mixed with the indicated purified TniQ protein, bound to beads, and then eluted complexes were analyzed by gel. Raw gel image and gel showing purified proteins used in the pull-down assay are in SD 2.

(D) Molecular details of the interaction region from the docking prediction between Tn6022 TnsF (pink) and TniQ (green). TniQ is predicted to interact via its C-terminal helix with the pCAT domain of TnsF. The interaction involves 2 salt bridges (pale blue areas) and multiple hydrophobic interactions (pale orange areas) (left).

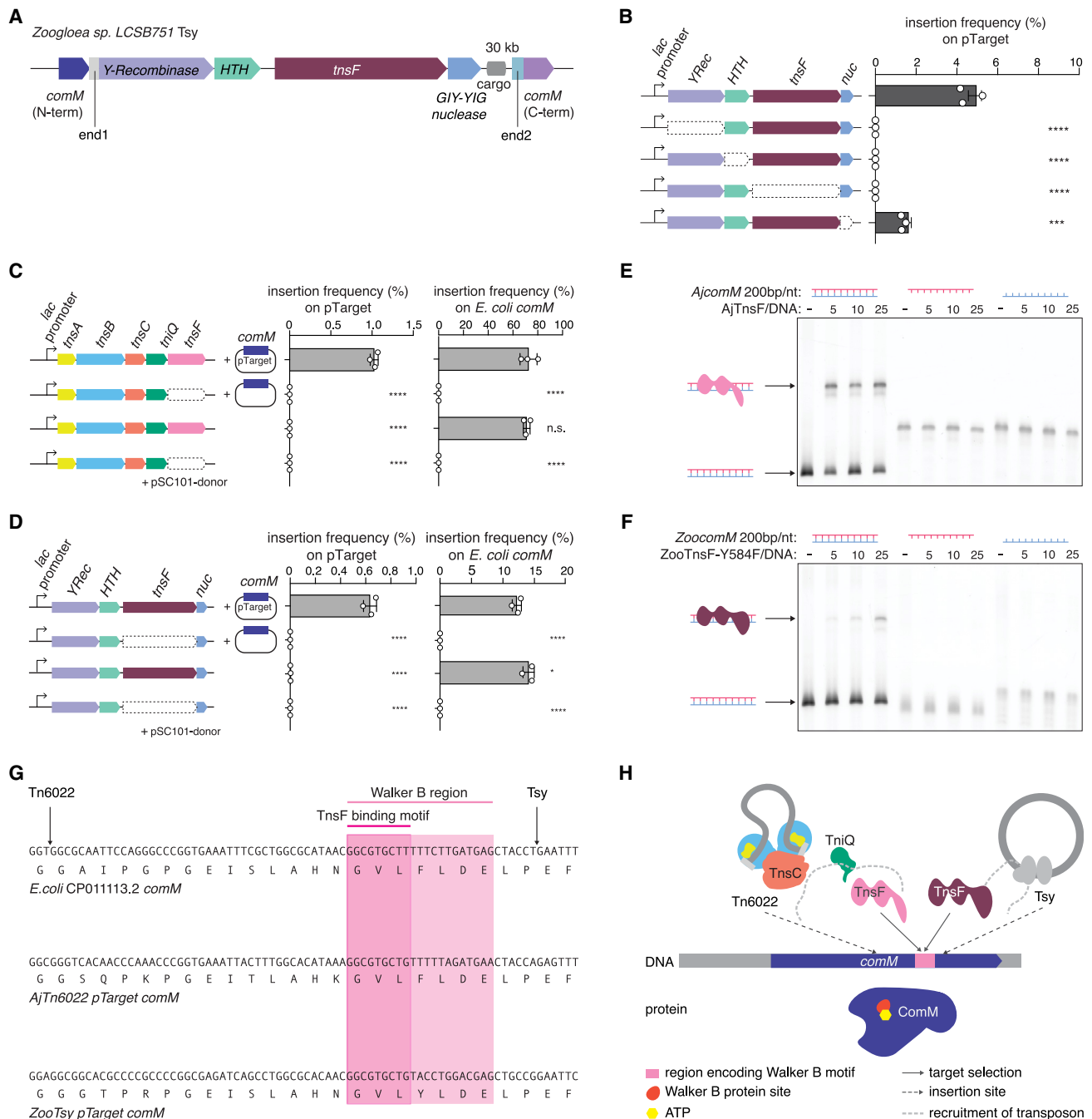
(E) ddPCR experiments showing the insertion frequency of AjTn6022 into pTarget with a 200 bp-fragment of *comM* with indicated TnsF and TniQ mutants. For TniQ mutants, overlapped TniQ and TnsF sequences were separated with a ribosome-binding sequence (RBS). ddPCR experiments were performed with three biological replicates.

All data points are shown with an error bar showing standard deviation, and statistical significance was assessed by t test; \* $p < 0.05$ ; \*\* $p < 0.01$ ; \*\*\* $p < 0.0001$ ; n.s., not significant.

derived from the pDonor. We constructed a derivative pDonor that contained the ColE1 origin of replication and kanamycin resistance gene as cargo as well as *lacZ $\alpha$*  but no other origin of replication. Upon circularization, this pDonor will lose *lacZ $\alpha$*  and can be isolated from white colonies by traditional blue-white screening (Figure S5A). Using this assay, we obtained white colonies (98% of the total) in a pHelper-dependent manner after retransformation with the extracted plasmids and successfully isolated a smaller plasmid which had lost the 0.6-kb backbone region of the pDonor (Figure S5A; Data S3). We confirmed the smaller plasmid as a CI by nanopore long-read sequencing

and observed the connected end2 (...AATCCCAGTC) and end1 (AAGTTCTGAT...) junction by Sanger sequencing (Figure S5A; Data S3). To determine the structure of the ZooTsy insertions, we performed nanopore long-read sequencing and found simple insertions (62.3% of total insertions) (Figures S5C and S5D).

To narrow down the requirements for transposition, we generated variants of the cargo with serial deletions of end1 from 135 to 0 bp, finding that truncation of this end gradually decreased the rate of simple insertions to zero (Figure S5E). By contrast, only 20 bp at end2 are required for transposition (Figure S5E). By systematically combining these optimized parameters, we found



**Figure 6. TnsF targets a conserved Walker B motif in *comM***

(A) Schematic of the locus architecture of *Zoogloea* sp. LCSB751 target selector based on tyrosine (Y) recombinase transposon (ZooTsy).  
 (B) Genetic requirement of YRec, HTH, and TnsF on ZooTsy transposition activity, as assayed by quantification of upstream-end1 junction formation by ddPCR. Deleted genes are indicated by a dashed outline.  
 See also Figure S5H.

(C) ddPCR experiments showing the insertion frequency of AjTn6022 into pTarget with a 200 bp-fragment of *comM* (left) and *E. coli* endogenous *comM* (right) in the absence of TnsF and/or presence of the pTarget(AjTn6022). pSC101 donor was used.  
 (D) ddPCR experiments showing the insertion frequency of ZooTsy into pTarget with a 200 bp-fragment of *comM* (left) and *E. coli* endogenous *comM* (right) in the absence of TnsF and/or presence of the pTarget(ZooTsy). pSC101 donor was used.

(E) Electrophoretic mobility shift assay (EMSA) to assess the interaction between a 200-bp or 200-nt fragment of AjcomM and purified AjTn6022-TnsF.  
 (F) EMSA to assess the interaction between a 200-bp or 200-nt fragment of ZoocomM and purified ZooTsy-TnsF\_Y584F.

(legend continued on next page)

that 12 bp of homology arm 1 (hom1), 135-bp end1, and 20-bp end2 are sufficient for the transposition of ZooTsy (Figure S5F).

To determine the genetic requirements for ZooTsy transpositions in *E. coli*, we constructed a series of pHelper plasmids with deletions of each gene. ZooTsy achieves transposition at 5.0% efficiency, and removal of any component (YRec, HTH, TnsF, or GIY-YIG nuclease) from the system impaired transposition, as quantified by ddPCR for the upstream-end1 junction formation (Figures 6B and S5G). The GIY-YIG nuclease, however, is not essential for transposition, which is supported by the identification of Tsy relatives lacking this component (Figure S4E). By contrast, we found that the tyrosine recombinase catalytic activities of both YRec and TnsF are essential for transposition (Figure S5H).

### TnsF targets a conserved region of *comM*

Both AjTn6022 and ZooTsy use TnsF to target *comM* of their respective hosts, although the directionality of insertion is different, suggesting different modes of insertion. ComM has been reported to facilitate recombination of sequences acquired by transformation,<sup>35</sup> and disruption of the *comM* gene by transposon insertion could inactivate this functionality, limiting transformation of other MGEs. To explore *comM* targeting in greater depth, we sought to compare TnsF-mediated targeting of this gene by AjTnsF and ZooTnsF. Using our heterologous *E. coli* expression systems, we looked for AjTn6022 and ZooTsy targeting of genomic *comM*, finding that they can both target *E. coli* endogenous *comM* in addition to their respective *comM* sites on pTarget, highlighting the broad recognition of *comM* by both TnsFs (Figures 6C and 6D). To determine the specificity of *comM* gene targeting by TnsF, we performed tagmentation-based tag integration site sequencing (TTISS).<sup>36</sup> For AjTn6022, in the absence of pTarget, we observed that 96.7% of insertions were at the *E. coli* endogenous *comM* locus; in the presence of pTarget, 56.6% of insertions were on pTarget and 40.8% were on the genomic *comM* (97.4% on target in total) (Figure S6A). For ZooTsy, we observed similarly high levels of specificity, indicating that TnsF is highly selective for *comM* (Figure S6B). We confirmed that both TnsF proteins bind 200-bp dsDNA fragments corresponding to their respective *comM* target sequences (Figures 6E and 6F).

To further characterize the attachment sites of these proteins, we constructed additional pTarget variants with different lengths of *comM* gene fragments to map the target-site specificity of both TnsFs at a greater resolution. For AjTn6022, deletion of the 40–50 bp of either upstream or downstream sequences of the insertion site substantially reduced transposition, indicating that the insertion site was located within the TnsF attachment site (Figure S6C). For Tsy, we found the attachment site is within a 40-bp upstream region from the insertion site (Figure S6D). Mapping these refined target sites on the respective *comM* genes showed that they overlap with a conserved

10-bp region within the Walker B motif (a highly conserved ATPase motif<sup>37</sup>) of the ComM protein (Figure 6G). We confirmed that this 10-bp region is necessary for both AjTnsF and ZooTnsF binding (Figures S7A and S7B). Furthermore, we found that mutating this region abolished the transposition activity of both AjTn6022 and ZooTsy (Figures S7C and S7D). Targeting the Walker B motif of *comM* might provide a natural conserved anchor for TEs to spread across species, paralleling Tn7 targeting of the catalytic site in *glmS*. Together, these results demonstrate that TnsF is a target selector related to the tyrosine recombinase family and involved in the transposition of at least two distinct groups of transposons.

### DISCUSSION

Target-site selection is a crucial step in the life cycle of TEs because insertion of the element can dramatically affect the fitness of both the host and the TE depending on the insertion site.<sup>38,39</sup> Moreover, the choice of target site is critical for the TE to spread horizontally. Here, we expand the current understanding of target-site selection mechanisms, identifying previously uncharacterized target-site selectors and a distinct family of TE. Together, our results reveal the modular architecture used by Tn7-like transposons to bridge target-site selection with transposase activity, providing TEs with maximal targeting flexibility (Figure 7).

This flexibility is manifested in the various proteins that Tn7-like transposons have co-opted and adopted for target-site selection. Although the CRISPR systems in most CASTs have lost their ancestral interference and nuclease functions, the CyCAST system we describe here has an interference component and a Cas10d protein with an active HD-nuclease domain. Very recently, other CAST I-D systems have been reported—but these lack Cas3d, and Cas10d is inactivated.<sup>40</sup> Thus, CyCAST seems to represent an evolutionary intermediate where the CRISPR system has been co-opted by the TE but has not fully lost its native function. Along similar lines, we found that the Tn7-like transposon Tn6022 has co-opted TnsF, a catalytically inactive derivative of a tyrosine-recombinase-containing protein, whereas the Tsy transposon we describe here encodes the apparent ancestor of TnsF, a catalytically active recombinase.

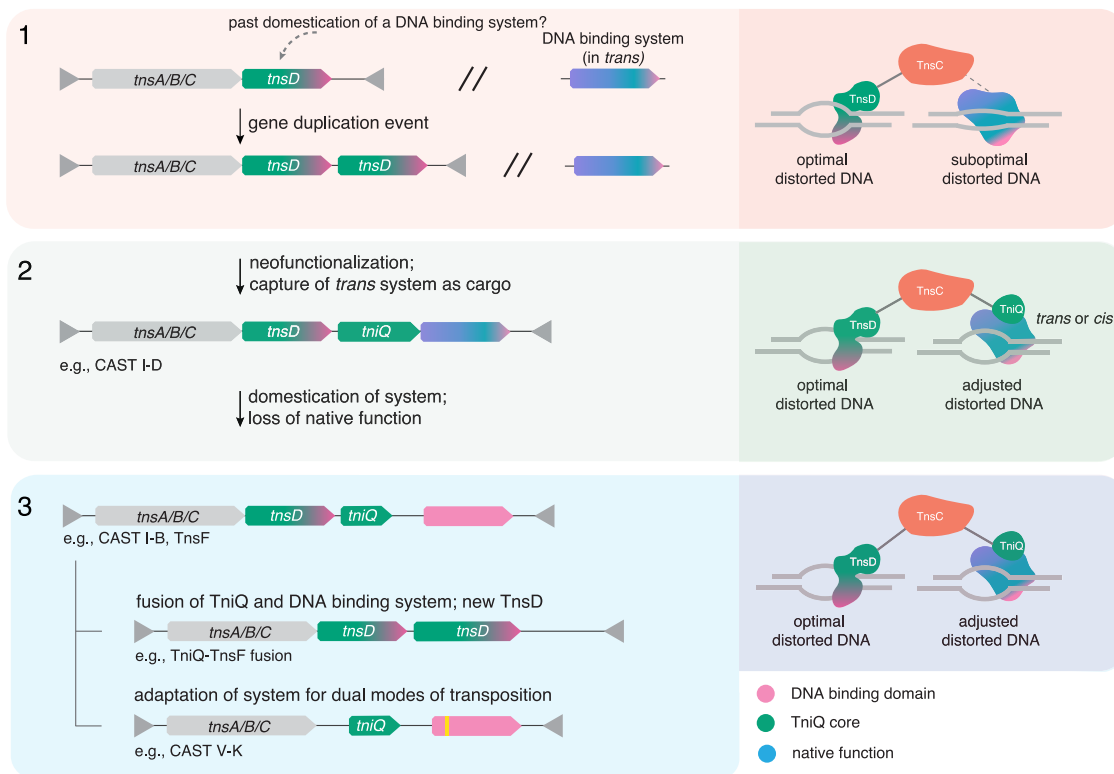
Our finding that additional enzymes have been co-opted as target selectors by Tn7-like transposons raises the question of how these recruited proteins evolve target-site selection capacity. In Tn7, to recruit the transposase machinery to the target site, TnsD binds to and induces a local distortion in the target DNA, which then attracts TnsC. Artificial induction of such DNA distortion has been shown to attract TnsC independently of TnsD.<sup>23,41</sup> Native Cas effectors<sup>42</sup> and TnsF likely induce similar distortions in the target DNA, through R-loop formation by Cas effectors and DNA bending by tyrosine-recombinase-containing proteins such

(G) Insertion sites of Tn6022 and ZooTsy on *E. coli* endogenous *comM* and *comM* of their respective hosts used in the pTarget. ComM protein sequences (translated *comM*) are shown below the nucleotide sequence. The pink rectangle indicates the genomic location of the Walker B of the AAA ATPase encoded by *comM*; the red rectangle shows the probable hot spot binding region of both TnsFs.

(H) Model of TnsF target selection and insertions for Tn6022 and Tsy.

ddPCR experiments were performed with three biological replicates. All data points are shown with an error bar showing standard deviation, and statistical significance was assessed by t test. \*p < 0.05; \*\*\*p < 0.001; \*\*\*\*p < 0.0001; n.s., not significant.

See also Figures S5–S7.



**Figure 7. Evolution of the functional versatility of TniQ-interacting target selectors**

Evolutionary scenarios for various Tn7-like transposons with distinct modes of target selection. Locus architecture is shown on the left and mechanics of target-site selection on the right. (1) An ancestral Tn7-like transposon might have used TnsD for site-specific target selection and a DNA-bending protein or complex (e.g., Cas effector, transcription factor, or tyrosine recombinase) in *trans* as a second mode of target-site selection. These DNA-bending proteins would create a distortion in the DNA that TnsC would recognize, albeit with a low efficiency. Gene duplication produced a second copy of TnsD. (2) Neofunctionalization of the second copy of TnsD yielded TniQ, which evolved to optimize the interaction between a *trans* DNA-bending target selector and the target site. The *trans* system could also be captured by the transposon as cargo, as was the case with CRISPR-Cas systems. (3) Further domestication of the target selectors would then occur, eventually leading to the loss of the native function of the system (e.g., CASTs I-B and I-F), fusion to TniQ generating a distinct TnsD (e.g., TniQ-TnsF fusion), or adaptation of the system for dual modes of transposition as in CAST V-K, which relies entirely on the CRISPR system for both homing and jumping. Pink indicates DNA-binding function; green indicates TniQ core; blue indicates native function of DNA-binding system.

as TnsF. Although a canonical tyrosine recombinase would entirely cover the distorted DNA, precluding TnsC access, the distinct architecture of TnsF, with two different DNA-binding domains, might bend the target DNA while still allowing partial or full access for TnsC. Thus, proteins that provide more efficient target selection could supersede the target-selecting role of the C-terminal region of TnsD, ultimately leading to their domestication and loss of their native enzymatic activity (Figure 7). CAST-I-D and the TnsF homolog in the Tsy system are examples of apparent intermediate stages on the evolutionary path to domestication. We also detected numerous loci encoding a TniQ that is too short to enable target selection and no other identifiable target-selector partner. Such elements might recruit target selectors in *trans*, perhaps representing the initial step in the evolution of new target selectors or an even greater flexibility in target-selector recruitment.

#### Limitations of the study

Given the apparent fast evolution of target-selector proteins, our sequence-based mining might have limited the scope of our ana-

lyses. The recent advances in protein-structure prediction now enable structure-based mining, which could yield candidates beyond those reported here. Indeed, although a sequence-based search did not identify TnsF homologs in other systems, a structural-mining approach might shed light on the potential origins of target selectors, as exemplified by the relationship between TnsE and PriB (Figure S3). These types of evolutionary and structural analyses, combined with further study of the mechanisms of TnsF and Tsy, will shed light on this distinct mode of target-site selection. The identification of distinct target selectors described here highlights the remarkable plasticity of the insertion machinery of Tn7-like transposons, but further research will likely reveal additional mechanisms of transposon targeting.

#### STAR★METHODS

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

- KEY RESOURCES TABLE

● RESOURCE AVAILABILITY

- Lead contact
- Materials availability
- Data and code availability

● METHOD DETAILS

- Identification of Tn7-like transposons
- Construction of the phylogenetic tree of TnsC homologs
- Annotation of the genomic neighborhoods of TnsC homologs
- Protein structure prediction and analysis
- Phylogenetic analysis of TniQ
- Analysis of candidate partners of TniQ
- TnsF analysis
- Mining of TnsF and phylogenetic analysis of TnsF homologs
- Determination of transposon ends for Tn6022 and Tsy
- Computational analysis of CAST I-D
- Plasmid construction
- *E. coli* transposition assays
- Nanopore long-read sequencing to determine plasmid structure
- Tsy circular intermediate isolation assays
- TTISS for insertion specificity analysis
- Droplet digital PCR reactions
- Purification of TniQ and TnsF proteins
- Pull-down experiments for detecting AjTn6022 TniQ-TnsF interactions
- Electrophoretic mobility shift assay for detecting TnsF-DNA interactions

● QUANTIFICATION AND STATISTICAL ANALYSIS

- Quantification of ddPCR data

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.molcel.2023.05.013>.

ACKNOWLEDGMENTS

We thank S. Kannan and C. Frangieh for assistance with TTISS analysis, S. Maguire for assistance with cloning, Zhang lab members and J. Schmid-Burgk for useful discussions, and A. Farina for preparation of reagents. M.S. is supported by a Long-Term Fellowship from the Human Frontier Science Program. E.V.K. is supported by the Intramural Research Program of the National Institutes of Health (National Library of Medicine). F.Z. is supported by an NIH grant (2R01HG009761-05), the Howard Hughes Medical Institute, the Bill and Melinda Gates Foundation, the Poitras Center for Psychiatric Disorders Research at MIT, the Hock E. Tan and K. Lisa Yang Center for Autism Research at MIT, the K. Lisa Yang and Hock E. Tan Center for Molecular Therapeutics at MIT, the K. Lisa Yang Brain-Body Center at MIT, Broad Institute Programmable Therapeutics gift donors, the Pershing Square foundation, W. Ackman and N. Oxman, J. Poitras and P. Poitras, K.C. Griffin, the BT Charitable Foundation, the Asness Family Foundation, the Phillips family, D. Cheng, and R. Metcalfe.

AUTHOR CONTRIBUTIONS

G.F., S.B., Y.I.W., H.A.-T., and K.S.M. performed phylogenomic analysis. G.F. and S.B. performed structural analyses. G.F., S.B., E.N., J.S., and K.M. performed computational curation of systems. M.S. and I.P. performed and analyzed experiments characterizing CAST I-D, Tn6022, and Tsy. D.L. provided code for analyzing NGS data. F.Z. supervised the research and experi-

mental design with support from R.K.M. G.F., M.S., R.K.M., E.V.K., and F.Z. wrote the manuscript with input from all authors.

DECLARATION OF INTERESTS

F.Z. is a scientific advisor and cofounder of Editas Medicine, Beam Therapeutics, Pairwise Plants, Arbor Biotechnologies, Proof Diagnostics, and Aera Therapeutics. F.Z. is a scientific advisor for Octant. G.F., M.K., and F.Z. are co-inventors on a provisional patent application filed by Broad relating to this work.

INCLUSION AND DIVERSITY

One or more of the authors of this paper self-identifies as an underrepresented ethnic minority in their field of research or within their geographical location. One or more of the authors of this paper self-identifies as a gender minority in their field of research.

Received: September 23, 2022

Revised: January 17, 2023

Accepted: May 9, 2023

Published: June 1, 2023

REFERENCES

1. Dyda, F., and Hickman, A.B. (2015). Mechanism of spacer integration links the CRISPR/Cas system to transposition as a form of mobile DNA. *Mobile DNA* 6, 9.
2. Waddell, C.S., and Craig, N.L. (1988). Tn7 transposition: two transposition pathways directed by five Tn7-encoded genes. *Genes Dev.* 2, 137–149.
3. Parks, A.R., and Peters, J.E. (2009). Tn7 elements: engendering diversity from chromosomes to episomes. *Plasmid* 61, 1–14.
4. Peters, J.E. (2014). Tn7. *Microbiol. Spectr.* 2, 1–20. <https://doi.org/10.1128/microbiolspec.MDNA3-0010-2014>.
5. Shen, Y., Gomez-Blanco, J., Petassi, M.T., Peters, J.E., Ortega, J., and Guarné, A. (2022). Structural basis for DNA targeting by the Tn7 transposon. *Nat. Struct. Mol. Biol.* 29, 143–151.
6. Choi, K.Y., Spencer, J.M., and Craig, N.L. (2014). The Tn7 transposition regulator TnsC interacts with the transposase subunit TnsB and target selector TnsD. *Proc. Natl. Acad. Sci. USA* 111, E2858–E2865.
7. 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. USA* 114, E7358–E7366.
8. 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.
9. 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.
10. 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.e18.
11. 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. USA* 118, e2112279118.
12. 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.e18.
13. Faure, G., Shmakov, S.A., Yan, W.X., Cheng, D.R., Scott, D.A., Peters, J.E., Makarova, K.S., and Koonin, E.V. (2019). CRISPR-Cas in mobile genetic elements: counter-defence and beyond. *Nat. Rev. Microbiol.* 17, 513–525.
14. 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.e5.

15. 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.
16. Kieper, S.N., Almendros, C., Behler, J., McKenzie, R.E., Nobrega, F.L., Haagsma, A.C., Vink, J.N.A., Hess, W.R., and Brouns, S.J.J. (2018). Cas4 facilitates PAM-compatible spacer selection during CRISPR adaptation. *Cell Rep.* *22*, 3377–3384.
17. Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Židek, A., Bates, R., Blackwell, S., Yim, J., et al. (2021). Protein complex prediction with AlphaFold-Multimer. <https://doi.org/10.1101/2021.10.04.463034>.
18. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Židek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* *596*, 583–589.
19. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nat. Methods* *19*, 679–682.
20. 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.
21. Park, J.U., Tsai, A.W.-L., Mehrotra, E., Petassi, M.T., Hsieh, S.-C., Ke, A., Peters, J.E., and Kellogg, E.H. (2021). Structural basis for target site selection in RNA-guided DNA transposition systems. *Science* *373*, 768–774.
22. Querques, I., Schmitz, M., Oberli, S., Chanez, C., and Jinek, M. (2021). Target site selection and remodelling by type V CRISPR-transposon systems. *Nature* *599*, 497–502.
23. Kuduvalli, P.N., Rao, J.E., and Craig, N.L. (2001). Target DNA structure plays a critical role in Tn7 transposition. *EMBO J.* *20*, 924–932.
24. Shi, Q., Straus, V., Caron, J.J., Wang, H., Chung, Y.S., Guarné, A., and Peters, J.E. (2015). Conformational toggling controls target site choice for the heteromeric transposase element Tn7. *Nucleic Acids Res.* *43*, 10734–10745.
25. Holm, L. (2022). Dali server: structural unification of protein families. *Nucleic Acids Res.* *50*, W210–W215.
26. Fujiyama, S., Abe, Y., Tani, J., Urabe, M., Sato, K., Aramaki, T., Katayama, T., and Ueda, T. (2014). Structure and mechanism of the primosome protein DnaT-functional structures for homotrimerization, dissociation of ssDNA from the PriB-ssDNA complex, and formation of the DnaT-ssDNA complex. *FEBS J.* *287*, 5356–5370.
27. Huang, C.-Y., Hsu, C.-H., Sun, Y.-J., Wu, H.-N., and Hsiao, C.-D. (2006). Complexed crystal structure of replication restart primosome protein PriB reveals a novel single-stranded DNA-binding mode. *Nucleic Acids Res.* *34*, 3878–3886.
28. Lovett, S.T. (2005). Filling the gaps in replication restart pathways. *Mol. Cell* *17*, 751–752.
29. Partridge, S.R., Kwong, S.M., Firth, N., and Jensen, S.O. (2018). Mobile genetic elements associated with antimicrobial resistance. *Clin. Microbiol. Rev.* *31*, e00088–e00017. <https://doi.org/10.1128/CMR.00088-17>.
30. Petassi, M.T. (2020). Dual Pathway Transposition with Tn7-Like Elements: Safe Sites, Mobile Plasmids, and CRISPR-Cas (Cornell University).
31. Peters, J.E., Fricker, A.D., Kapili, B.J., and Petassi, M.T. (2014). Heteromeric transposase elements: generators of genomic islands across diverse bacteria. *Mol. Microbiol.* *93*, 1084–1092.
32. Smyshlyayev, G., Bateman, A., and Barabas, O. (2021). Sequence analysis of tyrosine recombinases allows annotation of mobile genetic elements in prokaryotic genomes. *Mol. Syst. Biol.* *17*, e9880.
33. Bebel, A., Karaca, E., Kumar, B., Stark, W.M., and Barabas, O. (2016). Structural snapshots of Xer recombination reveal activation by synaptic complex remodeling and DNA bending. *eLife* *5*, e19706. <https://doi.org/10.7554/eLife.19706>.
34. Rubio-Cosials, A., Schulz, E.C., Lambertsen, L., Smyshlyayev, G., Rojas-Cordova, C., Forslund, K., Karaca, E., Bebel, A., Bork, P., and Barabas, O. (2018). Transposase-DNA complex structures reveal mechanisms for conjugative transposition of antibiotic resistance. *Cell* *173*, 208–220.e20.
35. Nero, T.M., Dalia, T.N., Wang, J.C.-Y., Kysela, D.T., Bochman, M.L., and Dalia, A.B. (2018). ComM is a hexameric helicase that promotes branch migration during natural transformation in diverse Gram-negative species. *Nucleic Acids Res.* *46*, 6099–6111.
36. Schmid-Burgk, J.L., Gao, L., Li, D., Gardner, Z., Strecker, J., Lash, B., and Zhang, F. (2020). Highly parallel profiling of Cas9 variant specificity. *Mol. Cell* *78*, 794–800.e8.
37. Iyer, L.M., Leippe, D.D., Koonin, E.V., and Aravind, L. (2004). Evolutionary history and higher order classification of AAA+ ATPases. *J. Struct. Biol.* *146*, 11–31.
38. Durrant, M.G., Li, M.M., Siranosian, B.A., Montgomery, S.B., and Bhatt, A.S. (2020). A bioinformatic analysis of integrative mobile genetic elements highlights their role in bacterial adaptation. *Cell Host Microbe* *28*, 767.
39. Touchon, M., Bobay, L.-M., and Rocha, E.P.C. (2014). The chromosomal accommodation and domestication of mobile genetic elements. *Curr. Opin. Microbiol.* *22*, 22–29.
40. Hsieh, S.-C., and Peters, J.E. (2023). Discovery and characterization of novel type I-D CRISPR-guided transposons identified among diverse Tn7-like elements in cyanobacteria. *Nucleic Acids Res.* *51*, 765–782. <https://doi.org/10.1093/nar/gkac1216>.
41. Rao, J.E., Miller, P.S., and Craig, N.L. (2000). Recognition of triple-helical DNA structures by transposon Tn7. *Proc. Natl. Acad. Sci. USA* *97*, 3936–3941.
42. Jiang, F., and Doudna, J.A. (2015). The structural biology of CRISPR-Cas systems. *Curr. Opin. Struct. Biol.* *30*, 100–111.
43. Eddy, S.R. (2011). Accelerated profile HMM searches. *PLoS Comput. Biol.* *7*, e1002195.
44. Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., et al. (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* *9*, 386.
45. Vos, M., Quince, C., Pijl, A.S., de Hollander, M., and Kowalchuk, G.A. (2012). A comparison of rpoB and 16S rRNA as markers in Pyrosequencing studies of bacterial diversity. *PLoS One* *7*, e30600.
46. Ogier, J.-C., Pagès, S., Galan, M., Barret, M., and Gaudriault, S. (2019). rpoB, a promising marker for analyzing the diversity of bacterial communities by amplicon sequencing. *BMC Microbiol.* *19*, 171.
47. Steinegger, M., and Söding, J. (2018). Clustering huge protein sequence sets in linear time. *Nat. Commun.* *9*, 2542. <https://doi.org/10.1038/s41467-018-04964-5>.
48. Wolf, Y.I., Kazlauskas, D., Iranzo, J., Lucía-Sanz, A., Kuhn, J.H., Krupovic, M., Dolja, V.V., and Koonin, E.V. (2018). Origins and evolution of the global RNA virome. *mBio* *9*, e02329–e02318. <https://doi.org/10.1128/mBio.02329-18>.
49. Edgar, R.C., Taylor, J., Lin, V., Altman, T., Barbera, P., Meleshko, D., Lohr, D., Novakovskiy, G., Buchfink, B., Al-Shayeb, B., et al. (2020). Petabase-scale sequence alignment catalyses viral discovery. <https://doi.org/10.1101/2020.08.07.241729>.
50. Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S.J., and Söding, J. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* *20*, 473.
51. Söding, J., Biegert, A., and Lupas, A.N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* *33*, W244–W248.
52. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* *25*, 1972–1973.
53. Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* *5*, e9490.

54. 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.
55. 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.
56. Mariani, V., Biasini, M., Barbato, A., and Schwede, T. (2013). IDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* *29*, 2722–2728.
57. Andreani, J., Faure, G., and Guerois, R. (2012). Versatility and invariance in the evolution of homologous heteromeric interfaces. *PLoS Comput. Biol.* *8*, e1002677.
58. Faure, G., Andreani, J., and Guerois, R. (2012). InterEvol database: exploring the structure and evolution of protein complex interfaces. *Nucleic Acids Res.* *40*, D847–D856.
59. Suzek, B.E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C.H. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* *23*, 1282–1288.
60. Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* *25*, 1422–1423.
61. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* *17*, 261–272.
62. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* *30*, 772–780.
63. Jones, D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* *292*, 195–202.
64. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* *25*, 3389–3402.
65. Altschul, S.F., and Lipman, D.J. (1990). Protein database searches for multiple alignments. *Proc. Natl. Acad. Sci. USA* *87*, 5509–5513.
66. Frickey, T., and Lupas, A.N. (2004). PhyloGenie: automated phylome generation and analysis. *Nucleic Acids Res.* *32*, 5231–5238.
67. Jayaram, M., Ma, C.H., Kachroo, A.H., Rowley, P.A., Guga, P., Fan, H.F., and Voziyanov, Y. (2015). An overview of tyrosine site-specific recombination: from an *flp* perspective. *Microbiol. Spectr.* *3*, MDNA3-0021-2014. <https://doi.org/10.1128/microbiolspec.MDNA3-0021-2014>.
68. Bland, C., Ramsey, T.L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N.C., and Hugenholtz, P. (2007). CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* *8*, 209.
69. Altae-Tran, H., Kannan, S., Demircioglu, F.E., Oshiro, R., Nety, S.P., McKay, L.J., Dlakić, M., Inskeep, W.P., Makarova, K.S., Macrae, R.K., et al. (2021). The widespread IS200/IS605 transposon family encodes diverse programmable RNA-guided endonucleases. *Science* *374*, 57–65.



STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
<b>Bacterial strains</b>		
One Shot™ PIR1 Chemically Competent E. coli	ThermoFischer	C101010
BL21(DE3) Competent E. coli	New England Biolabs	C2527
BL21(DE3) Electrocompetent Cells	Millipore Sigma	CMC0016
Endura™ ElectroCompetent Cells	Lucigen	60242
One Shot™ Stbl3™ Chemically Competent E. coli	ThermoFischer	C737303
NEB® 10-beta Competent E.coli (High Efficiency)	New England Biolabs	C3019
<b>Chemicals, peptides, and recombinant proteins</b>		
NEBNext® High-Fidelity 2X PCR Master Mix	New England Biolabs	M0541
Q5® High-Fidelity 2X Master Mix	New England Biolabs	M0492
KAPA HiFi HotStart ReadyMix	Roche	KK2602
KOD Hot Start DNA Polymerase	Millipore Sigma	71086-3
E-Gel™ EX Agarose Gels, 1%	ThermoFischer	G401001
E-Gel™ EX Agarose Gels, 2%	ThermoFischer	G401002
Novex™ TBE Gels, 6%, 15 well	ThermoFischer	EC62655
Novex™ Hi-Density TBE Sample Buffer (5X)	ThermoFischer	LC6678
SYBR™ Gold Nucleic Acid Gel Stain (10,000X Concentrate in DMSO)	ThermoFischer	S11494
Wizard® SV Gel and PCR Clean-Up System	Promega	A9282
Wizard® Genomic DNA Purification Kit	Promega	A1120
QIAquick PCR Purification Kit	Qiagen	28106
QIAprep Spin Miniprep Kit	Qiagen	27106
PureYield™ Plasmid Midiprep System	Promega	A2495
Gibson Assembly® Master Mix	New England Biolabs	E2611
NEBuilder® HiFi DNA Assembly Master Mix	New England Biolabs	E2621
Q5® Site-Directed Mutagenesis Kit	New England Biolabs	E0554
BstZ171-HF®	New England Biolabs	R3594
NruI	New England Biolabs	R0192
BsaI-HF®v2	New England Biolabs	R3733
BbsI-HF®	New England Biolabs	R3539
AarI	ThermoFischer	ER1581
T4 DNA Ligase	New England Biolabs	M0202
IPTG	Goldbio	I2481C
S-Gal®/LB Agar Blend	Millipore sigma	C4478
ddPCR Supermix for Probes (No dUTP)	Bio-Rad	#1863024
Droplet Generation Oil for Probes	Bio-Rad	#1863005
NEBNext® Companion Module for Oxford Nanopore Technologies® Ligation Sequencing	New England Biolabs	E7180S
Ligation Sequencing Kit	Oxford Nanopore Technologies	SQK-LSK109
Flow Cell (R9.4)	Oxford Nanopore Technologies	FLO-MIN106
AMPure XP for PCR Purification	Beckman Coulter	A63881
Tn5	Schmid-Burgk et al. <sup>36</sup>	N/A
cComplete Protease Inhibitor Cocktail	Millipore sigma	4693116001

(Continued on next page)

**Continued**

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Strep-Tactin® Sephaarose® Resin	IBA	2-1201-002
Strep-Tactin® Magnetic Microbeads	IBA	6-5510-050
D-Desthiobiotin	Millipore sigma	71610-3
Ulp1 SUMO protease	F. Zhang Lab	N/A
SENP protease	F. Zhang Lab	N/A
Amicon Ultra-15 Centrifugal Filter Units 10kDa NMWL	Millipore sigma	UFC901024
Amicon Ultra-15 Centrifugal Filter Units 50kDa NMWL	Millipore sigma	UFC905024
NuPAGE™ 4–12% Bis-Tris Protein Gels, 1.0 mm, 12-well	ThermoFischer	NP0322BOX
NuPAGE™ LDS Sample Buffer (4X)	ThermoFischer	NP0007
Imperial™ Protein Stain	ThermoFischer	24615
Ampicillin, sodium salt	AmericanBio	Ab00115
Carbenicillin disodium salt, 89.0–100.5% anhydrous basis	Millipore sigma	C1389
Spectinomycin dihydrochloride pentahydrate	Millipore sigma	S4014
Kanamycin sulfate from Streptomyces kanamyceticus	Sigma	K4000
Chloramphenicol	Sigma	C0378
MiSeq Reagent Kits v2	Illumina	MS-102
NextSeq 500/550 High Output Kit v2, 75 cycles	Illumina	FC-404-2005
<b>Critical commercial assays</b>		
Qubit 1X dsDNA HS (High-Sensitivity) Assay Kit	ThermoFischer	Q33231
eStain L1 Protein Staining System	GenScript	N/A
<b>Deposited data</b>		
Deep sequencing data	SRA	SRA: PRJNA913200
InsertionReadsCounter	Zenodo	10.5281/zenodo.7872374
<b>Oligonucleotides</b>		
CTTTCCCTACACGACGCTCTCCG ATCTgagcaagagattacgagcagac	Genewiz	NGS pTarget 6N-upstream primer
CTTTCCCTACACGACGCTCTCCG ATCTctaccgcattaaagctccgcc	Genewiz	NGS pTarget 6N-downstream primer
GACTGGAGTTCAGACGTGTGCTC TTCCGATCTgcctattgcttcgctctatctgtcc	Genewiz	NGS pDonor(CyCAST)-LE primer
GACTGGAGTTCAGACGTGTGCTCTT CCGATCTcctaaggcaacacaacggctg	Genewiz	NGS pDonor(CyCAST)-RE primer
CTTTCCCTACACGACGCTCTCCG ATCTgtaagccttgaggtcgattctcg	Genewiz	NGS tRNA-leu primer
ctaccgcattaaagctccgcc	Genewiz	pTarget ddPCR 6N-downstream primer
cctaaggcaacacaacggctg	Genewiz	pDonor(CyCAST) ddPCR for RE primer
cctggtgtccctgtgataccg	Genewiz	pDonor(CyCAST) ddPCR tRNA-leu downstream primer
cgacagcatgccagctcactatg	Genewiz	pTarget ddPCR for TcR1 primer
caagtagcgaagcgagcaggac	Genewiz	pTarget ddPCR for TcR2 primer
CTTTCCCTACACGACGCTCTCCGA TCTagctccttagctcctgaaatctcg	Genewiz	NGS AjTn6022-comM upstream primer

(Continued on next page)

**Continued**

REAGENT or RESOURCE	SOURCE	IDENTIFIER
CTTCCCTACACGACGCTCTCCGA TCTgaaaatgagacgttgatcgccac	Genewiz	NGS AjTn6022-comM downstream primer
GACTGGAGTTCAGACGTGTGCTCT TCCGATCTgccactcttgcttattactgtc	Genewiz	NGS pDonor(AjTn6022)-RE primer
GACTGGAGTTCAGACGTGTGCTCTT CCGATCTtgcaggcatactcttttagtg	Genewiz	NGS pDonor(AjTn6022)-LE primer
gccactcttgcttattactgtc	Genewiz	pDonor(AjTn6022) ddPCR for RE primer
gaaaatgagacgttgatcgccac	Genewiz	pTarget ddPCR Aj-comM downstream primer
CTTCCCTACACGACGCTCTCCGA TCTctgggtgaaggctctcaaggc	Genewiz	NGS ZooTsy-comM upstream primer
CTTCCCTACACGACGCTCTCCGA TCTgtgtgcttcaaatgcctgaggttc	Genewiz	NGS ZooTsy-comM downstream primer
GACTGGAGTTCAGACGTGTGCTCTT CCGATCTctcaagtaatgcgacattgg	Genewiz	NGS ZooTsy-end1 (for upstream-end1 junction) primer
CTTCCCTACACGACGCTCTCCG ATCTcctttatagtcagtggttatccg	Genewiz	NGS ZooTsy-end1 (for circularized donor junction) primer
GACTGGAGTTCAGACGTGTGCTCT TCCGATCTtgcagatcatgcatgagctcac	Genewiz	NGS KanR-cargo upstream primer
GACTGGAGTTCAGACGTGTGCTCTT CCGATCTgaggatccaacattccaatcactag	Genewiz	NGS KanR-cargo downstream primer
ctgcaagtaatgcgacattgg	Genewiz	pDonor(ZooTsy) ddPCR for end1 primer
atcaaaactggtgaaactaccaccag	Genewiz	pTarget ddPCR for CmR1 primer
gtgtcaccctgttacaccgttttc	Genewiz	pTarget ddPCR for CmR2 primer
ggtgatgacggtgaaaacctctgac	Genewiz	ZooTsy_CI-sequencing primer
ggtaaatgctgaatcagtacaaaaacaatg	Genewiz	E.coli comM upstream primer for AjTn6022
gattcaatcggtctctgcaaggc	Genewiz	E.coli comM downstream primer for AjTn6022
agtgtcgcgataattaagtctggtaaatg	Genewiz	E.coli comM upstream primer for ZooTsy
ccgagagccggtgagataacg	Genewiz	E.coli comM downstream primer for ZooTsy
atgtcactgtcaattgtcataccgc	Genewiz	E.coli comM ddPCR forward primer
ttttttccggatattcatatccgc	Genewiz	E.coli comM ddPCR reverse primer
gtctcgtgggctcgagatgtgtataagagacag	Genewiz	TTISS 1st PCR common primer
gtggcaacctattgtttcttatcatgac	Genewiz	AjTn6022-TTISS_RE 1st primer
AATGATACGGCGACCACCGAGATCTA CACAAGTAGAGACTCTTTCCCTACA CGACGCTCTCCGATCTgacactcttatcta ttgctgtaaatgac	IDT	AjTn6022-TTISS_RE 2nd F1 primer
tcctttatagtcagtggttatccg	Genewiz	ZooTsy-TTISS_End1 1st primer
AATGATACGGCGACCACCGAGATCTACA CAAGTAGAGACTCTTTCCCTACACGAC GCTCTCCGATCTgcctgttctgtctacattggc	IDT	ZooTsy-TTISS_End1 2nd F1 primer
CAAGCAGAAGACGGCATAACGAGATCATG ATCGGTCTCGTGGGCTCGGAGATGTGT	Genewiz	TTISS 2nd PCR common R1 primer
CAAGCAGAAGACGGCATAACGAGATAGGAT CTAGTCTCGTGGGCTCGGAGATGTGT	Genewiz	TTISS 2nd PCR common R2 primer
CAAGCAGAAGACGGCATAACGAGATGACAG TAAGTCTCGTGGGCTCGGAGATGTGT	Genewiz	TTISS 2nd PCR common R3 primer

(Continued on next page)

**Continued**

REAGENT or RESOURCE	SOURCE	IDENTIFIER
CAAGCAGAAGACGGCATACGAGATCCTATG CCGTCTCGTGGGCTCGGAGATGTGT	Genewiz	TTISS 2nd PCR common R4 primer
TGCGTTGATGCAATTTCTATGCGCACCCGT	IDT	pTarget ddPCR TcR probe
TGTCCACACCCATGAGTGGACAACCTTATGC	IDT	pDonor(CyCAST) ddPCR for RE probe
TCATGATAAGAAAACAATAGTTGCCACTC	IDT	pDonor(AjTn6022) ddPCR for RE probe
CAACGAACAAGGCAAAAATTCGGATAACCC	IDT	pDonor(ZooTsy) ddPCR for end1 probe
AGGTTTTACCGTAACACGCCACATCTTGC	IDT	pTarget ddPCR CmR probe
AGGTCTACCCGGCTTAACGATGGTGGGCT	IDT	E.coli comM probe
<b>Recombinant DNA</b>		
pUC19 Vector	New England Biolabs	N3041
pBluescript II SK (+)	Agilent	212205
pCDFDuet™-1 DNA	Millipore sigma	71340
pCOLADuet™-1 DNA	Millipore sigma	71406
pACYCDuet™-1 DNA	Millipore sigma	71147
pSC101-Donor	Addgene	#140630
pXT131_TwinStrep-SUMO-ShTniQ	Addgene	#135527
<b>Software and algorithms</b>		
Geneious	<a href="https://www.geneious.com/">https://www.geneious.com/</a>	v2022
Blast+	<a href="https://blast.ncbi.nlm.nih.gov/Blast.cgi">https://blast.ncbi.nlm.nih.gov/Blast.cgi</a>	v2.9.0
Hmmer	<a href="http://hmmer.org/">http://hmmer.org/</a>	v3.1
iTOL	<a href="https://itol.embl.de/">https://itol.embl.de/</a>	N/A
MAFFT	<a href="https://mafft.cbrc.jp/">https://mafft.cbrc.jp/</a>	v7.505
Muscle	<a href="https://drive5.com/muscle5/">https://drive5.com/muscle5/</a>	v5
PyMOL	<a href="https://pymol.org/2/">https://pymol.org/2/</a>	v1.2
Fasttree	<a href="http://www.microbesonline.org/fasttree/">http://www.microbesonline.org/fasttree/</a>	v2.1.10
DALI	<a href="http://ekhidna2.biocenter.helsinki.fi/dali/README.v5.html">http://ekhidna2.biocenter.helsinki.fi/dali/README.v5.html</a>	v5
HHsuite	<a href="https://github.com/soedinglab/hh-suite">https://github.com/soedinglab/hh-suite</a>	v3.1
psipred	<a href="https://github.com/psipred/psipred">https://github.com/psipred/psipred</a>	v2.6
HHpred webserver	<a href="https://toolkit.tuebingen.mpg.de/tools/hhpred">https://toolkit.tuebingen.mpg.de/tools/hhpred</a>	N/A
Mmseqs2	<a href="https://github.com/soedinglab/mmseqs2">https://github.com/soedinglab/mmseqs2</a>	v12
trimal	<a href="http://trimal.cgenomics.org/trimal">http://trimal.cgenomics.org/trimal</a>	v1.2
AlphaFold2 Colabfold	<a href="https://github.com/sokrypton/ColabFold">https://github.com/sokrypton/ColabFold</a>	N/A
AlphaFold2 Multimer colabfold	<a href="https://github.com/sokrypton/ColabFold">https://github.com/sokrypton/ColabFold</a>	N/A
Biopython	<a href="https://biopython.org/">https://biopython.org/</a>	v1.78
CLANS	<a href="http://ftp.tuebingen.mpg.de/pub/protevo/CLANS/">http://ftp.tuebingen.mpg.de/pub/protevo/CLANS/</a>	N/A
InsertionReadsCounter	<a href="https://doi.org/10.5281/zenodo.7872374">https://doi.org/10.5281/zenodo.7872374</a>	v3.0.0
QuantaSoft™ Software	Bio-Rad	N/A
<b>Other</b>		
Bench Protocol	This paper	STAR Methods

## RESOURCE AVAILABILITY

### Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Feng Zhang ([zhang@broadinstitute.org](mailto:zhang@broadinstitute.org)).

### Materials availability

Plasmids generated in this study have been deposited to Addgene ([Data S3](#)).

### Data and code availability

- All Illumina NGS and Oxford Nanopore Technologies (ONT) sequencing data generated from this publication have been deposited and are publicly available as of the date of publication. Accession numbers are listed in the [key resources table](#).
- All original code for transposition junction NGS reads analysis has been deposited to GitHub and Zenodo. DOI are listed in the [key resources table](#).
- Any additional information required to reanalyze the data reported in this paper is available from the [lead contact](#) upon request.

## METHOD DETAILS

### Identification of Tn7-like transposons

HMMs of Tn7 proteins (TnsA (PF08721.12, PF08722.12), TnsB (PF00665.27), TnsC (PF05621.12, PF11426.9, PF13401.7), TnsD (PF06527.12, PF15978.6), and TnsE (PF18623.2)) were used to search homologs with *hmmsearch* software<sup>43</sup> (using option *ga\_cut*) within predicted protein sequences derived from publicly available microbial contigs in the NCBI Genbank and WGS databases, JGI database (projects with stated permission to use), and the MG-RAST database<sup>44</sup> (all frozen in November 2020). The full database encompasses 521,828,662 contigs in total (contigs greater than 1.5kb), covering 6,932,321,054,498 bp of genomic DNA. 1,617,895 contigs have detectable *rpob* genes (detected with *hmmsearch* from TIRGR02013.1 profile), suggesting the diversity is probably reflected by 1.6 million genomes.<sup>45,46</sup> Loci were built by mapping the location of the hits into the contig and aggregating hits when they are no further than 20 kb from each other. Loci were selected if they satisfied the following criteria: (i) at least 2 hit genes homologous to 2 distinct Tn7 components, (ii) 2 of the hit genes are in a putative operon, which is operationally defined as 2 codirectional ORFs separated by less than 50 bp of non-coding sequence, and (iii) the hit genes are less than 3 kb from the contig boundary (to remove likely incomplete transposons). Using these criteria, 80,028 loci were obtained from which we extracted and translated 86,517 TnsC homologs. We clustered these homologs at 80% sequence identity over 75% of the protein length (coverage) using *MMSeqs2* (v. 12- 113e3)<sup>47</sup> and obtained 7,789 TnsC homolog representatives.

### Construction of the phylogenetic tree of TnsC homologs

The protein sequences of representative TnsC homologs were aligned using a method described previously.<sup>48</sup> Briefly, the protein sequences were clustered at 50% identity and cluster members were extracted and aligned using *MUSCLE* (version 5).<sup>49</sup> An all-versus-all comparison of the multiple sequence alignments (MSAs) was computed using *HHsearch*.<sup>50</sup> An unweighted pair group method with arithmetic mean (UPGMA) dendrogram was constructed using the *HHSearch* similarity scores. The dendrogram was used to guide the iterative pairwise alignment of cluster MSAs using *HHalign*.<sup>50,51</sup> Clusters were discarded if they could not be aligned using this approach, leaving 6,988 sequences in the single final alignment. The alignment was first filtered to remove sites with conserved gaps using *trimal* version 1.2 with the option *gappyout*.<sup>52</sup> Finally, aligned sequences from which Walkers A and B were not aligned correctly (any gap in one of the positions) were discarded. This led to an alignment of 6,384 sequences that are used for this study. The alignment was input into *FastTree2*<sup>53</sup> with the Whelan-Goldman models of amino acid evolution and gamma-distributed site rates. The tree was visualized and annotated using the interactive tree of life (*itol*).<sup>54</sup>

### Annotation of the genomic neighborhoods of TnsC homologs

Representative TnsC homologs were mapped on the respective genomic contigs and genes within 50 kb were extracted. These genes were translated and annotated for specific genes of interest including Cas effectors (Cascade and RAMP components and single protein effector from Class 2 CRISPR-Cas systems) and Cas6 that were annotated using the profiles extracted from *DefenseFinder*<sup>55</sup> and *hmmsearch* with a threshold set at a score of 25 whereby all hits with a score of 25 or greater were selected. Hits were mapped onto leaves of the TnsC homolog tree and subsequently the TniQ homolog tree and displayed with *itol* ([Figures 2 and S2](#)). As TniQ/TnsD can be extremely divergent and therefore not always detected by *hmmsearch*, a profile-profile comparison was performed using *HHalign* software, and the TniQ PF06527 profile was used to annotate distant TniQ homologs. TniQ hits were selected if they have a *hhsearch* probability  $\geq 80$ . Given the abundance of TniQ, false positive detection of TnsC (AAA-ATPase among cargo), and sometimes multiple Tn7 transposons co-occurring, it is challenging to know which *tniQ/tnsD* is associated with which *tnsC*. For a given *tnsC*, we defined an association with *tniQ/tnsD(s)* if no other *tnsC* homologs were found in the vicinity. If there was a *tnsC* homolog closer than the given *tnsC*, we associated *tniQ/tnsD* to it only if there was a *tnsB* gene operonized with

this closer *tnsC*. In total 5,072 *tniQ/tnsD* were found to be associated with TnsC homolog representatives. The presence of these *tniQ/tnsD* were mapped onto the TnsC homolog tree. When multiple *tniQ/tnsD* were found in the vicinity, the two largest were selected and the second selected *tniQ/tnsD* were mapped on an additional barplot indicating the size of the second protein in aa. *tnsE* were detected in relatives of *E. coli* Tn7 but also weakly in Tn6022-like transposons where it is encoded in the far vicinity of the Tn core components (in contrast to *E. coli* Tn7 relatives where *tnsE* is part of a full operon encompassing all Tn core components and *tnsD*). Manual inspection indicated these TnsE remote homologs are encoded near the left end of the transposon far from the other transposon components. Structural comparison from the models predicted by AlphaFold2<sup>18</sup> (Figure S3) confirmed that these hits are TnsE relatives. These remote TnsE were used as a new seed to annotate additional TnsEs using blastp on all translated genes in the vicinity of *tnsC*. 552 TnsE in total were detected, and their presence was mapped on the TnsC homolog tree and on the TniQ homolog tree and displayed using the itol framework.

### Protein structure prediction and analysis

All structural models were built using AlphaFold2 (AF2) software under the colabfold framework installed locally.<sup>17–19</sup> Multiple sequence alignments were constructed using colabfold\_search on the colabfold database that includes Uniref and environmental protein sequences. Alignments were input into AF2, and three models were generated with 35 recycles. All models were examined using the PyMOL framework (The PyMOL Molecular Graphic System Version 1.2, Schrodinger, LLC), mapping the predicted local distance difference test (pLDDT),<sup>18,56</sup> a local measure of prediction confidence, on each residue. Regions of the proteins with pLDDT less than 50 were not considered. Protein docking prediction was performed using AF2 with the multimeric model.<sup>17</sup> Results of protein docking were analyzed by examining the predicted aligned error (PAE) matrix<sup>18</sup> and visualizing the interaction area on PyMOL. The spatial distributions of specific chemical interactions found in protein-protein interactions<sup>57,58</sup> were analyzed using PyMOL to validate models when PAE was weak. Predicted structures with high confidence (typically average pLDDT>70) were considered for downstream analysis. Searches for structural similarity were performed using DALI software version 5<sup>25</sup> using the PDB50 (non-redundant at 50% of sequence identity) and a custom database made from the EBI AlphaFold2 database. This database contains AlphaFold2 models of Uniref50<sup>47,59</sup> extracted from EBI AlphaFold2 database (<https://alphafold.ebi.ac.uk/>) restricted to models having at least 30aa with a pLDDT greater than 50 (hereafter called AF2DB50). Hits with Z-scores greater than 5 were retained, and every hit was manually verified by building a structural alignment in PyMOL editing mode.

### Phylogenetic analysis of TniQ

5,072 *tniQ/tnsD* detected in the vicinity of the Tn7 *tnsC* homolog representatives were translated and aligned using MUSCLE version 5 using the option super5 to cluster the sequences. Alignments were then created for each cluster and merged into a single alignment. The alignment was first filtered to remove sites with conserved gaps using trimal version 1.2 with the option gappypout. As TniQ/TnsD can harbor very divergent C-terminal regions in terms of both sequence and size, the alignment was restricted to the core region of TniQ. To determine the core (roughly corresponding to the first 300aa), the structure of all dual TniQ-TnsD CAST systems was predicted with AF2 and manually aligned structurally (as described above). Core positions were mapped to the sequence alignment and the downstream regions (C-terminal region) were trimmed out. Finally, to filter out misaligned TniQ cores, aligned sequences from which the first CxxC motif of the zing finger were not well aligned (any gap in one of the positions) were discarded. The 4,916 remaining aligned sequences were used to build a tree using FastTree2 with the Whelan-Goldman models of amino acid evolution and gamma-distributed site rates. The presence of *cas* effector genes and *cas6* as well as *tnsE* in the vicinity was shown similarly as for the TnsC tree. Operonized genes with *tniQ/tnsD* (see TniQ partner candidate analysis) were also displayed (light green ring). To investigate the origin of dual TniQ-TnsD, a visual approach and a statistical approach were used. A connector between dual TniQ-TnsD leaves of the tree was drawn and colored based on a rainbow gradient spread across all leaves from the left to the right of the tree. The connector has a uniform color picked from the color assigned to the left-most leaf. If a connector has a color that matched the color both at the starting leaf (left) and the arriving leaf (right), the proteins are closely related. Conversely, any contrast between the colors indicates the proteins are not closely related. For the statistical approach, the branch distances of all dual TniQ-TnsD were extracted and compared with 1000 random branch distances involving nondual TniQ-TnsD across the tree. Branch distances were calculated using the Phylo package from the Biopython library. Biopython package.<sup>60</sup> Comparison of these distances was done via a T-Test using SciPy python library version 1.0.<sup>61</sup>

### Analysis of candidate partners of TniQ

Genes operonized with *tniQ/tnsD* were extracted and translated if they were not related to any Tn7 components (TnsA, TnsB, TnsC, TniQ/TnsD, or TnsE). 782 proteins were clustered at 30% of sequence identities and 30% sequences coverages using MMSeqs2 (v. 12- 113e3).<sup>47</sup> From each cluster, members were aligned together using mafft-linsi.<sup>62</sup> From each alignment, secondary structures were predicted psipred version 2.6<sup>63</sup> to ensure compatibility with hhpred, and HMM profiles were built for each using hhmake.<sup>51</sup> Each hhpred profile was compared to the Pfam protein domain database (preformatted for hhpred and available at [https://wwwuser.gwdg.de/~compbiol/data/hhsuite/databases/hhsuite\\_dbs/](https://wwwuser.gwdg.de/~compbiol/data/hhsuite/databases/hhsuite_dbs/)) using hhsearch. Hits with hhsearch probability  $\geq 90$  were considered. Candidates operonized with *tniQ/tnsD* were mapped onto the tree (light green ring). The 6 largest groups of candidates were selected based on the conservation of the operonized genes across several adjacent leaves in the TniQ tree and analyzed further. For each candidate, we performed profile analysis (using hhpred webserver) to assess their potential function and structural

docking with TnsC and TniQ/TnsD to test for potential interactions (interaction with TnsC could highlight a novel target selector independent from TniQ/TnsD, whereas interaction with TniQ/TnsD could suggest a partner target selector) using AlphaFold multimer.

### TnsF analysis

TnsF from *Acinetobacter johnsonii* Tn6022 (hereafter AjTn6022 and AjTnsF) was chosen as the representative for computational and experimental analysis. HHpred was used to annotate the domain architecture region of AjTnsF and detect the presence of multiple zinc fingers in the N-terminal region (positions 1-180 has a hit to a LIM Zinc-binding domain-containing protein hhprob=97.23) and similarity with tyrosine recombinase in the C-terminal region (positions 328-493 has a hit to a site-specific recombinase IntI4 hhprob=99.14). A structural model of TnsF was obtained using AF2 and split into 3 domains defined by long linkers connecting globular regions and a long N-terminal region encoding several zinc fingers. Each domain was used as a seed for structural similarity search using DALI software across the PDB50 (as described above). Top hits were inspected manually with PyMOL.

### Mining of TnsF and phylogenetic analysis of TnsF homologs

To search for TnsF relatives, AjTnsF was used as a seed for a psiblast<sup>64</sup> search for 3 iterations on the NCBI NR database (in August 2022). 22,419 protein hits were extracted and clustered at 80% of sequence identity and 70% of coverage with MMSeqs2 (v. 12-113e3). A blastall comparison<sup>65</sup> was performed, and e-values associated with each comparison were input into CLANS software<sup>66</sup> to cluster hits according to their e-values and draw a graph network representation. Several clusters (point density connected and close to each other) are also connected to each other. AjTn6022 TnsF was mapped onto the graph to identify the cluster to which it belongs in order to define the Tn6022 TnsF cluster. The Tn6022 TnsF cluster connected to another cluster from which several members were extracted and mapped onto genomic contigs. Genomic comparison between these contigs and Tn6022 contigs reveals a distinct system with partial *comM* surrounding the system and with no apparent Tn7 components but other genes operonized with *tnsF*. HHpred webserver<sup>51</sup> was used to annotate these genes, revealing the presence of a gene encoding a tyrosine recombinase (*yrec*), a gene encoding a helix turn helix domain (*hth*), and a gene encoding a GIY-YIG nuclease. The system was named transposon using Target Selector based on tyrosine (Y) recombinase (Tsy) based on the components operonized with *tnsF*. Hits belonging to the CLANS Tn6022 TnsF cluster and the adjacent cluster harboring the Tsy TnsF were extracted and aligned using MUSCLE version 5 with the super5 algorithm. The alignment of 1,095 protein sequences was further trimmed using trimal version 1.2 (gappout option) and was input to FastTree2 with the Whelan-Goldman models of amino acid evolution and gamma-distributed site rates. The tree was visualized and annotated using the interactive tree of life (itol). *tnsF* genes were mapped to genomic contigs, and genes in the vicinity (20 kb) were extracted, translated, and further clustered at 30% of sequence identity retaining 50% of coverage using MMSeqs2 (v. 12-113e3), and each cluster was converted into HMM profile using hhmake and compared to the profile pfam database using HHsearch. The top populated clusters were Tn7 components (*tnsA*, *tnsB*, *tnsC*, and *tniQ/tnsD*) and candidates operonized with Tsy TnsF (*yrec*, *GIY-YIG*). The presence of *tniQ/tnsD* as a marker of Tn7, and the presence of the *yrec* and the *GIY-YIG* nuclease were mapped on the TnsF tree as distinct rings. Split *comM* genes were extracted and translated from AjTn6022 and one Tsy locus and used as seeds to search for the full protein version of Mg chelatase for Tn6022 and Tsy using blastp. The closest Mg chelatase was selected for each of the two systems and used as a seed to detect *comM* pieces in the nucleotide vicinity of each locus using tblastn. The presence of a *comM* hit is indicated as a ring on the TnsF tree. Inspection of the tree shows Tn6022 TnsF is monophyletic (branch support = 0.976). A TnsF from Tsy was extracted from *Zoogloea sp. LCSB751* (ZooTnsF) and the structure was predicted using AF2. Structural comparison between AjTnsF and ZooTnsF was performed manually using a PyMOL framework. The structural similarity search was done using DALI on AF2DB50 (as described above). Hits with a zscore greater than 8 were inspected manually to search for tandem CB+CAT domain architecture using PyMOL.

### Determination of transposon ends for Tn6022 and Tsy

Transposon ends for AjTn6022 were determined using Geneious searching for at least one distinct cluster of short repeats (12 nt with 3 mismatches maximum, repeated at least twice in each end) that surround the transposon components including the *tnsA*, *tnsB*, *tnsC*, *tniQ*, *tnsF*, and *tnsE*. Exact end boundaries were then adjusted manually based on local alignment of the clustered repeats area and search for target site duplications. Transposon ends for ZooTsy were determined by prediction based on previous findings about YRec combined with experimental validation (Figure S5). YRec usually works as a dimer to recognize a region with two DNA motifs (each bound by the CB domain of each monomer) and cleave the middle region surrounding these motifs during recombination.<sup>67</sup> Based on this, we reasoned that in an excision scenario where the cleavage site for excision is at the edge of the partial *comM* gene, one motif should be located within the *comM* gene while the other motif should be located downstream of *comM* in the end of the transposon which would lead to a cleavage site at the transition area between the *comM* gene and the end region. To test this, we first cloned the ends of ZooTsy – 135-bp end1 (the region upstream of YRec extending to the border of the 5'-terminal portion *comM*) and 39-bp end2 (the region from the end of the cargo extending to the border of the 3'-terminal portion of *comM*). We then performed transposition assays, initially testing five different extensions (called homology arms (hom): 100-, 50-, 25-, 12-, 0-bp) for each end with the *comM* sequence upstream end1 (hom1) and downstream end2 (hom2) to determine if *comM* itself encoded a motif for end recognition. Based on these results, we concluded that 12 bp are required for hom1 (the requirement for hom2 was inconclusive). We then further refined this initial construct (end1:135bp, end2:39bp, hom1:12bp and hom2:12bp) to determine the minimal requirements for transposition (Figure S5F).

### Computational analysis of CAST I-D

Given the rarity of CAST I-D in our frozen database, we performed a blastp from CAST I-D Cas10d on the NCBI NR database to fetch additional loci. Only one candidate, from the *Cyanotheca sp. PCC 7425* genome (CyCAST), harbors a predicted active HD and Cas3'. The CRISPR-array of CyCast was predicted using minced<sup>68</sup> with default parameters, and transposon ends were determined as described above. Exact end boundaries were then adjusted manually based on local alignment of the clustered repeats area and search for target site duplication. CAST I-D Cas10d was used as a seed to search for homologs using blastp on the NCBI NR database. An initial phylogenetic tree (using FastTree) was done from the sequence alignment (using mafft-linsi) of the hits, and the subtree harboring all CAST I-D Cas10d and a few non-CAST I-D Cas10d were extracted. A structural model of CyCAST Cas10d was obtained using AF2 and used to determine the position of the catalytic sites of the HD nuclease domain. These positions were mapped onto the alignment from which a sub-alignment encompassing these positions and their local sequential vicinity was built and juxtaposed next to the tree.

### Plasmid construction

All plasmids used in this study are described in [Data S3](#). For type I-D CyCAST experiments, genes encoding *Cyanotheca sp. PCC 7425* (ASM2204v1\_genomic|CP001344.1|5374574) TnsAB-TnsC-TniQ-TnsD, and Cas10d-Csc2-Cas5-Cas6 were synthesized (Twist Bioscience) and cloned into pCDFDuet-1 with the appropriate spacer flanked by two CRISPR direct repeats, yielding pHelper(CyCAST). Gene fragments encoding both transposon ends (Genewiz) were cloned into pBluescript II SK (+) (Agilent), yielding pDonor(CyCAST). For PAM screens, a 0.5-kb exon fragment amplified from human *EMX1* was inserted between the transposon ends as a mock non-functional cargo in *E. coli*. For AjTn6022 experiments, genes encoding *Acinetobacter johnsonii* Tn6022 locus TnsA-TnsB-TnsC-TniQ-TnsF were synthesized (Twist Bioscience) and cloned into pCDFDuet-1. Predicted AjTn6022 transposon ends (RE and LE) were synthesized (Genewiz) and cloned into pBluescript II SK (+) with a kanamycin resistance gene to obtain pDonor(AjTn6022). For ZooTsy experiments, genes encoding *Zoogloea sp. LCSB751* Tsy locus YRec-HTH-TnsF-GIY-YIG nuclease were synthesized (Twist Bioscience) and cloned into pCDFDuet-1. Predicted ZooTsy transposon ends (end1 and end2) were synthesized (Genewiz) and cloned to obtain pDonor(ZooTsy). Throughout the study, pBluescript-donor was used for CyCAST, and pR6K-donor (backbone: Addgene#127924) was used for AjTn6022 and ZooTsy experiments, except for Tagmentation-based Tag Integration Site Sequencing (TTISS) analyses, for which pSC101-donor was used (backbone: Addgene#140630). To construct pHelper variants for CyCAST, AjTn6022, and ZooTsy, the Q5 Site-Directed Mutagenesis Kit (NEB) was used. For pHelper(AjTn6022) TniQ point mutants, a ribosome binding site and a partial TnsF N-terminal coding sequence was inserted to separate overlapped ORFs of TniQ and TnsF, and mutations were introduced into TniQ. The tRNA-Val gene of pTarget(PmcCAST) (Addgene#168163) was replaced by either tRNA-Leu gene of *Cyanotheca*, *comM* gene fragment of *A. johnsonii*, or *comM* gene fragment of *Zoogloea sp. LCSB751*, yielding pTarget(CyCAST), pTarget(AjTn6022), or pTarget(ZooTsy), respectively. For TniQ and TnsF protein purification, TniQ of AvCAST, PmcCAST, CyCAST, Tn7017, and AjTn6022 were individually cloned into ColE1-based pTwinStrep-SUMO bacterial expression vector, and ShCAST (Addgene#135527) were used. TnsF of AjTn6022 and nuclease-dead TnsF-Y584F mutant of ZooTsy were individually cloned into ColE1-based pTwinStrep-bdSUMO bacterial expression vector.

### *E. coli* transposition assays

For CyCAST, 100 ng of pHelper(CyCAST) was co-electroporated with 100 ng of pDonor(CyCAST) and 100 ng of pTarget(CyCAST) into BL21(DE3) electrocompetent cells (Sigma) and plated on 100 µg/ml carbenicillin, 50 µg/ml spectinomycin, and 50 µg/ml chloramphenicol containing LB-agar plates. After incubation for 17 hours at 37°C, all colonies were scraped from the plates, and a portion was re-plated on 0.1 mM IPTG supplemented triple antibiotic LB-agar plates to induce protein expression. Cells were incubated for an additional 17 hours at 37°C. All colonies were scraped, and the plasmid DNA was purified using QIAprep Spin Miniprep Kit (QIAGEN). Insertions were identified by PCR using Q5 High-Fidelity DNA polymerase (NEB). Insertion products containing the 6N sequence were amplified and sequenced using a MiSeq Reagent Kit v2, 300-cycle (Illumina). PAM and insert position were characterized as previously described.<sup>8</sup> The frequency of insertions was determined with droplet digital PCR (ddPCR) as described below with 1 pg to 10 ng template plasmid DNA for 20 µL ddPCR reaction. AjTn6022 and ZooTsy were analyzed in the same manner, but a constitutive lac promoter was used for expression of proteins. In brief, co-electroporated transformants were incubated with triple antibiotic (50 µg/ml kanamycin, 50 µg/ml spectinomycin, and 50 µg/ml chloramphenicol) containing LB-agar plates for 17 hours at 37°C, and all colonies were scraped for downstream analysis. Experiments were performed with three biological replicates. All data points are shown with an error bar showing standard deviation, and statistical significance was assessed by t-test.

### Nanopore long-read sequencing to determine plasmid structure

To isolate plinsert and determine its structure, the pR6K-donor was utilized. For AjTn6022, pHelper(AjTn6022), pR6K-donor (AjTn6022), and pTarget(AjTn6022) were co-electroporated into Pir1+ *E. coli*. Cells were recovered for 1 hour and plated on 50 µg/ml kanamycin, 50 µg/ml spectinomycin, and 50 µg/ml chloramphenicol containing LB-agar plates. Cells were harvested 5 days after incubation at room temperature and subjected to mini-prep by QIAprep Spin Miniprep Kit (QIAGEN). Note that we avoided incubation at 37°C to prevent transposed products from being actively resolved to enable detection of replicative transposition. To recover transposed plinsert products, 100 ng of mini-prep product was electroporated into Endura Competent Cells (Lucigen). Cells were recovered for 1 hour and plated on 50 µg/ml kanamycin and 50 µg/ml chloramphenicol containing LB-agar plates,



and further incubated at room temperature for 5 days. Donor insertion on pTarget was confirmed by Sanger sequencing of mini-prep products for 4 colonies. In parallel, all colonies were harvested and subjected to mini-prep, followed by amplification-free nanopore sequencing library preparation (Oxford Nanopore Technologies SQK-LSK109). Briefly, mini-prep products were linearized by BstZ171, followed by end-prep and subsequent ligation of sequencing adapters. Resulting libraries were loaded on a MinION R9 flow cell and sequenced. Sequence reads containing 25 bp of *comM* gene fragment, 25 bp of LE, and 25 bp of RE (up to 2 bp mismatches were allowed for each component) were filtered for further analysis, thereby discarding low-quality and contaminating gDNA reads. ZooTsy was analyzed in the same manner with cognate plasmids. For quality control, reads over 2.5 kb with QScore >7 were filtered. To determine the structure of plinsert, the reads were mapped on the expected simple insertion product (Figure S5D). To count simple insertions, nanopore sequencing reads were analyzed using the following pipeline (Figures S5C and S5E): the reads which contain both (1) the downstream edge 30-bp sequence of cloned *comM* 100-bp fragment (AGCGGGCCGGGAAC TCGGCGCGGCGGGCTG) and (2) 30-bp end2 sequence (GGACTGGGATTCTCCAATATTCCTTAGCGC) were further extracted by the insertion reads counter. Among those reads, the ones which have (3) the upstream edge 30-bp sequence of the cloned *comM* fragment (CACGGCTTCGACCGCAGCACTGGTCGGTGG) and (4) the 30-bp sequence of the cargo (GTCGGGGG ATCCACTAGTGAGCTCATGCA) were extracted by the counter and used for mapping on the expected simple insertion. Based on the distance between the upstream edge of *comM* and cargo, each selected read was counted as a simple insertion.

### Tsy circular intermediate isolation assays

100 ng of pHelper(ZooTsy) was co-electroporated with 100 ng of lacZa backbone pDonor(ZooTsy) and 100 ng of pACYC backbone vector into BL21(DE3) electrocompetent cells (Sigma) and plated on 50 µg/ml kanamycin, 50 µg/ml spectinomycin, and 50 µg/ml chloramphenicol containing LB-agar plates. After incubation for 17 hours at 37°C, all colonies were scraped from the plates, and the plasmid DNA was purified using QIAprep Spin Miniprep Kit (QIAGEN). 100 pg of mini-prep product was used for re-transformation of NEB 10-beta Competent *E. coli* (NEB), and the cells were plated on S-Gal (Sigma) and 50 µg/ml kanamycin containing LB-agar plates for blue/white selection. After incubation for 17 hours at 37°C, white colonies were picked up and further cultured for 17 hours at 37°C, and the plasmid DNA was purified using QIAprep Spin Miniprep Kit (QIAGEN). The original lacZa backbone pDonor(ZooTsy) and the circularized intermediate (CI) were digested with NruI (NEB) for an hour at 37°C and loaded on E-Gel Ex Agarose Gels 1%. The junction of end1 and end2 was confirmed by Sanger sequencing of mini-prep products for 4 white colonies. In parallel, isolated products from 24 white colonies were combined and subjected to amplification-free nanopore sequencing library preparation (Oxford Nanopore Technologies SQK-LSK109). Briefly, mini-prep products were linearized by NruI, followed by end-prep and subsequent ligation of sequencing adapters. Resulting libraries were loaded on a MinION R9 flow cell and sequenced.

### TTISS for insertion specificity analysis

Endura Competent Cells were transformed for transposition assay as mentioned above, with the following modifications: 100 ng of a temperature-sensitive pSC101-donor was used. After incubation on triple antibiotic LB-agar plates, cells were re-plated and grown for 12 hours at 43°C to prevent unintended amplification of donor plasmids in the following Tagmentation-based Tag Integration Site Sequencing (TTISS) analysis.<sup>36</sup> Genomic DNA was extracted from hundreds of colonies on a LB-agar petri dish, and 500 ng genomic DNA (from approximately 10<sup>8</sup> *E. coli* cells) was tagmented with Tn5, and re-purified by Wizard SV Gel and PCR Clean-Up System (Promega). Tagmented DNA samples were amplified using two rounds (in total 37 cycles) of PCR with KOD Hot Start DNA Polymerase (Millipore) using a Tn5 adapter-specific primer and nested primers within the DNA donor (Data S3). PCR products from 4 different experimental conditions were pooled together, purified, and the resulting libraries were sequenced using a NextSeq v2 kit (Illumina), 75 cycle kit with 45 forward cycles and 30 reverse cycles. Read pairs with R1 containing the terminal 29 bp of the AjTn6022 transposon RE sequence were filtered for further analysis and trimmed of the transposon sequence for alignment to the *E. coli* genome (CP011113.2) and pTarget(AjTn6022). Filtered and trimmed reads were aligned using the established BWA aligner pipeline.<sup>69</sup> The resulting SAM files were exported for further analysis. Aligned R1 reads with length of 16 bp (remaining R1 length after trimming of 29 bp of transposon RE sequence) and SAM flags 99 and 147 (for mapped reads within the insert size and in correct orientation) were used to determine the correct transposon insertion reads. Reads with a single insertion position at the *AjcomM* on pTarget and *E. coli* endogenous *comM* insertion sites were considered on-target, while remaining reads were counted as off-targets. For ZooTsy, the resulting libraries were sequenced using a MiSeq v2 kit with 75 forward cycles and 75 reverse cycles. Read pairs with R1 containing the terminal 36 bp of the ZooTsy end1 sequence were filtered for further analysis and trimmed of the transposon sequence for alignment to the *E. coli* genome and pTarget(ZooTsy). Filtered and trimmed reads were aligned, and the resulting SAM files were exported for further analysis. Aligned R1 reads with length of 39 bp (remaining R1 length after trimming of 36 bp of transposon end1 sequence) and SAM flags 99 and 147 were used to determine the correct transposon insertion reads.

### Droplet digital PCR reactions

Insertion events were quantified using insertion specific primers and a donor specific probe (Data S3). ddPCR Supermix for Probes (No dUTP) (BioRad), primers (900 nM each), a probe (250 nM), and template DNA were combined into 20 µL reactions, and droplets were generated with 70 µL of Droplet Generation Oil for Probes (BioRad) using the QX200 Droplet Generator (BioRad). Thermal cycling conditions for ddPCR reactions were as follows: 1 cycle, 95°C, 10 min; 40 cycles, 94°C, 30 s, 58°C, 1 min; 1 cycle, 98°C, 10 min; 4°C hold. PCR products were read with a QX200 Droplet Reader, and the absolute concentrations of inserts were

determined using QuantaSoft (v1.6.6.0320). Total template (genome or target plasmid) amount was also quantified through this process, and insertion frequency was calculated as inserts/template.

### Purification of TniQ and TnsF proteins

Each protein expression vector was transformed into BL21(DE3) Competent Cells (NEB). 4 mL of starter culture was grown in TB supplemented with 100  $\mu$ g/ml ampicillin for 12 h, which was used to inoculate 2 L of TB for growth at 37°C and 150 rpm until an OD<sub>600</sub> of 0.6. Protein expression was induced by supplementation with IPTG to a final concentration of 0.5 mM. The cells were incubated at 16°C for 16 h for protein expression, and then harvested by centrifugation for 20 min at 4°C at 4000 rpm (Beckman Coulter Avanti J-E, rotor JLA9.100). All subsequent steps were performed at 4°C. Cell pellet was resuspended in 200 mL of lysis buffer (50 mM Tris-HCl, 500 mM NaCl, 5% glycerol, 1 mM DTT, pH 8.0) supplemented with cOmplete Protease Inhibitor Cocktail (Millipore sigma 4693116001). Cells were disrupted by the LM20 Microfluidizer system at 28,000 PSI. Lysate was cleared by centrifugation for 30 min at 4°C at 9000 rpm (Beckman Coulter Avanti J-E, rotor JLA-10.500). The cleared lysate was applied to 1 mL of packed Strep-Tactin Sepharose resin (IBA) and incubated with rotation for 1 h, followed by washing of the protein-bound beads in 50 mL of lysis buffer. The proteins were cleaved off by Ulp1 SUMO protease OR SENP at 4°C for 16 h. Resulting proteins were concentrated by an Amicon Ultra Centrifugal Filter Units (Millipore), and protein concentration was estimated by NuPAGE (Invitrogen) and eStain L1 Protein Staining System (GenScript). The concentrated protein was loaded onto a gel filtration column (Superdex 200 Increase 10/300 GL, GE Healthcare) equilibrated with storage buffer (50 mM Tris-HCl, 250 mM NaCl, 5% glycerol, 1 mM DTT, pH 8.0) via FPLC. The resulting fractions from gel filtration were analyzed and the fractions containing the protein were pooled and stored at -80°C. To purify TwinStrep-bdSUMO-TnsF from AjTn6022, the bound bdSUMO-tagged proteins were eluted in 10 mL of lysis buffer supplemented with 2.5 mM desthiobiotin (Sigma) instead of digestion by SENP.

### Pull-down experiments for detecting AjTn6022 TniQ-TnsF interactions

TwinStrep-bdSUMO-TnsF was mixed with each purified TniQ protein in the assembly buffer (25 mM Tris-HCl, 250 mM NaCl, 1 mM DTT, pH 8.0) and incubated at 37°C for 1 hour. The reaction was mixed with pre-washed Strep-Tactin Magnetic Microbeads (IBA) and further incubated at 4°C for 30 min. The beads were washed three times with the wash buffer (25 mM Tris-HCl, 250 mM NaCl, 1 mM DTT, 0.05% Tween20, pH 8.0). Then, the protein complexes on the beads were eluted and denatured in NuPAGE LDS sample buffer at 95°C for 10 min. Samples were loaded on NuPAGE 4 to 12%, Bis-Tris Gel, separated by electrophoresis, and stained with Imperial Protein Stain (ThermoFisher Scientific).

### Electrophoretic mobility shift assay for detecting TnsF-DNA interactions

Purified PCR amplicons and oligos (HPLC purified, IDT) were used as DNA probes for EMSA assays. For DNA-protein complex assembly, 20 ng DNA probe was mixed with purified AjTn6022 TnsF protein or ZooTsy nuclease-dead TnsF-Y584F mutant protein at different DNA:protein molecular ratio in assembly buffer (25 mM Tris-HCl, 50 mM NaCl, 1% glycerol, 5 mM MgCl<sub>2</sub>, pH 8.0), and incubated at 37°C for 1 hour. Samples were mixed with Novex Hi-Density TBE Sample Buffer (ThermoFischer Scientific) and loaded on pre-equilibrated Novex 6% TBE Gel (ThermoFischer Scientific) in 0.5X TBE buffer for electrophoresis. Gel images were captured by SYBR Gold-staining.

## QUANTIFICATION AND STATISTICAL ANALYSIS

### Quantification of ddPCR data

Frequency of transposition event was determined by ddPCR using QuantaSoft (v1.6.6.0320). Copy number of generated transposon and target junction was measured by transposon end specific FAM probes ([Data S3](#)). Copy number of the target site (on genome or target plasmid) was also quantified through this process using specific FAM probes at independent ddPCR reactions and insertion frequency was calculated as inserts/template \* 100 and displayed in %. All data points are shown with an error bar showing standard deviation, and statistical significance was assessed by two-tailed t-test.