Recruitment of CRISPR-Cas systems by Tn7-like transposons

Joseph E. Peters[1#], Kira S. Makarova[2], Sergey Shmakov[3], Eugene V. Koonin[2#]

[1] Department of Microbiology, Cornell University, Ithaca, NY, 14853, USA

[2] National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD

[3] Skolkovo Institute of Science and Technology, Skolkovo, 143025, Russia

Address correspondence to koonin@ncbi.nlm.nih.gov or joe.peters@cornell.edu

Running title = Tn7 and CRISPR-Cas

Abstract

A survey of bacterial and archaeal genomes shows that many Tn7-like transposons contain 'minimal' type I-F CRISPR-Cas systems that consist of fused cas8f and cas5f, cas7f and cas6f genes, and a short CRISPR array. Additionally, several small groups of Tn7-like transposons encompass similarly truncated type I-B CRISPR-Cas systems. This gene composition of the transposon-associated CRISPR-Cas systems implies that they are competent for pre-crRNA processing yielding mature crRNAs and target binding but not target cleavage that is required for interference. Here we present phylogenetic analysis demonstrating that evolution of the CRISPR-Cas containing transposons included a single, ancestral capture of a type I-F locus and two independent instances of type I-B loci capture. We further show that the transposon-associated CRISPR arrays contain spacers homologous to plasmid and temperate phage sequences, and in some cases, chromosomal sequences adjacent to the transposon. A hypothesis is proposed that the transposon-encoded CRISPR-Cas systems generate displacement (R-loops) in the cognate DNA sites, targeting the transposon to these sites and thus facilitating their spread via plasmids and phages. This scenario fits the "guns for hire" concept whereby mobile genetic elements can capture host defense systems and repurpose them for different stages in the life cycle of the element.

Importance

CRISPR-Cas is an adaptive immunity system that protects bacteria and archaea from mobile genetic elements. We present comparative genomic and phylogenetic analysis of degenerate CRISPR-Cas variants associated with distinct families of transposable elements and develop the hypothesis that such repurposed defense systems contribute to the transposable element propagation by facilitating transposition into specific sites. Such recruitment of defense systems by mobile elements supports the "guns for hire" concept under which the same enzymatic machineries can be alternately employed for transposon proliferation or host defense.

Key words: CRISPR-Cas system, Tn7 transposon, transposition strategy, crRNA guide, target-site selection

Introduction

CRISPR (Clustered Regularly Interspaced Short Palindromic Repeat)-Cas (CRISPR-ASsociated proteins) systems are the only adaptive immune systems identified in prokaryotes (1). CRISPR-Cas systems possess modular organization which roughly reflects the three main functional stages of the CRISPR immune response: i) spacer acquisition (known as adaptation), ii) pre-crRNA processing and iii) interference (1). CRISPR-Cas systems are highly diverse but can be partitioned into two distinct classes based on the organization of the effector module that is responsible for processing and adaptation (2). Class 1 CRISPR-Cas systems are further divided into 3 types and 12 subtypes in all of which the effector modules are multisubunit complexes of Cas proteins (2). In contrast, in the currently identified 3 types and 12 subtypes of Class 2, the effector modules are represented by a single multidomain protein, such as the thoroughly characterized Cas9 (3-5).

At the adaptation stage, the Cas1-Cas2 protein complex, in some instances with additional involvement of effector module proteins, captures a segment of the target DNA (known as the protospacer) and inserts it at the 5'end of a CRISPR array (6-9). In the second, processing stage, a CRISPR array is transcribed into a long transcript known as pre-CRISPR (cr) RNA that is bound by Cas proteins and processed into mature, small crRNAs. In most Class 1 systems, the pre-crRNA processing is catalyzed by the Cas6 protein that, in some cases, is loosely associated with the effector complex (1, 10). The final, interference step involves binding of the mature crRNA by the multisubunit effector complex, scanning a DNA or RNA molecule for a sequence matching the crRNA guide and containing a protospacer adjacent motif (PAM), and cleavage of the target by a dedicated nuclease domain(s) (1, 10-12). The identity of this nuclease(s) differs between type I and type III CRISPR-Cas systems. In type I, the protein

responsible for target cleavage is Cas3 which typically consists of a Superfamily II helicase and HD-family nuclease domains (13). After the effector complex, which is denoted Cascade (CRISPR-associated complex for antiviral defense (14)) in type I systems, recognizes the cognate protospacer in the target DNA, it recruits Cas3, after which the helicase unwinds the target DNA duplex, and the HD nuclease cleaves both strands (15, 16). Type III systems lack Cas3 , and the protein responsible for target cleavage is Cas10 which contains a polymerase-cyclase and HD-nuclease domains that are both required for the target degradation (17, 18).

In some of the CRISPR-Cas systems, the adaptation genes are encoded separately or even are missing from the genome containing effector complex genes. Among these non-autonomous CRISPR-Cas systems, those of type III systems have been characterized in most detail (1). It has been shown that class III effector complexes can utilize crRNA originating from CRISPR arrays associated with type I systems and thus do not depend on their own adaptation modules (19-23). Furthermore, the CRISPR-Cas systems of type IV, that are often encoded on plasmids, typically consist of the effector genes only (2). No adaptation genes and no associated nuclease domains could be found in the type IV loci although, occasionally, CRISPR arrays and *cas6*-like genes are present. The type IV systems have not yet been studied experimentally, so their mode of action remains unknown. Finally, several variants of type I systems, similarly to type IV, lack adaptation genes and genes for proteins involved in DNA cleavage. A “minimal” variant of subtype I-F has been identified in the bacterium *Shewanella putrefaciens*, with an effector module that consists only of Cas5f, Cas6f and Cas7f proteins, and lacks the large and small subunit present in other Cascade complexes (24). Even more dramatic minimization of subtype I-F has been observed in another variant of subtype I-F that lacks the adaptation module and consists solely of three effector genes , namely a fusion of *cas8f* (large subunit) with *cas5f* that is unique for this variant, *cas7f* and *cas6f* (Figure 1A) (2). Given the composition of their Cascade complex, these Cas1-

less minimal subtype I-F systems can be predicted to process pre-crRNA, yielding mature crRNAs, and recognize the target. However, they lack the Cas3 protein and therefore cannot be expected to be competent for target cleavage. Here we report an exhaustive *in silico* analysis of this system showing that it is strongly linked to a specialized group of transposons related to the well-studied Tn7.

The canonical Tn7 is notable because of the level of control it exerts over the target site selection (25). Three transposon-encoded proteins form the core transposition system including a heteromeric transposase (TnsA and TnsB) and a regulator protein (TnsC) (Figure 1B). In addition to the core TnsABC transposition proteins, Tn7 elements encode dedicated target site selection proteins (TnsD and TnsE) that only allow transposition when specific types of target sites are available. In conjunction with TnsABC, the sequence-specific DNA-binding protein TnsD directs transposition into a conserved site referred to as the Tn7 attachment site, *attTn7* (26). Transposition mediated by TnsABC + TnsE is preferentially directed into mobile plasmids and bacteriophages owing to the ability of TnsE to recognize complexes formed during specific types of DNA replication (27-30). Mobile elements related to Tn7 and encoding proteins homologous to TnsA, TnsB, and TnsC have been described that appear to use distinct attachment sites recognized by TnsD/TniQ-like proteins, but do not encode a TnsE-like protein (31).

Our analysis shows that the Cas1-less I-F systems associate with a distinct group of Tn7-like elements. These transposons encode TnsD/TniQ-like proteins and utilize novel attachment sites but lack TnsE-like proteins that normally promote horizontal transfer of the elements. Several identified matches for the spacers from the transposon-associated CRISPR arrays suggest that this system might function by targeting transposition to target sites enabled by guide

crRNAs. We hypothesize that the subtype I-F CRISPR-Cas machinery recruited by these elements facilitates their horizontal dissemination, mostly via plasmids and/or phages.

Results and discussion

A variant of the type I-F CRISPR-Cas system is specifically associated with a distinct family of Tn7-like elements. For the purpose of comprehensive identification type I-F CRISPR-Cas loci, we chose the Cas7f protein as the probe given that it is the most conserved component in all systems of this subtype including the "minimal" variant lacking *cas1* and *cas2-cas3* genes. Using a PSI-BLAST search started with Cas7f profiles, we obtained 2905 Cas7f protein sequences, mapped them onto the respective genomes and annotated the genes in the 10 kb up- and downstream neighborhoods of the *cas7f* genes using PSI-BLAST against the conserved domain database (CDD). These 20 kb loci are long enough to cover a typical complete I-F system that consists of 6 genes (2). We then reconstructed a phylogenetic tree from all identified Cas7f protein sequences (Figure 2A, Supplemental Table S1, see respective Newick tree at ftp://ftp.ncbi.nih.gov/pub/makarova/supplement/Peters_at_al_2017/). Mapping gene neighborhoods on the tree revealed a single, monophyletic, strongly supported branch including all *cas1*-less I-F variants. As of this analysis, the branch encompassed 423 sequences from 19 genera of Gammaproteobacteria and appears to derive from a typical, complete I-F system (Figures 1A and 2A). Indeed, all other branches in the tree consist of Cas7f homologs from complete I-F systems containing a *cas1* gene within the locus. A few exceptions that are scattered in the tree are from either small contigs or disrupted *cas* loci. In the vast majority of the loci corresponding to the *cas1*-less branch, a *tniQ/tnsD* gene is located next to the *cas* genes (Figure 3).

To determine whether the association of the Cas1-less I-F systems with Tn7-like elements was unique or emerged independently on several occasions, we analyzed the TniQ/TnsD and TnsA families. The TnsA protein is the most highly conserved gene of the Tn7-like elements and is responsible for the unique behavior of the elements with heteromeric transposases (31-34). We collected and annotated 10,349 loci containing at least *tniQ/tnsD* or *tnsA* (Supplemental Table S2) and reconstructed a tree for both protein families (Figure 2 B and C and see respective Newick trees at ftp://ftp.ncbi.nih.gov/pub/makarova/supplement/Peters_at_al_2017/). In both trees, the loci containing *cas* genes of the *cas1*-less I-F variant mapped to strongly supported monophyletic branches (Figure 2 B and C). Thus, phylogenetic analysis of both Cas7f and the associated transposon-encoded proteins reveals a strong link between a specific group of Tn7-like elements and a distinct variant of the subtype I-F CRISPR-Cas system. The Tn7-like elements in the clade that includes Tn6022 were identified as the outgroups to the respective branches in both the TnsA and TniQ/TnsD trees, suggesting that a member of the Tn6022 family is the ancestor of the CRISPR-associated variety of Tn7-like transposons (Figure 2B and C). Both clades include multiple, deep branches that are not associated with *cas* genes in the respective loci suggesting that the link with I-F system evolved relatively late in the history of this group of Tn7-like elements (see respective Newick trees at ftp://ftp.ncbi.nih.gov/pub/makarova/supplement/Peters_at_al_2017/). In several cases, however, the *cas* genes seem to have been lost from the vicinity of the conserved transposon genes (eg. *Shewanella baltica* OS678 and *Thiomicrospira crunogena* XCL_2), suggesting that the CRISPR-Cas system is not essential for the transposon survival. However, there are no intact *cas1*-less I-F

systems outside this transposon neighborhood, with the implication that this CRISPR-Cas variant is functional only when associated with a Tn7-like element.

We further investigated the *tniQ/tnsD/tnsA* loci in order to identify any other CRISPR-Cas systems that might be linked to Tn7-like transposons. Only a few such instances were identified, mostly complete loci containing the adaptation genes. The respective *tnsA* and/or *tniQ/tnsD* genes are scattered in the phylogenetic trees suggesting that most of these associations are effectively random and might be transient (Supplemental Table S2). However, some of such loci do show a degree of evolutionary coherence. Specifically, they form two small, unrelated branches in both the TnsA and the TniQ/TnsD trees (See I-B in Figure 2B-C). All these CRISPR-*cas* loci are present in different cyanobacteria, belong to the I-B subtype and lack adaptation genes as well as the *cas3* gene that is required for DNA cleavage in type I systems. Thus, to a large extent, these type I-B variants mimic the organization of the more common transposon-associated, *cas1*-less I-F variant (See below).

The *cas1*-less type I-F CRISPR-Cas system is mobilized together with conserved transposition genes. We analyzed the transposon end-sequences in the loci containing the I-F and I-B CRISPR-Cas variants in order to determine whether the *cas* genes were located within the boundaries of these elements or are simply adjacent to the transposon. The structure of the left and right ends of canonical Tn7 has been defined previously (Supplemental Figure S1). Tn7 ends are marked by a series of 22 bp TnsB-binding sites (35-37). Flanking the most distal TnsB-binding sites is an 8 base pair terminal sequence ending with 5'-TGT-3'/3'-ACA-5'. Tn7 contains 4 overlapping TnsB-binding sites in the ~90 bp right end of the element and three dispersed sites in the ~150 bp left end of the element, but the number and distribution of TnsB-binding sites can vary among Tn7-like elements (25, 31). End-sequences of Tn7-related elements can be

183 determined by identifying the directly-repeated 5 base pair target site duplication, the terminal 8
184 base-pair sequence, and 22 base pair TnsB-binding sites (the latter two found in an inverted
185 configuration in the left and right ends of the element) (Supplemental Figure S1). Compared with
186 the canonical Tn7 and Tn6022, Tn7-like elements show extensive variation in size and gene
187 complements as illustrated by a representative set of 12 complete elements ranging in size from
188 22 kb to almost 120 kb (38, 39)(Figure 3 and Table 1). One of these elements has been
189 previously identified in *Vibrio parahaemolyticus* RIMD2210633 as a member of the Tn7
190 superfamily and encodes the *Vibrio* pathogenicity determinant, the thermostable direct
191 hemolysin (TDH) (40).

192 In our analysis of CRISPR-Cas systems, two groups of type I-B variants were identified
193 in association with Tn7-like elements (Figure 2B-C). Similar to the type I-F CRISPR-Cas
194 variant, these I-B systems are expected to be functional for maturing CRISPR transcripts and
195 forming crRNA complexes at protospacers but lack adaptation genes and Cas3, and accordingly,
196 are likely to be defective for interference. Furthermore, these type I-B CRISPR-Cas variants are
197 associated with small CRISPR arrays (Figure 3). One group of the type I-B associated
198 transposons encodes a TnsD/TniQ protein that belongs to the same clade as TnsD from canonical
199 Tn7 and resides in the *attTn7* site downstream of *glmS*. An example from this subgroup has been
200 previously identified in *Anabaena variabilis* (# 5291 in Figure 3), but the minimal Cas system
201 contained in the element was not analyzed (41). The second group encodes a TniQ protein that
202 belongs to a new family of elements encoding fused TnsA-TnsB proteins (#2757 in Figure 3).

203 Taken together, these findings indicate that the type I-F and I-B CRISPR-Cas variants
204 identified in this work are part of the core gene repertoire in multiple clades of Tn7-like
205 elements.

Chromosomal insertions of the I-F CRISPR-Cas-associated elements show three recognizable attachment sites likely accessed by dedicated TniQ/TnsD proteins. The canonical Tn7 element has been studied extensively, especially the transposition pathway that directs the element into the *attTn7* site located downstream of the conserved *glmS* gene. The Tn7 TnsD(TniQ) protein is a sequence-specific DNA-binding protein that recognizes a highly conserved 36 bp sequence in the downstream region of the *glmS* gene coding sequence (26, 42). Transposition events promoted by TnsABC+D are directed into a position 23 bp downstream of the region bound by TnsD. Tn7 transposition is orientation-specific in all transposition pathways; the transposon end proximal to the *tnsA* gene (the "right" end of the element) is adjacent to the DNA sequence or a specific protein complex recognized in each pathway (29, 42-44).

We analyzed the region adjacent to the point of insertion of the elements and identified three new attachment sites for the *cas1*-less, type I-F-associated transposons. Similar to Tn7 insertions, one subgroup of the elements occurred downstream of a gene, but instead of *glmS*, these insertions were found downstream of an inosine-5′-monophosphate dehydrogenase gene (Table 1, Figure 3 and 4). The configurations found with the other recognizable attachment sites were new for Tn7-like elements. In one case, the attachment site was located upstream of the *yciA* gene, which encodes an acyl-CoA thioester hydrolase (Table 1, Figure 3 and 4). Presumably, insertion of the element into this attachment site would ablate the normal promoter, but changes in expression remain to be demonstrated experimentally. The third attachment site identified for the *cas1*-less type I-F-associated elements is the first example where a non-protein-encoding gene was recognized, namely, the gene for the signal recognition particle RNA (SNP-RNA) (Table 1, Figure 3 and 4). The concordance between the phylogeny of the TniQ/TnsD proteins and the attachment site used by the element is consistent with the hypothesis that each

attachment site is recognized by a cognate TniQ/TnsD protein (Figure 4). Despite this concordance, many transposons appear to be inserted in random sites (Figure 4). It remains unclear how insertions were directed into these sites because they are unlikely to be specifically recognized by TniQ/TnsD proteins encoded by these elements, and these elements lack a homolog of the TnsE protein found in typical Tn7 transposons.

Analysis of CRISPR arrays associated with the *cas1*-less I-F systems. The great majority of the transposon-associated I-F and I-B systems encompass a CRISPR array downstream of the *cas6* gene (See Figure 3 for examples). In most cases, this array contains only one or two spacers, suggesting that spacer acquisition in these arrays occurs only rarely (Table 2 and Figure 3). Nevertheless, the spacers are typically unrelated even in closely related bacterial genomes indicating that, occasionally, new spacers are incorporated, and old ones are lost. Obviously, only adaptation genes acting *in trans* can insert new spacers into these arrays. Among the 14 complete bacterial genomes containing elements with the I-F CRISPR-Cas, only two encompass other CRISPR-Cas loci containing adaptation genes, namely, *Vibrio fluvialis* ATCC 33809 and *Pseudoalteromonas rubra* SCSIO6842 that possess I-F and I-C systems, respectively. Among draft genomes, there are more cases where additional, complete CRISPR-Cas systems, mostly I-F and I-E, are present in the same genomes. Nevertheless, most of the genomes that contain the I-F variant associated with Tn7-like transposons lack other CRISPR-Cas system that would be able to provide for adaptation, which might account for the short CRISPR arrays. All 4 complete genomes containing elements associated with I-B systems encompass additional CRISPR-Cas loci containing adaptation genes, often of subtype I-D, which is abundant in cyanobacteria (2).

Altogether, more than 800 spacers were identified in the transposon-associated I-F and I-B CRISPR arrays (see automatically and manually identified spacers at ftp://ftp.ncbi.nih.gov/pub/makarova/supplement/Peters_at_al_2017/). As in most analyses of CRISPR spacers (45-47), only a small fraction of these spacers yielded significant matches to sequences in public databases. However, the matches that could be detected were informative because they were to plasmids and bacteriophages associated with the same bacterial genera where the respective elements are found (Table 2). We identified two cases (in *Photobacterium kishitanii* and *Photobacterium leiognathi*) of potential special interest, where spacers matched the region adjacent to the *tnsA*-gene proximal side of the element (Table 2), i.e. the specific region where complexes involved in targeting transposition events interact with the target DNA (29, 48, 49). An additional spacer match was found inside the transposon boundaries in several *Vibrio parahaemolyticus* strains (Table 2). A similar situation might have also occurred in a Tn7-like transposon associated with a type I-B CRISPR-Cas variant in a *Cyanothece* PCC 7822 plasmid although end sequences could not be unambiguously defined for this element (Table 2).

A role for CRISPR-Cas in targeting transposition? Taking into account all the observations on the transposon-associated CRISPR-Cas systems and previous studies on the mechanism of target site activation, we propose a model for the involvement of Cas1-less CRISPR-Cas systems in targeting transposition to facilitate cell-cell transfer of the element (Figure 5). Canonical Tn7 encodes two targeting pathways that are both mediated by the same set of TnsABC proteins (Figure 1B). The TnsABC+TnsD(TniQ) pathway appears to be broadly conserved allowing high frequency transposition into an attachment site recognized by a cognate TnsD/TniQ protein (Figure 1 and 4, Table 1)(31). The *cas1*-less I-F CRISPR-Cas variant is encoded in the same location where the *tnsE* gene that promotes transposition into conjugal

plasmids and filamentous bacteriophages is typically located (Figure 3). Thus, it appears plausible that the CRISPR-Cas system functionally replaces *tnsE* as a mechanism facilitating horizontal transfer of the element. Support for this possibility comes from the observation that the transposon-associated CRISPR arrays largely carry plasmid and phage-specific spacers and could direct the transposon to the respective elements (Table 2).

Distortions in B form DNA induced by Cas-crRNA could play a role in recruiting transposition. Transposition into *attTn7* is well-understood at the molecular level; the DNA structure in the vicinity of the attachment site plays a central role in transposition (Figure 6 A, B and C). TnsD-binding induces an asymmetric distortion in the *attTn7* target DNA that is primarily responsible for attracting TnsC during transposition (48, 50)(Figure 6A). Protein-protein interaction between TnsD and TnsC have been detected (26) but multiple lines of evidence suggest a defining role for the distortion in the DNA for the target site selection (see below).

The TnsABC proteins are normally insufficient for Tn7 transposition *in vivo* or *in vitro* (51); however, certain gain-of-activity mutations in the regulator protein TnsC (TnsC*) allow transposition in the absence of TnsD or TnsE (52, 53). Untargeted transposition is observed *in vitro* and *in vivo* with TnsABC* in the absence of target site selection proteins (30, 52), but notably, transposition in this case is attracted to a specific location adjacent to a short segment of triplex-forming DNAs (44, 54). Analogous to transposition events found in *attTn7*, these events are targeted to a position on one side of the triplex-forming DNA in a unique orientation owing to the ability of TnsC to recognize the distortion formed at the triplex-to-B-form DNA transition (Figure 6C). Distortions induced in the target DNA are also implicated in transposition targeting by TnsABC+E (55) (Figure 6B). Given that distortions in B form DNA are also expected

297 adjacent to crRNA-bound effector complexes that generate R-loops through duplex formation
298 between the crRNA and the protospacer (12, 56), there could be a mechanistic link between the
299 well-understood Tn7 targeting process and DNA targeting by the CRISPR-Cas effector
300 complexes (Figure 6D). This analogy is consistent with the left to right orientation of the two
301 insertions located adjacent to spacer matches (Table 2).

302 **Evolution of the association between CRISPR-Cas variants and Tn7 like elements.**
303 Given that at least some type I CRISPR-Cas systems selectively integrate spacers from plasmids
304 and phages (6, 57), an attractive hypothesis is that the CRISPR-*cas* loci that randomly became
305 associated with the transposon were fixed through selection for their ability to facilitate
306 dissemination of transposons. As discussed above, because changes in DNA structure play an
307 important role in target site selection by Tn7, relatively little evolutionary adaptation might be
308 needed to allow the core TnsABC machinery to recognize crRNA-bound effector complexes for
309 targeting. In this light, it is intriguing that association between CRISPR-Cas systems and Tn7-
310 like elements occurred on multiple, independent occasions. The consistent minimalist features in
311 the organization of the transposon-associated type I-F and I-B variants imply that they co-
312 evolved with Tn7-like elements along parallel paths of reductive evolution. Both type I-F and
313 type I-B systems have lost the adaptation module (*cas1* and *cas2*) and the *cas3* gene, which is
314 required to cleave the target DNA in other type I systems (1). The absence of Cas3 implies that
315 these CRISPR-Cas systems recognize but do not cleave the target, a mode of action that would
316 allow the targeted DNA to serve as a vehicle for horizontal transfer of the respective Tn7-like
317 transposon.

318 The transposon-associated CRISPR arrays are short, and the respective bacterial genomes
319 often lack CRISPR adaptation modules. Thus, the majority of the CRISPR-containing

transposons are likely to be relatively recent arrivals to the respective genomes, conceivably, brought about by the plasmid or phage against which they carry a spacer. Once integrated into a new host attachment site, such transposons could "lie in waiting" for a horizontal transfer vehicle, either as a result of *in trans* acquisition of a new spacer that is specific to an endogenous plasmid or prophage or via the entry of an element that is already represented by a cognate spacer in the transposon-encoded CRISPR array. In some cases, an incoming plasmid or phage recognized by the CRISPR-Cas system and targeted for transposition would be incapacitated by the integration event. Nevertheless, even such unproductive integrations would still benefit the CRISPR-carrying transposon by protecting the host. In such cases, CRISPR-directed integration that is in keeping with a selfish behavior for the transposon would also qualify as altruistic behavior toward the host. It is unclear what advantage there could be to losing the capacity to undergo adaptation (i.e. lack of *cas1* and *cas2*). However, because Tn7 itself is a mobile element, sporadic access to adaptation systems *in trans* may prove safer for the element, limiting the acquisition of transposon-directed spacers in the element-encoded array. Occasionally, these elements appear to acquire spacers from the host chromosome, conceivably stimulating ectopic transposition within the same genome. Such a system could be beneficial in allowing transposition in hosts that lack attachment sites recognized by the element-encoded TniQ/TnsD protein.

Many questions remain regarding the functioning of the CRISPR-Cas in Tn7-like transposons including the possibility of direct interaction between the CRISPR effector complexes and either TnsD/TniQ, TnsABC, or other transposon-encoded accessory proteins. It is also unclear if these CRISPR-Cas variants might perform alternative or additional functions, beyond facilitation of transposition, such as gene silencing or protection of the transposon.

From the evolutionary standpoint, the transposon-associated CRISPR-Cas systems fit the "guns for hire" paradigm (58). Under this concept, genes of mobile genetic elements (MGE) genes are often recruited by host defense systems, and conversely, defense systems or components thereof can be captured by MGE and repurposed for counter-defense or other roles in the life cycle of the element. Recruitment of MGE apparently was central to the evolution of CRISPR-Cas, contributing to the origin of both the adaptation module and the Class 2 effector modules (3, 56, 59). On the other side of the equation, virus-encoded CRISPR-Cas systems have been identified and implicated in inhibition of host defense (60). The observations described here, if validated experimentally, seem to "close the circle" by demonstrating recruitment of CRISPR-Cas systems by transposons, conceivably, for a role in targeting transposition, a key step in transposon propagation. Finally, it has not escaped our notice that the transposon-encoded CRISPR-Cas systems described here potentially could be harnessed for genome engineering applications, namely, precise targeting of synthetic transposons encoding selectable markers and other genes of interest.

Methods

Prokaryotic Genome Database and open reading frame annotation

Archaeal and bacterial complete and draft genome sequences were downloaded from the NCBI FTP site (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/) in March 2016. For incompletely annotated genomes (coding density less than 0.6 CDS per kbp), the existing annotation was discarded and replaced with the Meta-GeneMark 1 (61) annotation using the standard model MetaGeneMark_v1.mod (Heuristic model for genetic code 11 and GC 30). Altogether, the

365 database includes 4,961 completely sequenced and assembled genomes and 43,599 partially
366 sequenced genomes.

367 Profiles for three families protein families, namely Cas7f (cd09737, pfam09615), TnsA
368 (pfam08722, pfam08721) and TnsQ/TnsD (pfam06527), that are available in the NCBI CDD
369 database (62) were used as queries for PSI-BLAST searches (E-value: 10^{-4}, other parameters
370 were default) to find respective homologs. All ORFs within 10 kb regions up- and downstream
371 of *cas7f* genes (to cover potential complete I-F system) and 20 kb regions up- and downstream of
372 *tnsQ/tnsD* and *tnsA* (to cover potential Tn7-like elements) were further annotated using RPS-
373 BLAST searches with 30,953 profiles (COG, pfam, cd) from the NCBI CDD database and 217
374 custom Cas protein profiles (2). The CRISPR-Cas system (sub)type identification for all loci was
375 performed using previously described procedures (2).
376
377 ***Protospacer analysis***
378 The CRISPRfinder (63) and PILER-CR (64) programs were used with default parameters to
379 identify CRISPR arrays in Cas7f and TnsA/TnsD loci. The MEGABLAST program (65) (word
380 size 18, otherwise default parameters) was used to search for protospacers in the virus subset of
381 the NR database and the prokaryotic genome database. Matches were considered only if they
382 showed at least 95% identity and at least 95% length coverage in the case of the NR database,
383 and 80% identity and 80% length coverage for the self-hits (hits were classified as self if they
384 matched the same genomes or genome of the same species disregarding the strain information).
385 Because the automatic approach missed several short CRISPR arrays, loci initially found to lack
386 CRISPR were analyzed manually by examining the intergenic region downstream of the *cas6f*

gene for repeats and using the BLASTN program with the default parameters to find matches to the spacer identified.

Clustering and Phylogenetic Analysis

To construct a non-redundant, representative sequence set, protein sequences within families of interest were clustered using the NCBI BLASTCLUST program. (ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html) with the sequence identity threshold of 90% and length coverage threshold of 0.9. Short fragments or disrupted sequences were discarded. Multiple alignments of protein sequences were constructed using MUSCLE (66) or MAFFT (67) programs. Sites with the gap character fraction values >0.5 and homogeneity <0.1 were removed from the alignment. Phylogenetic analysis was performed using the FastTree program (68), with the WAG evolutionary model and the discrete gamma model with 20 rate categories. The same program was used to compute bootstrap values.

Relationships within diverse sequence families were established using the following procedure: initial sequence clusters were obtained using UCLUST (69) with the sequence similarity threshold of 0.5; sequences were aligned within clusters using MUSCLE (66). Then, cluster-to-cluster similarity scores were obtained using HHsearch (70) (including trivial clusters consisting of a single sequence each), and a UPGMA dendrogram was constructed from the pairwise similarity scores. Highly similar clusters (pairwise score to self-score ratio >0.1) were aligned to each other using HHALIGN (70), and the procedure was repeated iteratively. At the last step, sequence-based trees were reconstructed from the cluster alignments using the FastTree program (68) as described above and rooted by mid-point; these trees were grafted onto the tips of the profile similarity-based UPGMA dendrogram.

410

411 ***Analysis of Tn7-like elements***

412 End-sequences of Tn7-like elements were determined by identifying the directly-repeated five
413 base pair target site duplication, the terminal eight base-pair sequence, and 22 base pair TnsB-
414 binding sites as described in the text using Gene Construction Kit 4.0 to manipulate DNA
415 sequences and search for DNA repeats. Sequence files were derived from matches to *cas7f, tnsA*
416 and *tniQ* as described above.

417

418 **Author contributions.** J.E.P, K.S.M. and S.S. performed genomic analysis. J.E.P, K.S.M and
419 E.V.K. designed the analysis, participated in the data interpretation and discussion, and wrote the
420 paper.

421

422 **Acknowledgements**

423 JEP was supported by the USDA National Institute of Food and Agriculture, Hatch project
424 NYC-189438. KSM, SS and EVK are supported by the intramural program of the U.S.
425 Department of Health and Human Services (to the National Library of Medicine).

426

427

428

429 **Competing interests.** The authors declare no competing financial interests.

430

1. Mohanraju P, *et al.* (2016) Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. *Science* 353(6299):aad5147
2. Makarova KS, *et al.* (2015) An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Microbiol* 13(11):722-736
3. Shmakov S, *et al.* (2017) Diversity and evolution of class 2 CRISPR-Cas systems. *Nat Rev Microbiol* 15(3):169-182
4. Burstein D, *et al.* (2016) New CRISPR-Cas systems from uncultivated microbes. *Nature*
5. Koonin EV, Makarova KS, & Zhang F (2017) Diversity, classification and evolution of CRISPR-Cas systems. *Curr Opin Microbiol* in press
6. Amitai G & Sorek R (2016) CRISPR-Cas adaptation: insights into the mechanism of action. *Nat Rev Microbiol* 14(2):67-76
7. Heler R, *et al.* (2015) Cas9 specifies functional viral targets during CRISPR-Cas adaptation. *Nature* 519(7542):199-202
8. Wei Y, Terns RM, & Terns MP (2015) Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. *Genes Dev* 29(4):356-361
9. Vorontsova D, *et al.* (2015) Foreign DNA acquisition by the I-F CRISPR-Cas system requires all components of the interference machinery. *Nucleic Acids Res* 43(22):10848-10860
10. Charpentier E, Richter H, van der Oost J, & White MF (2015) Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity. *FEMS Microbiol Rev* 39(3):428-441
11. Jackson RN & Wiedenheft B (2015) A Conserved Structural Chassis for Mounting Versatile CRISPR RNA-Guided Immune Responses. *Mol Cell* 58(5):722-728
12. Tsui TK & Li H (2015) Structure Principles of CRISPR-Cas Surveillance and Effector Complexes. *Annu Rev Biophys* 44:229-255
13. Makarova KS, Grishin NV, Shabalina SA, Wolf YI, & Koonin EV (2006) A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol Direct* 1:7
14. Brouns SJ, *et al.* (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* 321(5891):960-964
15. Hochstrasser ML, *et al.* (2014) CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference. *Proc Natl Acad Sci U S A* 111(18):6618-6623
16. Huo Y, *et al.* (2014) Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. *Nat Struct Mol Biol* 21(9):771-777
17. Samai P, *et al.* (2015) Co-transcriptional DNA and RNA Cleavage during Type III CRISPR-Cas Immunity. *Cell* 161(5):1164-1174
18. Zhang J, Graham S, Tello A, Liu H, & White MF (2016) Multiple nucleic acid cleavage modes in divergent type III CRISPR systems. *Nucleic Acids Res* 44(4):1789-1799
19. Majumdar S, *et al.* (2015) Three CRISPR-Cas immune effector complexes coexist in Pyrococcus furiosus. *RNA* 21(6):1147-1158
20. Deng L, Garrett RA, Shah SA, Peng X, & She Q (2013) A novel interference mechanism by a type IIIB CRISPR-Cmr module in Sulfolobus. *Mol Microbiol* 87(5):1088-1099
21. Staals RH, *et al.* (2013) Structure and activity of the RNA-targeting Type III-B CRISPR-Cas complex of Thermus thermophilus. *Mol Cell* 52(1):135-145
22. Staals RH, *et al.* (2014) RNA targeting by the type III-A CRISPR-Cas Csm complex of Thermus thermophilus. *Mol Cell* 56(4):518-530
23. Elmore JR, *et al.* (2016) Bipartite recognition of target RNAs activates DNA cleavage by the Type III-B CRISPR-Cas system. *Genes Dev* 30(4):447-459

24. Gleditzsch D*, et al.* (2016) Modulating the Cascade architecture of a minimal Type I-F CRISPR-Cas system. *Nucleic Acids Res* 44(12):5872-5882
25. Peters JE (2014) Tn7. *Microbiology Spectrum* 2(5):1-20
26. Mitra R, McKenzie GJ, Yi L, Lee CA, & Craig NL (2010) Characterization of the TnsD-attTn7 complex that promotes site-specific insertion of Tn7. *Mobile DNA* 1(1):18
27. Wolkow CA, DeBoy RT, & Craig NL (1996) Conjugating plasmids are preferred targets for Tn7. *Genes & Dev.* 10:2145-2157
28. Finn JA, Parks AR, & Peters JE (2007) Transposon Tn7 Directs Transposition into the Genome of Filamentous Bacteriophage M13 Using the Element-Encoded TnsE Protein. *Journal of Bacteriology* 189(24):9122-9125
29. Parks AR*, et al.* (2009) Transposition into replicating DNA occurs through interaction with the processivity factor. *Cell* 138(4):685-695
30. Shi Q*, et al.* (2015) Conformational toggling controls target site choice for the heteromeric transposase element Tn7. *Nucleic Acids Res*
31. Peters JE, Fricker AD, Kapili BJ, & Petassi MT (2014) Heteromeric transposase elements: generators of genomic islands across diverse bacteria. *Mol Microbiol* 93(6):1084-1092
32. May EW & Craig NL (1996) Switching from cut-and-paste to replicative Tn7 transposition. *Science* 272:401-404
33. Choi KY, Li Y, Sarnovsky R, & Craig NL (2013) Direct interaction between the TnsA and TnsB subunits controls the heteromeric Tn7 transposase. *Proceedings of the National Academy of Sciences* 110(22):E2038-2045
34. Hickman AB*, et al.* (2000) Unexpected structural diversity in DNA recombination: the restriction endonuclease connection. *Mol. Cell* 5(6):1025-1034
35. Arciszewska LK & Craig NL (1991) Interaction of the Tn7-encoded transposition protein TnsB with the ends of the transposon. *Nucl. Acids Res.* 19:5021-5029
36. Gary PA, Biery MC, Bainton RJ, & Craig NL (1996) Multiple DNA processing reactions underlie Tn7 transposition. *J. Mol. Biol.* 257:301-316
37. Holder JW & Craig NL (2010) Architecture of the Tn7 Posttransposition Complex: An Elaborate Nucleoprotein Structure. *Journal of Molecular Biology* 401(2):167-181
38. Rose A (2010) TnAbaR1: a novel Tn7-related transposon in Acinetobacter baumannii that contributes to the accumulation and dissemination of large repertoires of resistance genes. *Bioscience Horizons* 3(1):40-48
39. Hamidian M & Hall RM (2011) AbaR4 replaces AbaR3 in a carbapenem-resistant Acinetobacter baumannii isolate belonging to global clone 1 from an Australian hospital. *Journal of Antimicrobial Chemotherapy* 66(11):2484-2491
40. Sugiyama T, Iida T, Izutsu K, Park KS, & Honda T (2008) Precise region and the character of the pathogenicity island in clinical Vibrio parahaemolyticus strains. *J Bacteriol* 190(5):1835-1837
41. Parks AR & Peters JE (2009) Tn7 elements: engendering diversity from chromosomes to episomes. *Plasmid* 61(1):1-14
42. Bainton RJ, Kubo KM, Feng JN, & Craig NL (1993) Tn7 transposition: target DNA recognition is mediated by multiple Tn7-encoded proteins in a purified in vitro system. *Cell* 72(6):931-943
43. McKown RL, Orle KA, Chen T, & Craig NL (1988) Sequence requirements of Escherichia coli attTn7, a specific site of transposon Tn7 insertion. *Journal of Bacteriology* 170(1):352-358
44. Rao JE, Miller PS, & Craig NL (2000) Recognition of triple-helical DNA structures by transposon Tn7. *Proc. Natl. Acad. Sci. USA* 97:3936-3941
45. Mojica FJ, Diez-Villasenor C, Garcia-Martinez J, & Almendros C (2009) Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* 155(Pt 3):733-740

46. Shah SA, Hansen NR, & Garrett RA (2009) Distribution of CRISPR spacer matches in viruses and plasmids of crenarchaeal acidothermophiles and implications for their inhibitory mechanism. *Biochem Soc Trans* 37(Pt 1):23-28
47. Stern A, Keren L, Wurtzel O, Amitai G, & Sorek R (2010) Self-targeting by CRISPR: gene regulation or autoimmunity? *Trends Genet* 26(8):335-340
48. Bainton RJ, Kubo KM, Feng J-N, & Craig NL (1993) Tn7 transposition: target DNA recognition is mediated by multiple Tn7-encoded proteins in a purified in vitro system. *Cell* 72:931-943
49. Rao JE, Miller PS, & Craig NL (2000) Recognition of triple-helical DNA structures by transposon Tn7. *Proceedings of the National Academy of Sciences* 97(8):3936-3941
50. Kuduvalli P, Rao JE, & Craig NL (2001) Target DNA structure plays a critical role in Tn7 transposition. *EMBO J.* 20(4):924-932
51. Waddell CS & Craig NL (1988) Tn7 transposition: two transposition pathways directed by five Tn7-encoded genes. *Genes & Development*:137-149
52. Stellwagen A & Craig NL (1997) Gain-of-function mutations in TnsC, an ATP-dependent transposition protein which activates the bacterial transposon Tn7. *Genetics* 145:573-585
53. Biery MC, Steward F, Stellwagen AE, Raleigh EA, & Craig NL (2000) A simple *in vitro* Tn7-based transposition system with low target site selectivity for genome and gene analysis. *Nucleic Acids Res.* 28:1067-1077
54. Rao JE & Craig NL (2001) Selective recognition of pyrimidine motif triplexes by a protein encoded by the bacterial transposon Tn7. *J Mol. Biol.* 307:1161-1170
55. Peters JE & Craig Nancy L (2001) Tn7 recognizes transposition target structures associated with DNA replication using the DNA-binding protein TnsE. *Genes and Development* 15(6):737-747
56. Rutkauskas M*, et al.* (2015) Directional R-Loop Formation by the CRISPR-Cas Surveillance Complex Cascade Provides Efficient Off-Target Site Rejection. *Cell Rep*
57. Levy A*, et al.* (2015) CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature* 520(7548):505-510
58. Koonin EV & Krupovic M (2015) A moveable defense. *The Scientist* (January 2015)
59. Krupovic M, Makarova KS, Forterre P, Prangishvili D, & Koonin EV (2014) Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. *BMC Biol* 12(1):36
60. Seed KD, Lazinski DW, Calderwood SB, & Camilli A (2013) A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. *Nature* 494(7438):489-491
61. Besemer J, Lomsadze A, & Borodovsky M (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* 29(12):2607-2618
62. Marchler-Bauer A*, et al.* (2013) CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Res* 41(Database issue):D348-352
63. Grissa I, Vergnaud G, & Pourcel C (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* 35(Web Server issue):W52-57
64. Edgar RC (2007) PILER-CR: fast and accurate identification of CRISPR repeats. *BMC Bioinformatics* 8:18
65. Morgulis A*, et al.* (2008) Database indexing for production MegaBLAST searches. *Bioinformatics* 24(16):1757-1764
66. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5):1792-1797
67. Katoh K & Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30(4):772-780

68. Price MN, Dehal PS, & Arkin AP (2010) FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* 5(3):e9490

69. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19):2460-2461

70. Soding J (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics* 21(7):951-960

Figures legends

Figure 1 - Schematic representation of the complete and minimal type I-F CRISPR-Cas systems and Tn7 transposition. A. Gene organizations of a complete and a minimal type I-F CRISPR-Cas system lacking the genes for proteins responsible for adaptation and target cleavage. Minimal I-F systems contain fused *cas8f* and *cas5f* genes that are characteristic of this group (2). Together, these proteins can be predicted to be subunits of a minimal Cascade complex. B. Gene structure of the Tn7 genes flanked by left (L) and right (R) end sequences. Transposition catalyzed by the TnsABC+TnsD proteins directs the transposon into a single chromosomal site in bacterial genomes (*attTn7*). Transposition catalyzed by the TnsABC+TnsE proteins preferentially directs transposition into actively conjugating DNA and filamentous bacteriophage (shown by a red circle with arrows). The transposon is denoted by a rectangle in the attachment site. DNA sequence omitted in the graphic is denoted with two back slashes. See text for details.

Figure 2 Schematic evolutionary trees for the Cas7f, TnsA and TniQ/TnsD protein families.

A. The dendrogram was built using 2905 Cas7f proteins as described in Material and Method section (see complete tree at ftp://ftp.ncbi.nih.gov/pub/makarova/supplement/Peters_at_al_2017/). The major subtrees are collapsed and shown by triangles. The branch corresponding to the minimal I-F variant is colored in orange, and the bootstrap value for this subtree is shown.

B. The dendrogram was built using 7023 TnsA protein sequences (see complete at ftp://ftp.ncbi.nih.gov/pub/makarova/supplement/Peters_at_al_2017/). The branch corresponding to TnsA in the loci containing I-F variant *cas* genes is colored in orange

606 and I-B subtype *cas* genes are colored in green. The CRISPR-Cas subtypes are indicated
607 next to the respective branches. Distinct cyanobacterial strains are indicated next to the
608 respective I-B systems. The bootstrap value for TnsA branch associated with I-F *cas*
609 genes is shown.

610 **C.** The dendrogram was built using 7963 TniQ proteins (see complete tree at
611 ftp://ftp.ncbi.nih.gov/pub/makarova/supplement/Peters_at_al_2017/). The designations
612 are the same as for the TnsA dendrogram in B.

613

614

615 **Figure 3. Schematic representation of Tn7, Tn6022, and selected Tn7-like transposons**
616 **containing *cas* genes.** Genomic features recognized by the transposon-encoded TniQ protein
617 are indicated on the left (*glmS, yifB,* IMPDH, *yciA* and SRP-RNA). Color coding and labeling
618 are as in Figure 1. Elements other than Tn7 and Tn6022 are denoted by the respective TnsA tree
619 leaves (#XX)(Tn6022=Tree node #582)(Supplemental Table S2). Other genes are shown in
620 grey, and known Tn7 cargo genes are indicated. Black vertical bars indicate repeats in the
621 element-encoded arrays. DNA sequences omitted in the graphic is shown with two back slashes.
622 See text for details.

623

624 **Figure 4. Phylogenetic tree of selected representatives of type I-F-associated TniQ-like**
625 **proteins.** A maximum likelihood phylogenetic tree was built as described in Materials and
626 Methods for a selected set of TniQ-like proteins associated with the type I-F CRISPR-Cas
627 variant and a few sequences from the outgroup outlined in the Figure 2C (67 sequences
628 altogether). The numbers at internal branches indicate bootstrap support (percent); only values

greater than 70% are indicated. Elements located in one of the three attachment sites identified in this work are shown by color as indicated (*yciA*, IMPDH, and SRP-RNA); random sites . The TniQ tree leaves (#XX)(Supplemental Table S2) are shown in green.

Figure 5. Model of the two targeting pathways for Tn7 elements containing CRISPR-Cas system. Designations are the same as in Figure 1.

Figure 6. Models of the three previously described Tn7 targeting pathways and the proposed CRISPR-Cas-facilitated transposition pathway. Representations of TnsABC+TnsD (A) and TnsABC+TnsE (B) transposition pathways, the synthetic transposition pathway that targets triplex DNA complexes with a mutant form of TnsC, TnsABC* (C) and the proposed targeting pathway mediated by Cas interference complexes (D) are shown. Known host factors that participate in the TnsD and TnsE pathways are also shown (ACP, L29, and DnaN). See text for details and references.

651

652 **Table 1. Characteristics of selected Tn7 loci associated with *cas* genes.**

TnsA Node[1]	Strain	(Right)TSD.end.TnsB-binding/ //TnsB-binding.end.TSD (Left)[2]	Insertion position	Size (bp)
2	*Vibrio_parahaemolyticus_RIMD_2210633_GCA_000196095.1*	gagttgtaaatacaaccatacattgcaacaatac// //tataaatgtcactttatggttgtatcaacagagtt	*yciA* (17 bp upstream)	80,041
141	Vibrio_campbellii_ATCC_BAA_1116_GCA_000017705.1	ttaaatgtaaaaacaactaaacgttgaatttacga// //ttttataaaccatggttaattatttttacattaaa	IMPDH (21 bp downstream)	27,150
199	*Vibrio_parahaemolyticus_O1_Kuk_FDA_R31_GCA_000430405.1*	tgagttgttgatacaaccataaaatgataattaca// //tataaatatcactttatggttgtatcaacatgagt	*yciA* (17 bp upstream)	107,025
262	Vibrio_fluvialis_ATCC_33809_GCA_001558415.1	aaaattgttgaaacaaccataaattgatatttaca// //tatgaatatcaagatatggttgtatcaacaaaaat //cgtaaatatcaatttatggttgtatcaacagaaaa	IMPDH (21 bp downstream)	43,582 56,066
298	Vibrio_cholerae_VC35_GCA_000299495.2	ccatctgatgtttgcaaaataagttcgcataaatt// //tgtctatgcagacttatgctgcaagcatcaccatc	SRP-RNA (6 bp downstream)	28,955
330	Vibrio_natriegens_NBRC_15636_ATCC_14048_DSM_759_GCA_000417905	cattgtgaagcctgcaatatatgttcgcataaatt// //ggactatgctaaattacgttgcaggcatcacattg	SRP-RNA (6 bp downstream)	23,147

	5.1			
340	Vibrio_crassostreae_J5_20_GCA_001048515.1	cttaatgaagcctgcaatatatgttcgcataaatt// //tagttatgcaggcatatgttgcaagcatcacttaa	SRP-RNA (6 bp downstream)	26,389
375	Pseudoalteromonas_translucida_KMM_520_GCA_001465295.1	ttgggtgttgtttgaagtataagttgacatatctg// //tacttatgccaacttatacttcaaacaacattggg	SRP-RNA (22 bp downstream)	21,768
409	Vibrio_cholerae_YB2A06_GCA_001402375.1	catcttgtcgctgaaagcataaagtgtccaattaa// //agcataggacatcttatgctttcagcgacacatct	SRP-RNA (7 bp downstream)	34,953
424	Vibrio_hyugaensis_151112A_GCA_000818475.1	aaagctgtcgttaaaacgataacctgtcccaataa// //gtggttggacgcattatggttttagcgacaaaagc	SRP-RNA (11 bp downstream)	30,851
438	Vibrio_EJY3_EJY3_GCA_000241385.1	gctggtgtggctggaaccataagatgacattttag//? //taagggtgtcaccttatggtcctagccacagctgg	AraC type (16bp downstream)	117,187
446	Shewanella_ANA_3_ANA_3_GCA_000203935.1	tgagttgtcgctgaacccatacattgacataattg// //ctgctgtgtcatggtttggtttcagcgacatgagt	Conserved (inside the gene)	23,831
2757	Cyanothece_PCC_8801_PCC_8801_GCA_000021805.1	aagtttgtttttgccgtattttaaagtttgttttt // //aggcataaccttgccgatgcggcaaaaacaaactt	DNA methyltransferase (711 bp downstream)	16,228
5291	Anabaena_variabilis_ATCC_29413_GCA_000204075.1	ataaatgtggaggaataattcggttcaacaatata// //ataattgtttaaccgaattctttgtctacaaatat	*glmS* (5 bp)	31,369

[1] TnsA node number for each element cross-lists as indicated in Figures 3 and 4 and supplemental Table S2.

656 [2] The inferred target site duplication (TSD)(red and underlined), 8 base pair end sequence (blue),
657 and terminal most TnsB-binding site (in green) are indicated for the Right (Right) and Left (Left)
658 transposon ends identified for each element indicated in Figure 3. See supplemental Figure S1
659 for details.
660

Table 2. Spacers with identified matches

TnsA node	DNA sequence with element	Spacer (3bp 3') Match(PAM)[2]	Match	Hit accession number
108, 120, 122	Vibrio_parahaemolyticus_VIP4_0434_GCA_000500425.1 Vibrio_parahaemolyticus_07_2965_GCA_000960565.1 Vibrio_parahaemolyticus_S028_GCA_000491615.1	Catgtcaggggtaaaactcactgagccaacaa(att) Catatcaggggtaaaactcactgagccaacaa(tat)	**Chromosome**, element itself	CP011406.1 CP003973.1 BA000032.2
267	Vibrio_cholerae_HE_45_GCA_000279285.1	Attgaggacatagacaaaacttggatctttaa(gtt) Atcgaggacaaagacaaaacttggatctttaa(agt)	**Plasmid**, integrase (integron), *Vibrio tubiashii* ATCC 19109	CP009358.1
		Ctgaaaagcaatgaagccaaagcgtcctgtaa(gtt) Ctgaaaagcaatgaagcgaaaacgtcctgcag(ggc)	**Plasmid**, DNA Polymerase *Methylophaga frappieri* strain JAM7 plasmid	CP003381.1
271	Photobacterium_kishitanii_201212X_GCA_001455895.1	Aatgtctacacattacaaggcttacttgcggc(gtg) Aatgtctacgctttacaaggcttactcgctgc(aaa)	**Chromosome**, match 51 bp from right end of insertion	LNTE01000012
436	Photobacterium_leiognathi_mandapamensis_GCA_001558075.1	Aaagtatgggaacggagaaacggttctttgc(gtg) Agatattaggaacggagaaacggctctttgc(tta)	**Chromosome**, match 30 bp from right end of insertion	LNRA01000001
139	Vibrio_parahaemolyticus_S019_GCA_000491795.1	Aaatcattgaccaaacataccacaaaatcact(att) Aaatcattgaccagacatatcacaaaatcaca(ggt)	**Plasmid**, hypothetical protein, *Vibrio parahaemolyticus* strain VPS92 plasmid pVPS92-VEB	KU356480.1
		Aaatcattgaccaaacataccacaaaatcact(att) Aaatcattgaccagacatatcacaaaattatt(ggt)	**Plasmid**, hypothetical protein,	LN831184.1

			Vibrio cholerae 116-14, plasmid / pNDM-116-14	
145	Marinomonas_S3726_S3726_GCA_000967665.1	Tataccaccaaagaaaccgatggtagcttcaa(gtg) Tatactaccaaagaaacagatggtagcttcaa(gca)	**Chromosome**, catabolite regulation protein CreA, *Vibrio anguillarum* strain 90-11-286 chromosome II	CP011461.1 Many others
285	Vibrio_vulnificus_SC9740_GCA_000959765.1	Atatcagcaatagctagcatactcactgtgtc(gtg) Atatcagcaattgctagcatactcacagtgtc(cgt)	**Plasmid**, hypothetical gene, *Vibrio vulnificus* strain FORC_017	CP012741.1 2 other plasmids KX765275.1 CP009849.1
311	Vibrio_alginolyticus_UCD_30C_GCA_001306785.1	Aactagtaaatacactcgaaacatcattatta(gtg) Aactagtaaatacactcgaaacatcattatta(act)	**Plasmid**, RdgC superfamily protein, *Vibrio alginolyticus* strain ATCC 33787 plasmid pMBL287	CP013487.1
353	Vibrio_fischeri_MJ11_GCA_000020845.1	Ttatagagaaattcagcattcgcatgaaggt(gta) Ttatagagaaactcagcattcgcatgcaggt(aag)	**Plasmid**, Repair ATPase, Plasmid found in the same strain	CP001134.1
446	Shewanella_ANA_3_ANA_3_GCA_000203935.1	Gtttgattttaaagaaggtcttgataaatagt(gtg) Gtttgattttaaagaaggtttgataaatagt(ttt)	**Plasmid,** Hypothetical gene, *Shewanella xiamenensis* strain T17 / pSx1	CP013115.1
		Agcgtggggtaatcgaggtggtctgcttcatc(gtg) Agcgtggggtaatcgaggtggtcggcttcatc(cac)	**Bacteriophage**, MuB ATPase, *Shewanella oneidensis* MR-1	AE014299.2
263	Vibrio_parahaemolyticus_ISF_25_6_GCA_001267595.1	Acaaaacaaagttgtacactgttgtctaccgt(gtt) Acaaaacaaagttgaacgctgttatctaccgt(gtc)	**Plasmid,** *Vibrio coralliilyticus* strain	CP009619.1

			RE98 plasmid p380	
194	Neptunomonas_japonica_DSM_18939_GCA_000422765.1	Tgcaggctggttcgccaattgcaacgctgttt(gtg) Tgcaacctggttcgccaatcgcaacgctggtt(cag)	**Chromosome,** hemolysin D gene *Rhodopirellula baltica* SH 1	BX294151.1
399	Ferrimonas_senticii_DSM_18821_GCA_000422665.1	Tacagtgttgatgaggcgtttatgacgctggc(gtg) Gacagagtttatgacgcgtttatgacgctggc(gtg)	**Chromosome**, Intergenic, *Pusillimonas* sp. T7-7	CP002663.1
407	Vibrio_cholerae_YB3G04_GCA_001402275.1	Tcagagcgtattagacggacgcgcatcagaca(gtg) Tcagagcgtattagacggacgcgcatcagata(taa)	**Plasmid,** *Vibrio parahaemolyticus* plasmid pVPS91	KX957972.1
5284#	Cyanothece_PCC_7822_PCC_7822_GCA_000147335.1	Aatggaacgggcaaaatgtaaaatgtaaacgaatccc(tga) Aatgggacgggcaaaatgcaaaatataaacgtatccc(atg)	**Plasmid,** Cyanothece sp. PCC 7822 plasmid Cy782202	CP002200.1 Same plasmid

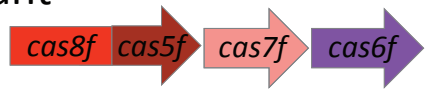
[1] TnsA node number for each element cross-lists as indicated in Supplemental Table S2.

2 Spacers were identified as described in the Materials and Methods section. Spacers (and 3 base pair located 3' in the repeat are indicated) for comparison with the match identified in the database (presumed PAM sequences). Mismatches between the spacer (3' sequence) and protospacer (PAM) are indicated in red.

A

canonical subtype I-F system

subtype I-F variant



B

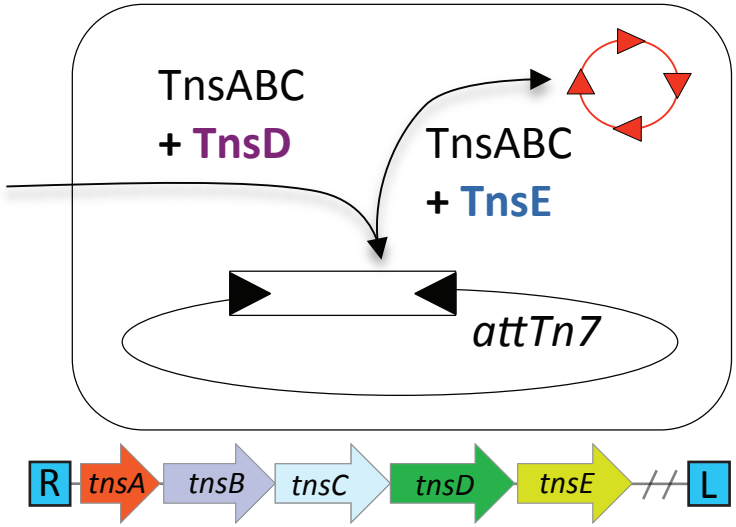
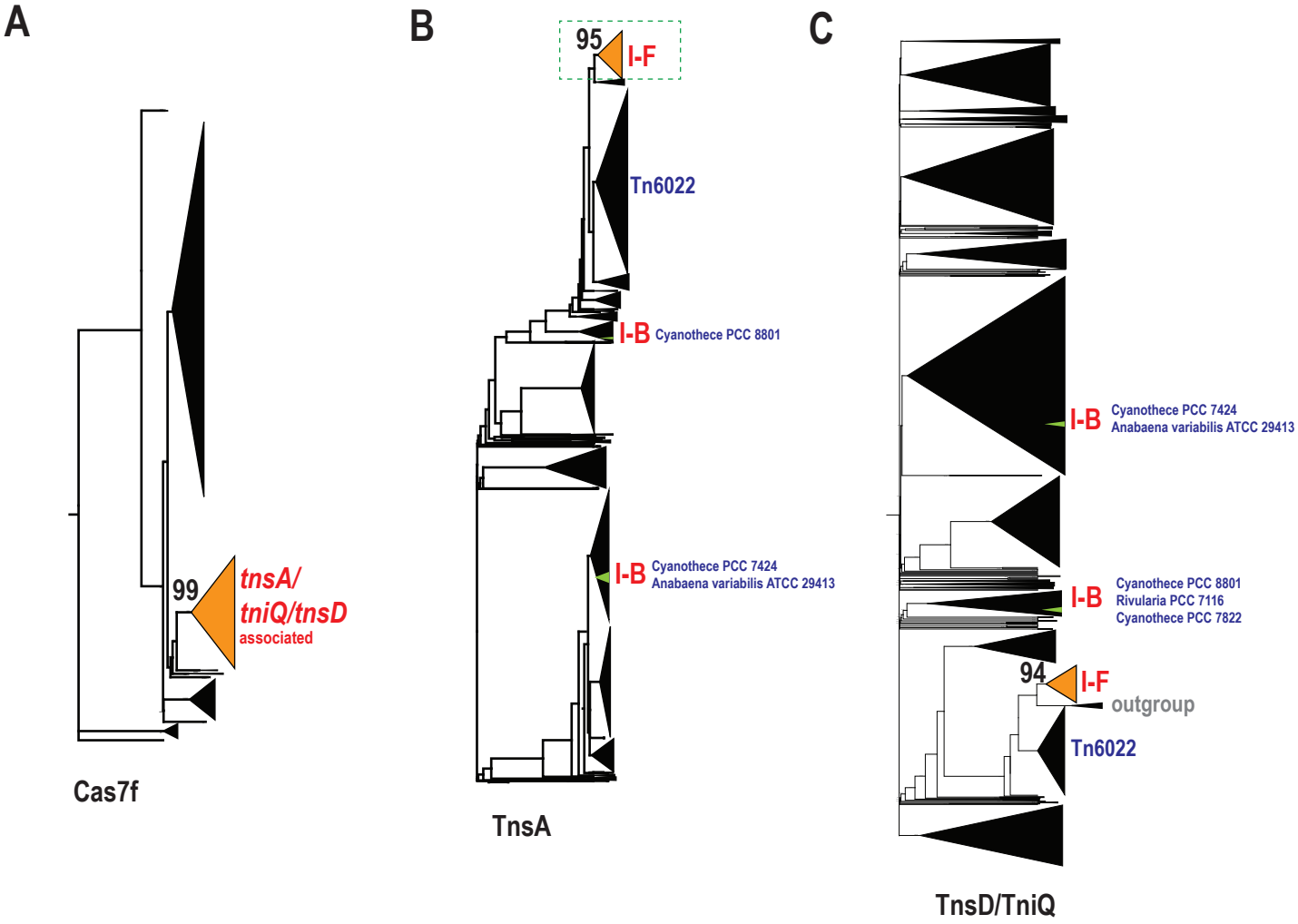


Figure 1



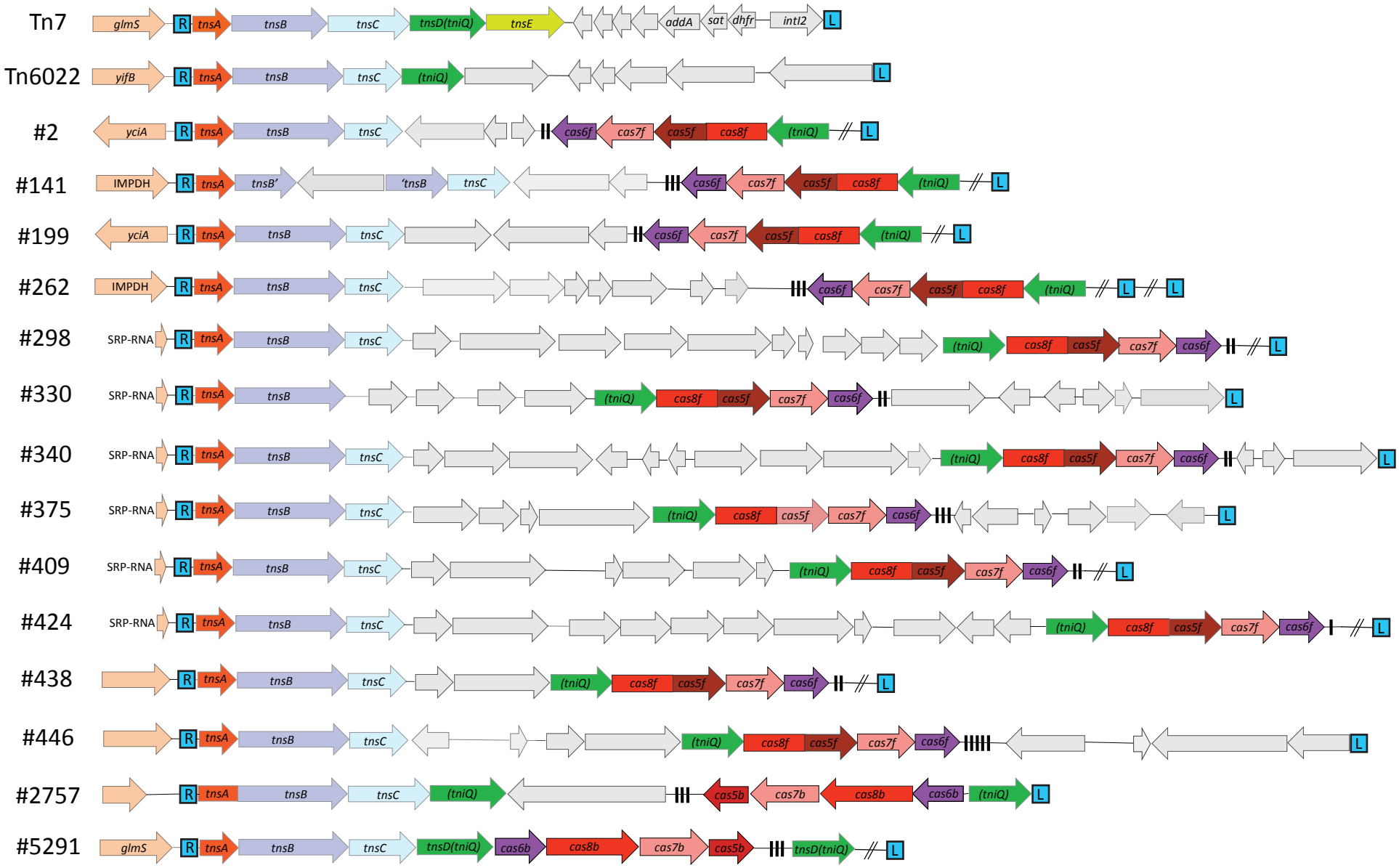

Tn7
glmS
R
tnsA
tnsB
tnsC
tnsD(tniQ)
tnsE
addA
sat
dhfr
intI2
L
Tn6022
yifB
(tniQ)
#2
yciA
cas6f
cas7f
cas5f
cas8f
#141
IMPDH
tnsB'
'tnsB
#199
#262
#298
SRP-RNA
#330
#340
#375
#409
#424
#438
#446
#2757
cas5b
cas7b
cas8b
cas6b
#5291

Figure 3

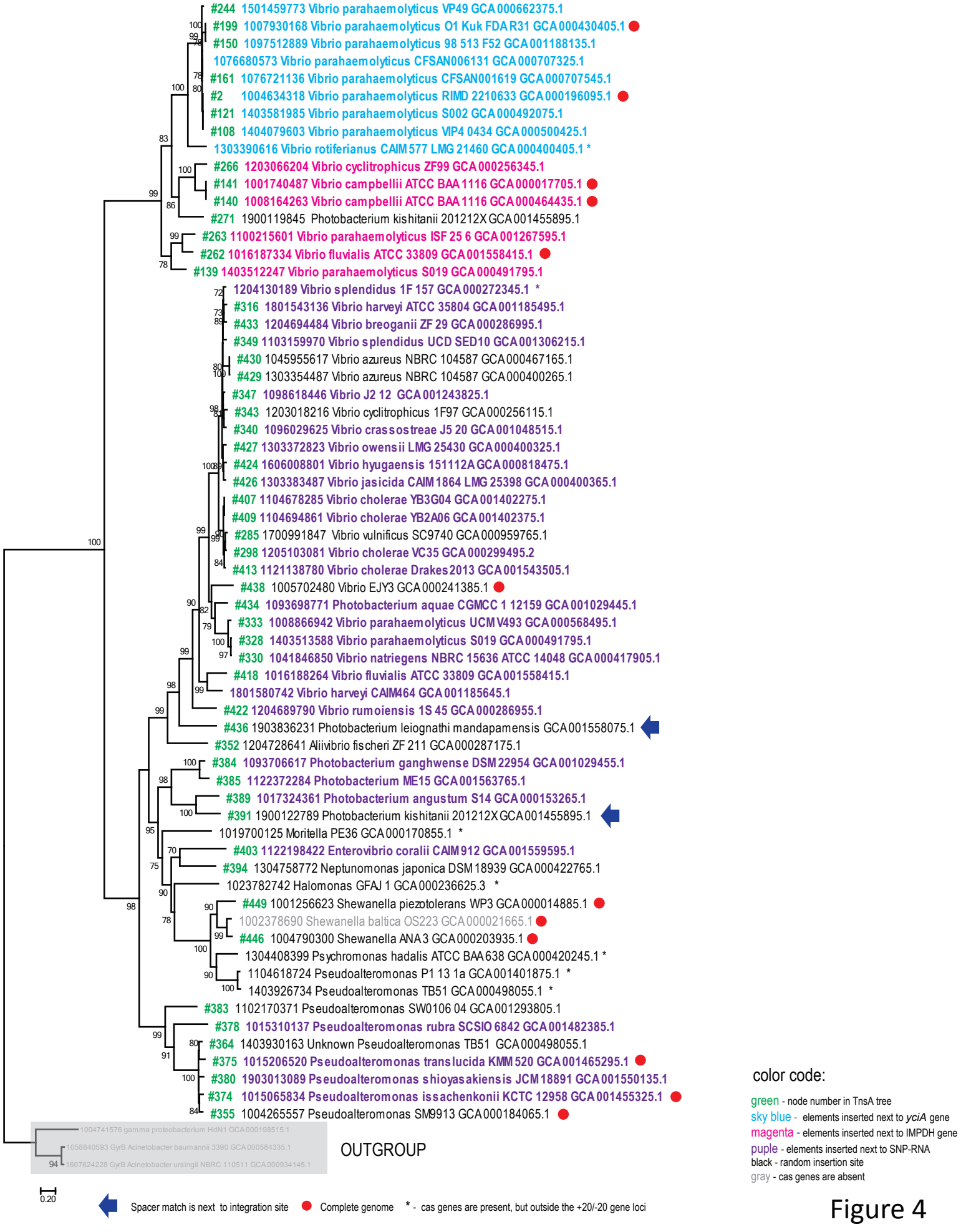


Figure 4

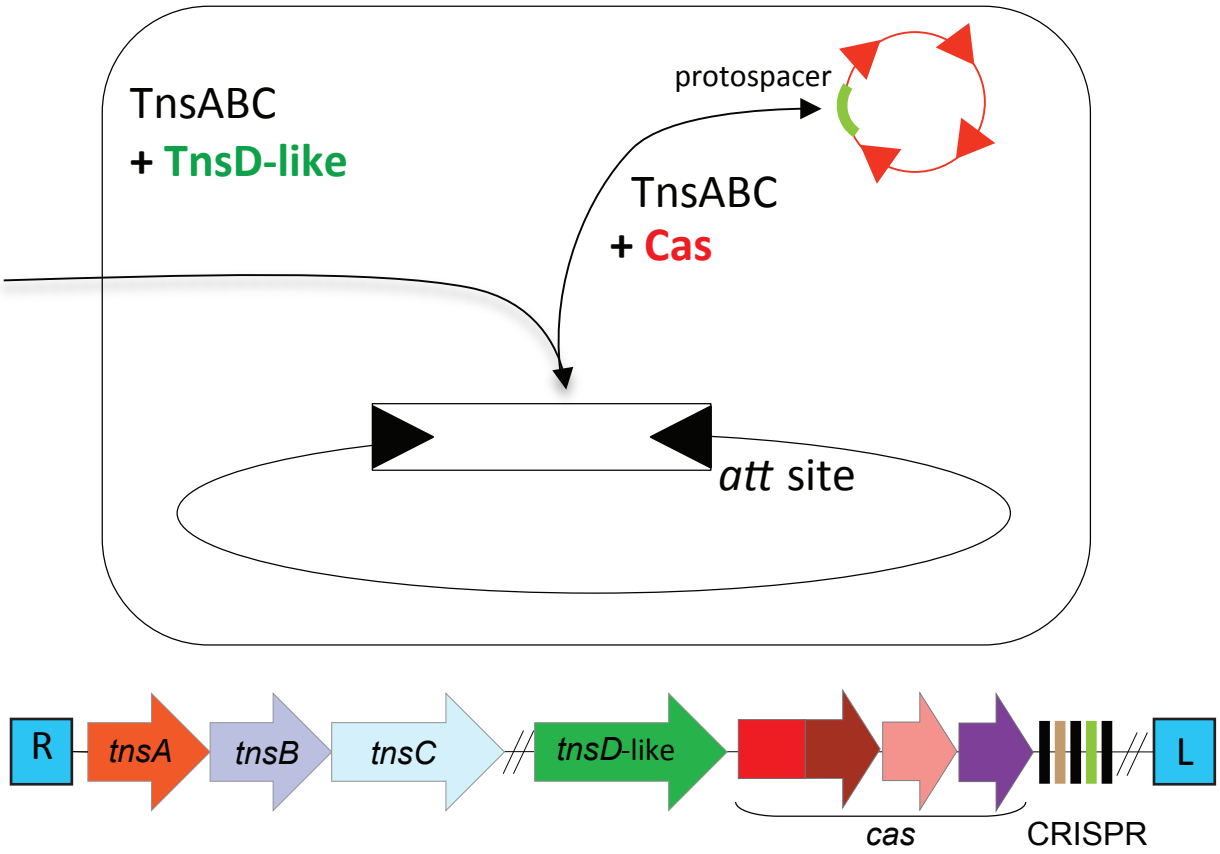


Figure 6

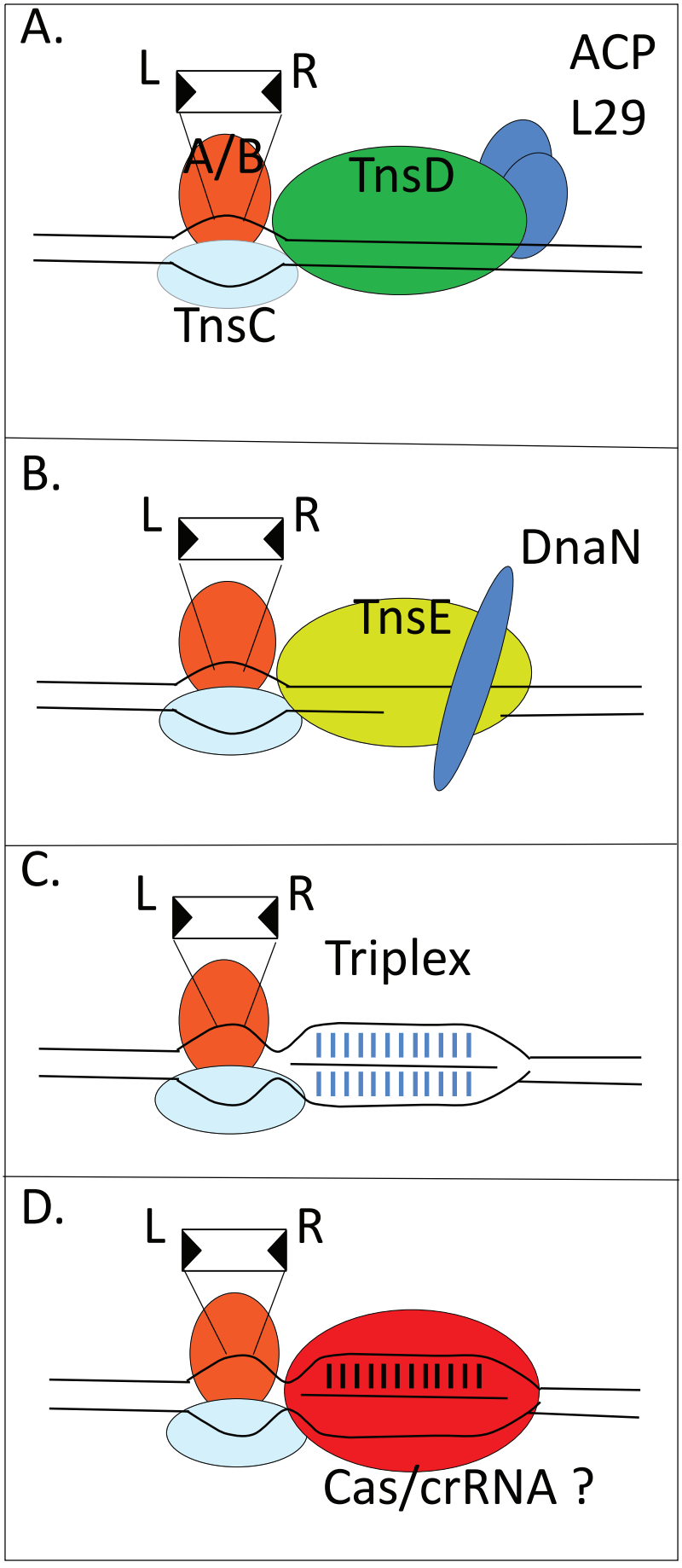
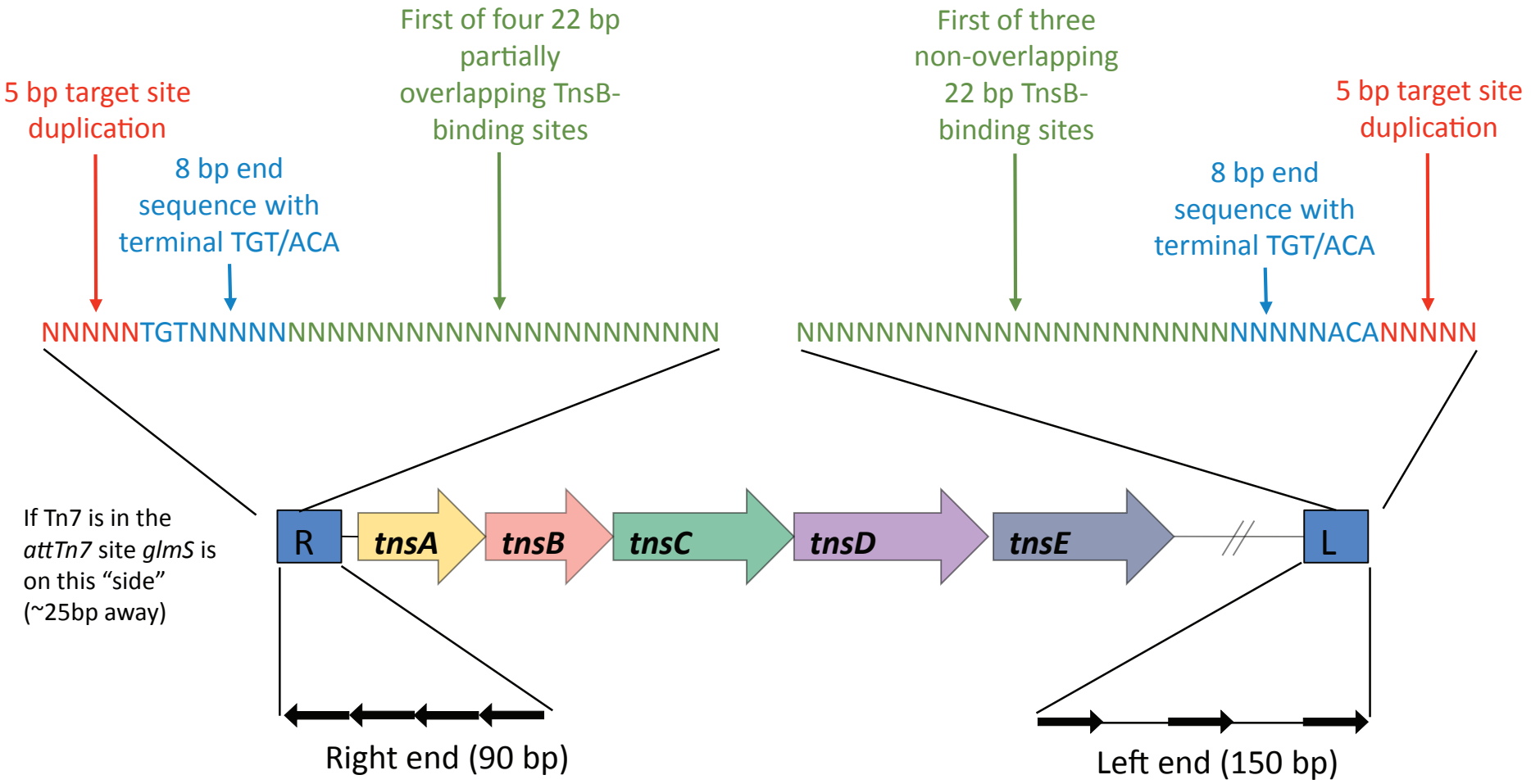


Figure 6



Supplemental Figure S1

Schematic representation of the end structure of Tn7-like elements – Anatomy of a Tn7 insertion. Typically the insertion occurs at a single site about 25 bp downstream from the last codon of glmS. The Tn7 end proximal to tnsA is closest to glmS (by convention it is referred to as the "right" end). Transposition generates a target site duplication (shown in red) of the chromosomal sequence that now forms a direct repeat on either side the element. In the case of insertion of the canonical Tn7element at the attTn7 site this sequence is of GCGGG.
There will be an 8 bp "end sequence" that starts with TGT/CAC, immediately after this will be the first 22 bp binding site for TnsB.