Structure and Function of a Transposon-Encoded CRISPR-Cas System

Tyler S. Halpin-Healy

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2021

Abstract

Structure and Function of a Transposon-Encoded CRISPR-Cas System

Tyler S. Halpin-Healy

CRISPR-Cas defense systems are employed by their hosts to prevent parasitization by mobile genetic elements. The discovery of nuclease-deficient CRISPR-Cas systems contained within transposon ends suggested a repurposing of the contained defense system. One such Type I-F3 CRISPR-Cas system was found inside Tn*6677*, a Tn*7*-like transposon within the genome of a *Vibrio cholerae* strain. Tn*6677* requires coordination between the contained CRISPR-Cas system and the transposition proteins for effective transposition. Isolation of this system, and reduction to its minimal components, enabled RNA-guided integration of donor DNA in *Escherichia coli*. Base-pairing interactions between the user-specified CRISPR RNA and the target sequence precede the integration of donor DNA approximately 49-bp downstream of the end of the target sequence. This system is specific regardless of the supplied RNA guide, and successfully integrates donors of different lengths. The donor DNA is indicated by flanking cognate transposon end sequences. While clearly functional, the mechanism by which the transposition proteins and the CRISPR-Cas proteins interact remained unclear. To this end we purified the multi-protein RNA-guided DNA binding complex (Cascade) from the transposon-encoded minimal I-F3 CRISPR-Cas system in complex with the transposition protein TniQ. *De novo* modeling revealed the unexpected dimerization of TniQ, and its location within the complex, bound to the Cas6-end of the transposon-encoded Type I-F3 Cascade. Additional models obtained from DNA-bound structures of the complex demonstrate initial steps in target binding alongside novel conformations of

Cascade subunits. This work reveals the mechanism by which the Tn*6677* components guide integration and will enable rational engineering of these systems for further experimentation and tool development.

Table of Contents

List of Charts, Graphs, Illustrations

Acknowledgments

The work described within would have been impossible without the support of the following people. Sam, for always making time, and providing more support than I would have even thought to ask for. His good-naturedness and curiosity created an environment I consider myself unreasonably lucky to have been a part of. Israel, for his generosity and willingness to show me the ins-and-outs of his specialty. Alex, for his patience and pragmatism with both my project and my incessant questions. Harris, for always keeping a level head and for his prescient suggestions. Sanne and Leo, for being stalwart colleagues, but more for being my friends and keeping me grounded. Marcus for his reliability and enthusiasm, but particularly for his patience with me as a mentor. Leslie, for showing me the many ways in which to conduct not only science, but oneself. Nora, for keeping this ship run tightly, and for her overall amiability. Florian, for his meticulousness and ensuring the lab is stocked with some sort of sugar. Ethan and Jason, again for the cheer they bring to the lab, but also their willingness to contribute where they may. Minjoo, for her rigor and inclination to assist wherever possible. Chance, for his encyclopedic knowledge, but also for his humility, reminding us there is always more to learn. Diego, for his outspokenness and, always kind, candor. Chris, for always providing a different view, and for his eagerness. Jerrin, for his insights and patience. Lastly, Rebeca and George, for helping me grow as a mentor, but also for their eagerness and diligence; I rest easily knowing I am leaving my unfinished work in their hands.

The program directors and coordinator, Ron Liem for keeping a steady hand during all conditions and guiding me in my early steps. Donna Farber, for her patience and the time she took to converse. And of course, Zaia, consistently checking in and ensuring I was on the right path.

My classmates, in particular Luke McGoldrick, Sreyan Chowdhury, Samhita Rao, and Charlotte Wayne. They provided a bulwark against the more difficult times. Micah Rapp, Felix Wu, Samuel Hayward, Davys Lopez, Corey Garyn, and Bex Cole for some escapism. Lastly, Albana Kodra, Virginia Aimiuwu, and Manny Tamargo for their good cheer.

My friends, who kept my head on my shoulders this entire time. Joe, Caroline, Gabriel, Caitlin, Ismael, Mike, and David, who all had to support me in different ways at different times. My sister, Anna, for always lending an ear when needed. My Mother and Father, for their unconditional support my entire life, but particularly so during the past five years. Bear, for ensuring I stay humble. But most importantly, Julie, who has been through this all with me, from before the beginning, and has done nothing but patiently support me through this entire endeavor.

Dedication

For Julie, who was always there for me.

And for my heroes, CAK, GDY, & TJHH.

Chapter 1: Transposon biology and general mechanisms of transposition

1.1: A primer on transposons

Transposons, or Transposable Elements (TEs), are obligate genomic parasites defined by their ability to move from one position to another[1]. When Dr. Barbara McClintock discovered the Ac/Ds system in Maize in 1950, she revealed a category of mobile DNA elements that we now have come to realize are not only ubiquitous throughout many organisms but are also exceedingly overrepresented compared to other DNA elements[2,3,4]. These mobile DNA elements are so common that most laypeople have encountered them, with jewel maize used as a decoration, the British peppered moth that is used as an example of evolution in nearly every secondary-school biology textbook, or the RAG proteins that drive antibody generation in our own immune system[2,5,6]. Correlative with transposon mobility is the resulting genomic rearrangement, driving both genomic and organismal evolution[7].

TEs vary widely in their features and their mechanisms of transposition. A consequence of this variation is that only the most general definition can describe TEs: that a TE is a discrete DNA sequence capable of moving or copying itself from one position in a DNA strand to another position in a DNA strand[8]. The TE may encode all necessary factors for its mobilization (an autonomous TE), or may rely on other factors not encoded within the discrete DNA segment (a nonautonomous TE)[9]. In moving or copying, the TE may go through intermediate stages as DNA (DNA transposons) or RNA (retrotransposons)[10]. The possibility of massive genomic damage or rearrangement is increased with the inclusion of TEs in a genome, stemming from insertion of a TE into an open reading frame, an abundance of homologous regions proximal to each other, abnormal transposition chemistry yielding novel genomic rearrangements, or a transposition mechanism that brings along adjacent (or intervening) DNA sequences[11]. However, TE integration is not intrinsically catastrophic and can even be beneficial depending on the genetic context; examples include the introduction of regulatory regions suppressing or enhancing novel reading frames, polyadenylation (pA) signals placed upstream of exons, the addition of splice donors or acceptors resulting in additional exons, or even the in-frame addition of a partial sequence of a gene, ultimately yielding a fusion protein[4,10–12]. Additionally, TEs are often found harboring passenger genes that may benefit the host, ensuring it is more beneficial for the host to keep the TE and its progeny by offsetting the metabolic burden of maintaining them[13]. TE-induced massive genomic reorganizations will inevitably drive evolution, solely as a consequence of the diversity being generated[11].

Despite TEs being capable of both beneficial and deleterious genomic rearrangement, the latter are catastrophic enough to warrant the existence of defense mechanisms to prevent transposition[10,14,15]. Downregulation of transposition often occurs through epigenetic silencing mechanisms such as N6-methyladenine in *E. coli* and 5-methylcytosine in eukaryotes, or through RNA interference (RNAi) pathways using PIWI-domain containing Argonaute-family proteins that quarantine the transposon by degrading transposon gene RNA transcripts[10,15,16]. However, due to the mutualism between prokaryotic TEs and their hosts, these TEs will often regulate their own activity[12]. This autoregulation can occur through various means, whether by an antisense RNA (asRNA), or ensuring specificity of the target site-determining pathway (e.g. conjugative plasmids or a non-essential region of its host's genome), or by the efficiency of their transposition reaction chemistry [8,13,17,18,19].

1.2: General mechanisms of transposition

Despite the diversity of transposition mechanisms, when sufficiently abstracted, common elements can be found. The first shared element is the same feature responsible for the discreteness of the TE, the ends of the transposon. The ends are the terminal sequences of the DNA sequence that establish the boundaries of a TE. Repetitive features within the ends, and what type of transposon they flank, determine what the ends are named: some are Inverted Terminal Repeats (ITRs), others are Long Terminal Repeats (LTRs), others are simply ends[1]. Whatever their name, the ends of the TE serve to distinguish the TE from the host's genome. A lack of TE ends all but ensures that the protein effectors responsible for mobilizing the TE would be unable to bind to the TE[1]. The second element, regardless of the mechanism of transposition, is an effector protein to mobilize the sequence between the ends[20]. This may be a reverse transcriptase that can identify the transcribed TE, or it may be a transposase (also referred to as an integrase) specific to the TE that encodes it[21,22]. Additionally, the TE requires a targeting module, an effector that recognizes preferred target sites of the TE, whether via sequence or topological specificity[23,24]. The TE will also require a target site. Lastly, the TE requires an effector, the transposase/integrase, to integrate the mobilized TE into the target site specified by the targeting effector[1].

Transposons are first segregated into Class I and Class II transposons[8]. Class I transposons, also called retrotransposons, are TEs that transpose through an RNA-intermediate and as such, tend to encode a reverse transcriptase[8]. Class II transposons, or DNA transposons, are named as such, as they are mobilized by their cognate transposase from one DNA locus to a target site without proceeding through any RNA intermediates[25]. DNA transposons may be further categorized based on the catalytic domains (e.g. DDE, HUH) present in their transposase, and whether their transposition reaction involves ssDNA or dsDNA[8]. Despite there being a myriad of specific transposition mechanisms for DNA TEs, most pertinent to this thesis are the cut-and-paste (sometimes called simple transposition) pathway and the replicative pathway.

An exemplar of canonical cut-and-paste transposition, Tn5, and its mechanism of transposition is generalizable to other cut-and-paste Class II TEs. To initiate productive transposition, (1) Tn*5*'s transposase, TnpA, recognizes the TE ends and binds them[26]; (2) the Tn*5*-end-bound TnpA monomers dimerize to form a loop called the synaptic complex or transpososome[27]; (3) TnpA cleaves at the boundaries of the Tn*5* ends, excising the TE[27]; (4) the synaptic complex of the TE and Tn*5* TnpA dimer is joined to the target DNA by the Tn*5* TnpA dimer[20]; (5) the TnpA dimer integrates the synaptic complex at the target site by catalyzing a transesterification reaction with the free 3'-OH's of the synaptic complex; (6) the integrated TE is immediately flanked by short single-stranded DNA (ssDNA) sequences; (7) host gap repair mechanisms fill in the ssDNA stretches creating a canonical Target Site Duplication (TSD)[8]. The length of the TSD is determined by the distance between the sites of the strand transfer attack and can differ between different TEs[4]. As the name of the general mechanism suggests, this process results in the complete removal of the TE from its initial locus and its insertion at a new one. An archetype of this process is shown in Figure 1.

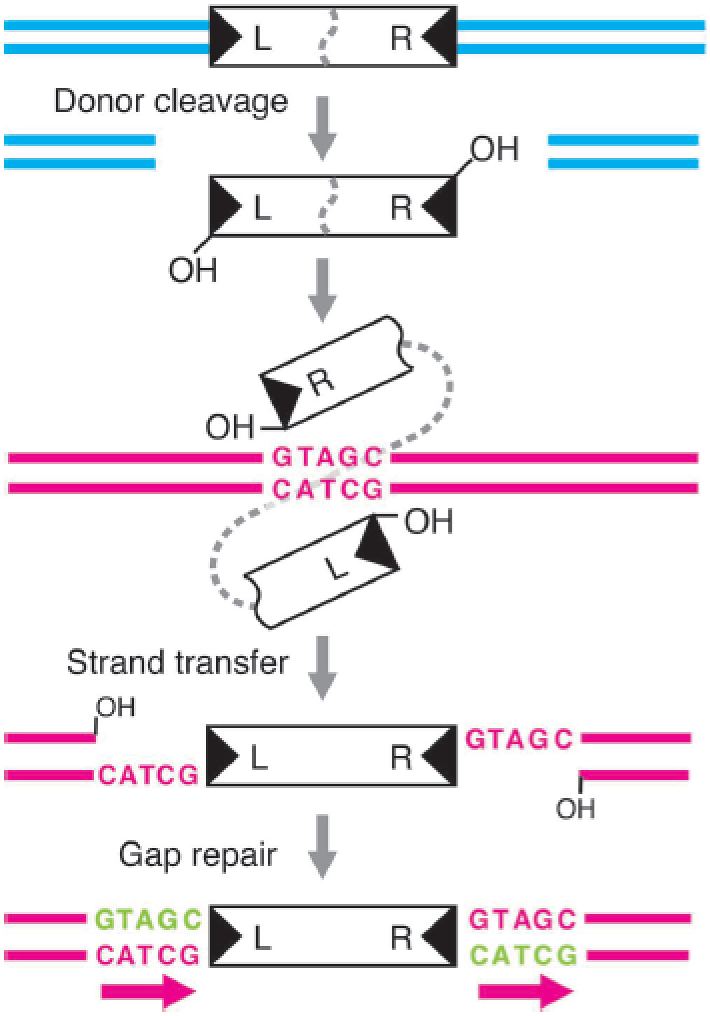


Figure 1: Mechanism of DNA cut-and-paste transposition by a DDE transposase

The DDE transposase recognizes the ends of the TE and binds them. Following binding, the transposase excises the TE. The bound transposase catalyzes the transesterification reaction and drives the free 3′-OH's attack of the target strand. The gaps formed as a byproduct of the integration reaction persist until host factors repair the ssDNA gap. Repair generates the hallmark target site duplications. Figure adapted from Craig, N. L. *et al. Mobile DNA III*[8].

Unlike cut-and-paste transposition, no step in replicative transposition ever yields a fully excised transposon[1]. The replicative transposition pathway, exemplified here by the Tn*3* TE and diagrammed in Figure 2,

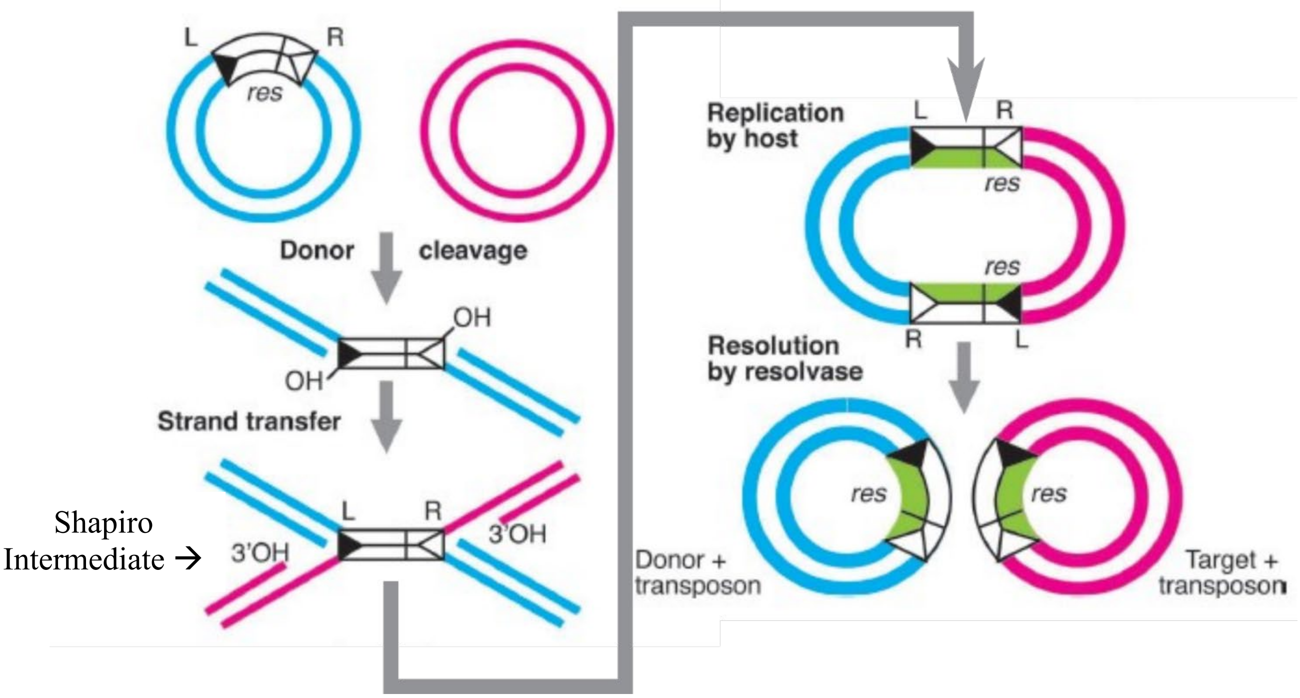


Figure 2: Mechanism of Replicative Transposition with a DDE transposase

When the TE and target site are proximal, the transposase binds and nicks the 3′-ends of the TE. The bound transposase catalyzes the free 3′-OHs attack of the target site. The 5′-ends of the TE are still bound to within the starting locus, resulting in a composite DNA molecule joined at a Shapiro Intermediate. Replication initiated in part by the topology of the Shapiro Intermediate produces a cointegrate. The cointegrate links the donor DNA and target DNA through the TE. Lack of a resolvase will cause the reaction to halt here. Introduction of a resolvase will effectively return the two dsDNA molecules to their beginning state, but with the TE replicated into the target DNA. Figure adapted from Craig, N. L. *et al. Mobile DNA III*[8].

proceeds as follows: (1) the Tn*3* transposase TnpA forms a complex with both the TE and the target DNA; (2) Tn*3* TnpA nicks the 3’ ends of the TE; (3) similarly to the cut-and-paste pathway, the nicked strands provide free 3’-OH’s resulting in invasion of the target DNA with an offset that will yield a TSD of the same length as the offset; (4) the invasion 3’-nicked dsDNA transposon creates a Shapiro Intermediate characterized by the topological similarity of the ends of the

TE:target DNA to replication forks; (5) replication is initiated by the host replisome at the free 3' ends of the target DNA, synthesizing the complementary strands of the TE, forming a cointegrate containing the entire donor molecule flanked by two repeats of the transposon sequence; (6) if this TE contains a resolvase (TnpR in the case of Tn*3*), the cointegrate may then undergo recombination to be separated into two double stranded (dsDNA) molecules[28]. The first molecule is an exact replica of the initial locus with the TE, and the second is the target locus interrupted with the replicated Tn*3*. Without the resolvase or other modes of assisted recombination, the cointegrate molecule will persist[29].

Both Tn*3* and Tn*5* utilize DDE transposases for their integration steps; named as such for their catalytic domain resembling an RNase H-like fold and the conservation of the eponymous amino acids[1,28]. This fold brings the DDE residues into proximity with each other, forming the active site of the transposase. This active site cleaves DNA phosphodiester bonds and forces the strand transfer of the invading TE DNA strand using the free 3'-OH on each strand as the nucleophile in the subsequent transesterification reaction[1,8].

1.3: Tn*7* transposon biology in its endogenous context

The transposon Tn*7* was first discovered in *E. coli* in 1976. Identified originally through its passenger genes that carry antibiotic resistance[30,31], it was later found in medical samples, hinting at its widespread distribution[8]. With the explosive growth of sequence repositories, Tn*7* and Tn*7*-like transposons have been found across many environments in widely diverged prokaryotes[32]. Tn*7* is the founding member of the Tn*7*-like transposon family that contains TEs with obvious Tn*7*-like characteristics, mainly the protein components encoded by *tnsA*, *tnsB*, *tnsC*, *tnsD*, and in the case of the seminal member, *tnsE*[8]. Broadly, Tn*7* transposition proceeds as follows: First, TnsB recognizes the ends of the TE and together with TnsA, forms the heteromeric transposase responsible for the unique transposition reaction mechanism of Tn*7*[33,34]. Second, the TnsA/TnsB heteromeric transposase brings the ends of the Tn*7* TE together to form the Paired-End Complex (PEC)[35]. The PEC is structurally and functionally analogous to Tn*5*'s synaptic complex given in the initial example of cut-and-paste transposition. Third, TnsC mediates the interaction between the PEC and the targeting proteins TnsD and TnsE (but never simultaneously)[36]. The heteromeric transposase TnsA/TnsB only excises the PEC when in complex with TnsC and either TnsD or TnsE at their specified target sites[37]. Tn*7*, like other DDE-family transposons, integrates the TE through a transesterification reaction and as such will yield a 5-bp TSD due to the 5-bp offset of the nucleophilic attacks from the TnsB integrases when attacking opposite strands of the target DNA[38]. Should this reaction proceed down the TnsE pathway, the PEC+TnsA/B/C is recruited to TnsE, a site-selecting protein that identifies and binds the 3'-recessed ends in the lagging-strand during DNA replication. The PEC+TnsA/B/C/E complex then integrates at the TnsE-bound free 3' DNA end[24]. This pathway facilitates horizontal gene transfer of Tn*7* by allowing Tn*7* to transpose onto conjugal plasmids[13]. Alternatively, should the transposition reaction proceed via TnsD instead of TnsE, the PEC+TnsA/B/C/D complex is recruited to 3'-end of the highly conserved glucosamine 6-phosphate synthase (*glmS*) gene through the sequence-specific binding of TnsD. The TnsD pathway results in Tn*7* integrating the TE just downstream of the coding region of *glmS* and allows Tn*7* to be transmitted vertically[22,39].

A curious feature of Tn*7* transposition is target site immunity. Wild type Tn*7* will transpose with severely reduced efficiency into a target site that has already been used for transposition and contains an integrated Tn*7* TE nearby[40]. This phenomenon can persist through closely related homologous Tn*7* systems but may be bypassed if the integrated TE and the transposing homolog are diverged enough[41]. Furthermore, target site immunity is not likely due to DNA-molecule specific effects, as the presence of Tn*7* ends on one of two spatially proximal, but separate plasmids intercalated with each other will reduce transposition into both[42]. Target immunity is hypothesized to be essential to preventing rampant genomic instability from sequential transposition into the same locus, as the target sequence recognized by TnsD is not disrupted by the first transposition event[8]. It is likely required due to 1) the high efficiency of Tn7 transposition, and 2) that a high concentration of homologous regions at a single locus will almost invariably result in recombination that may create deleterious chromosomal rearrangements[8].

1.4: Components and mechanistic details of *E. coli* Tn*7* transposition

Tn*7* ends

Denoted as Left and Right, the ends of *E. coli* Tn7 serve to distinguish the TE from the surrounding sequence. The Left end is 150 bp long and contains 3 TnsB binding sites, each 22 bp long. The Right end is shorter at 90 bp, but contains four TnsB binding sites, and is the end closest to *tnsA* (see Figure 3)[33,43,32]. When transposing through the TnsD pathway, the Tn*7* transposon will always integrate with the Right end proximal to *glmS* end[37].

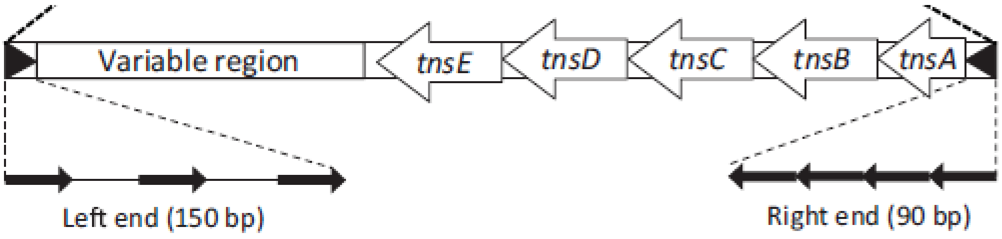


Figure 3: Distribution of TnsB binding sites in Tn7 ends

The bounds of the Tn7 are determined by the Left and Right end sequences. The Right end is always closer to *tnsA*. The Right end contains four 22-bp TnsB binding sites within its 90-bp length. The Left end contains only three, spread across its 150-bp length. Figure adapted from Craig, N. L. *et al. Mobile DNA III*[8].

TnsA

TnsA is the endonuclease responsible for cleaving the 5'-ends of the TE, at +3 bp just upstream of the Tn*7* ends[44]. TnsA may be specific to Tn*7* and its homologs, as it shows little sequence homology to other proteins outside of the Tn*7* family. A crystal structure of TnsA revealed structural similarity between the N-terminal region of TnsA and the active site of the endonuclease FokI[45]. The C-terminal domain (CTD) of TnsA displays structural similarity to the linker histone H5 and is responsible for interfacing not only with the Tn*7* ends, but also TnsB and TnsC[34,44]. CTD fragments of TnsC (TnsC$^{495\text{-}555}$) bind to TnsA and stimulate nonspecific DNA binding. Furthermore, the TnsC$^{495\text{-}555}$ fragment stimulates cleavage of DNA bound by a TnsAB heterotransposase[44]. TnsA stimulates formation of the PEC when bound to TnsB on Tn*7* ends[46]. Perhaps unsurprisingly, loss-of-function mutations (E63A, D114A, K132A) place the catalytic site of TnsA, thus the endonuclease domain, within the first half of the TnsA sequence[34,45]. Interestingly, a Tn*7* TE with a TnsA active site loss-of-function mutant (TnsA^{D114A}) will completely lose 5' cleavage, forcing the TE through replicative transposition in lieu of its normal transposition pathway[29]. See Figure 4 for noted features of *E. coli* TnsA.

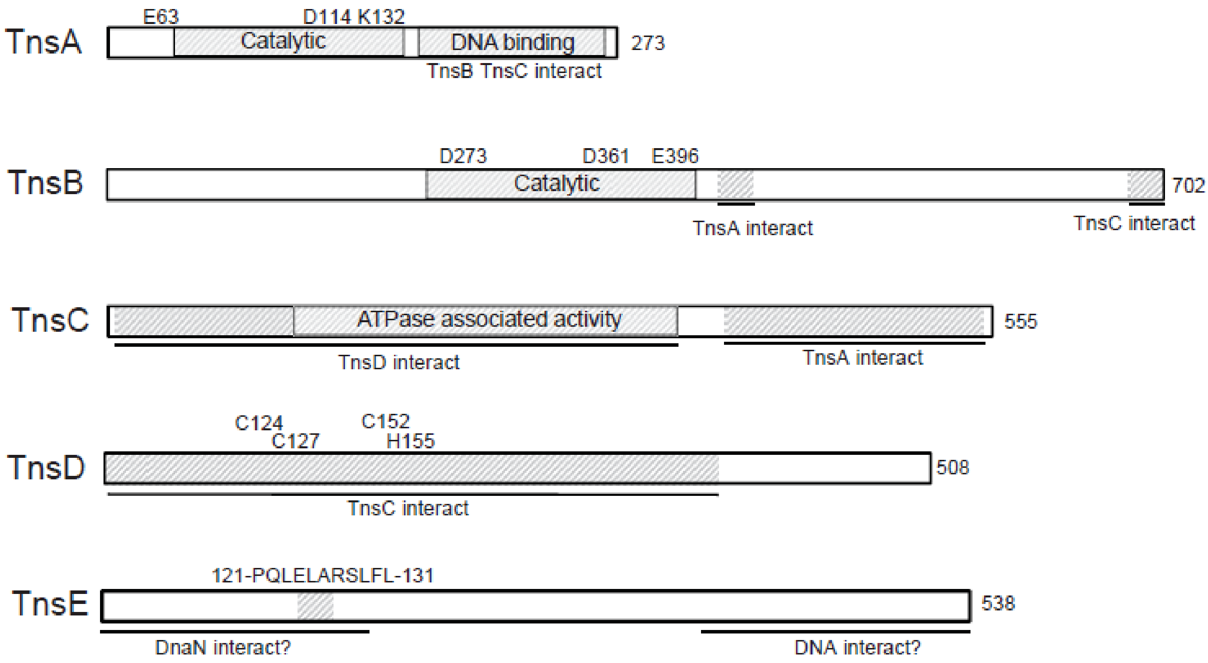


Figure 4: Length and key features of *E. coli* Tn*7* core components

Relevant features mentioned in the text are diagrammed here for *E. coli* Tn*7* proteins. Indicated amino acids comprise a classifying feature of the protein, e.g. the D273, D361, E396 that make up the DDE motif of the TnsB transposase. Figure adapted from Craig, N. L. *et al. Mobile DNA III*[8].

TnsB

As a DDE retroviral integrase, TnsB is the archetypal transposition protein in Tn*7*[8]. The DDE motif (D273, D361, E396) is contained within TnsB residues 266 to 406[47]. As with other DDE integrases, disrupting the motif's interaction with metal ions, whether by mutation or chelation, removes TnsB's ability to catalyze integration[34]. TnsB binds to repeated 22-bp sequences within the Left and Right ends of Tn*7*[33]. Following end binding, TnsB will recruit and complex with TnsA on the ends of Tn*7*, ultimately leading to the formation of the PEC[46]. When bound on the ends of the PEC, TnsB is poised to cleave the 3'-ends of the TE and catalyze the strand-transfer reaction to the target site, presumably upon signaling from TnsC. Truncating TnsB by removing the last 40aa in TnsB's CTD, creating TnsB$^{1\text{-}662}$, will abolish all integration. The removed 40 aa are implicated in TnsC interaction, as while there is no productive integration with the full TnsAB$^{1\text{-}662}$C+D complex, there is also no productive integration with the TnsAB$^{1\text{-}662}$C^{A225V} complex[35]. Additional support for TnsC interaction with TnsB through TnsB's CTD is that specific mutations within the CTD (P686S, V6899M, and P690L) produce a TnsB that fails to establish target immunity. Given that the model of target immunity (shown in Figure 5) is dependent on TnsB-TnsC interactions, TnsB's CTD is implicated in contacting TnsC[36]. See Figure 4 for noted features of *E. coli* TnsB.

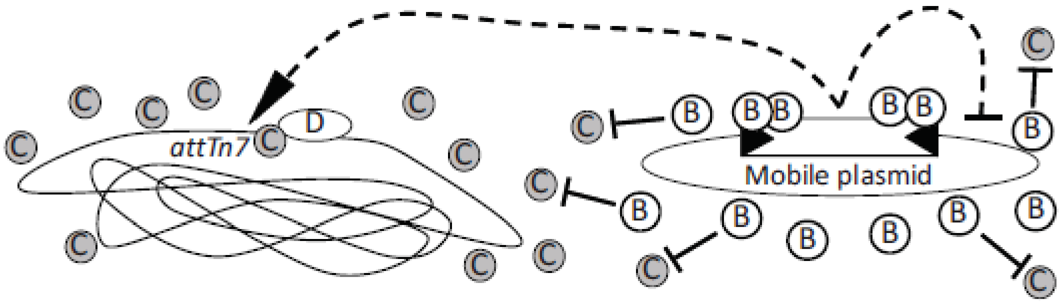


Figure 5: TnsB-mediated target immunity

Target immunity prevents the transposition of Tn*7* into a site that already contains an integrated Tn*7* nearby. TnsB binding at the ends of the integrated element results in an increased local concentration of TnsB. TnsB interaction with TnsC prevents productive complex formation of TnsC with either TnsD or TnsE if they are adjacent to the high concentration of TnsB. Figure adapted from Craig, N. L. *et al. Mobile DNA III*[8].

TnsC

TnsC is a AAA+ ATPase that functions primarily as a mediator between the TnsA/TnsB heteromeric transposase and the target site selectors, TnsD and TnsE [8]. *In vitro* reactions have shown that ATP is essential for Tn*7*-transposition and that it cannot occur in the presence of ADP[37,48]. However, when the same reactions are carried out using a non-hydrolyzable ATP analog, in the absence of TnsD & TnsE, random integration will occur[37]. Use of the TnsC^{A225V} ATPase active site mutant in transposition reactions will allow a TnsA/B/C^{A225V} complex to integrate the TE in a sequence-independent fashion. Strangely, the same mutation does not abolish TnsC's ability to bind to TnsD and TnsE and can still transpose through those targeting pathways[40,47,49]. Concomitant with the ease at which targeting pathways can be circumvented is the theory that the site selectivity of TnsD and TnsE must confer a survival advantage on Tn*7*. The CTD of TnsC (residues 504-555 of 555) interact with TnsA[34,44]. To date, there is no published work indicating what portion of TnsC interacts with TnsB.

Analysis of target immunity indicates a relationship between TnsB and TnsC. Currently, target immunity is thought to emerge from the increased local concentration of TnsB around Tn*7* ends, granting primacy to TnsC-TnsB interactions over DNA-bound TnsD/TnsE interactions with TnsC. As ATP is required not only for productive transposition, but for TnsC binding to DNA, it is likely that TnsB forces TnsC to hydrolyze any bound ATP, preventing transposition[40,42,50]. This model is further supported by the findings that target immunity can be bypassed with the use of non-hydrolyzable ATP analogs or the use of the TnsC mutant TnsC$^{S401Y\Delta402}$, which ablates the requirement for ATP and will allow productive transposition in the presence of ATP or ADP. While the TnsC$^{S401Y\Delta402}$ mutant bypasses target immunity, it also foregoes any site selectivity, integrating regardless of DNA sequence through TnsC-DNA interactions[40,42]. Furthermore, mutations in TnsC that abolish TnsB interaction (and vice-versa) create additional target immunity bypass Tn*7* variants[49]. Most oddly, the TnsC$^{S401Y\Delta402}$ mutation does not appear to affect ATP hydrolysis activity and is thought to instead force a permanent conformational change in TnsC that would otherwise occur upon TnsD/TnsE binding and subsequent ATP hydrolysis[40]. While TnsC recruitment to TnsD-*attTn7* (*attTn7* is the DNA sequence within *glmS* recognized by TnsD) is dependent upon the topological changes of the local DNA enforced by TnsD, it appears that simultaneous expression of TnsC and TnsD will increase recruitment of TnsD to *attTn7*[22]. Lastly, while in the TnsC-TnsD-*attTn7* complex, TnsC binds the minor groove of the DNA in between *attTn7* and the Tn*7* insertion site[22,39]. Footprinting assays suggest that it is TnsC that is responsible for the 25-bp separation of *attTn7* and the downstream insertion site[43]. See Figure 4 for noted features of *E. coli* TnsC.

TnsD

TnsD is one of two target-site selecting proteins in the Tn*7* transposon. TnsD recognizes the DNA sequence at *attTn7* and coordinates integration 25-bp downstream of the recognized site, just outside of the coding region of *glmS*. A zinc finger domain of CCCH can be found at positions C^{124}, C^{127}, C^{152}, and H^{155}. While mutations in these residues ablate recognition of the *attTn7* sequence, they do not prevent DNA binding, suggesting TnsD harbors additional DNA-binding motifs[22]. TnsD binding to *attTn7* requires the essential host factors ACP (acyl carrier protein) and L29 (large ribosomal subunit 29)[51]. Binding of TnsD to *attTn7* results in a distortion in the DNA that is required for TnsC recruitment[22]. Beyond the target site distortion, TnsD and TnsC interact through their N-terminal domains, as deletion of either abrogates *attTn7*:TnsD:TnsC complex formation[39]. See Figure 4 for noted features of *E. coli* TnsD.

TnsE

TnsE stimulates Tn*7* transposition into lagging-strand DNA during replication, specifically recognizing the 3'-recessed ends of DNA in the lagging strand. Mutations in the C-terminus of TnsE will generate hyperactive 3'-recessed end DNA-binding mutants, implicating the C-terminus of TnsE in DNA binding. However, initial targeting to replicating DNA likely occurs through a conserved sliding clamp interacting sequence from residues 121-131 that moderates the interaction of TnsE with DnaN. Electrophoretic Mobility Shift Assays (EMSAs) of TnsE incubated with DNA containing a 3' recessed end yield both shifted and supershifted species, indicating a potential for TnsE multimerization[24,52]. See Figure 4 for noted features of *E. coli* TnsE.

TnsD-mediated transposition

As mentioned before, Tn*7* may transpose through two different pathways. One is the TnsD driven pathway, in which the target-site binding protein TnsD binds to the *attTn7* TnsD attachment site at the end of the coding region of the highly conserved *glmS* gene (see Figure 6 and Figure 7)[38]. The binding of TnsD to this sequence is stimulated by the host factors ACP and L29[51]. Binding of TnsD to *attTn7* distorts the DNA topology at the distal end of *glmS* [22]. This distortion, in concert with the N-terminus of TnsD, recruits and binds TnsC through TnsC's N-terminal domain[39]. Simultaneously, the integrase TnsB has bound the TnsB binding sites located in the Right- and Left-ends of the Tn*7* TE[33]. TnsB binding to the TE ends recruits the endonuclease TnsA to assemble the heteromeric transposase TnsA/TnsB on the ends of the TE. TnsA binding to TnsB effects a topological change of the TE that results in the formation of the toroidal shape of the PEC[46]. With the PEC in proximity of the target-bound TnsD and TnsC, the C-terminal domain of TnsC binds to the C-terminal domain of TnsA, while undergoing a conformational change that allows the catalytic core of the TnsBs bound to each end of the PEC to access the downstream integration site with a 5-bp offset. TnsA cleaves the 5'-strands of the TE and TnsB performs a 5-bp staggered strand-transfer reaction with the 3'-ends of the TE post 3'-end cleavage[44]. At this point the TE has been completely excised from the donor DNA and has left two free DNA ends behind. The now integrated Tn*7* is flanked by two 5-bp ssDNA segments connecting it to the rest of the flanking dsDNA sequence. These ssDNA segments are a consequence of the transesterification reaction chemistry performed by TnsB. They will be filled in by a host polymerase forming the TSDs[1,8]. Perhaps as a consequence of the asymmetric distribution of TnsB binding sites within the ends of the TE, the transposon has integrated with the Right end proximal to the TnsD binding site[37]. With the ends of the Tn*7* TE now adjacent to the *attTn7*, the local concentration of TnsB is sufficiently high enough to discourage productive TnsC interaction with DNA or TnsD, resulting in the observed target immunity[40,42,50].

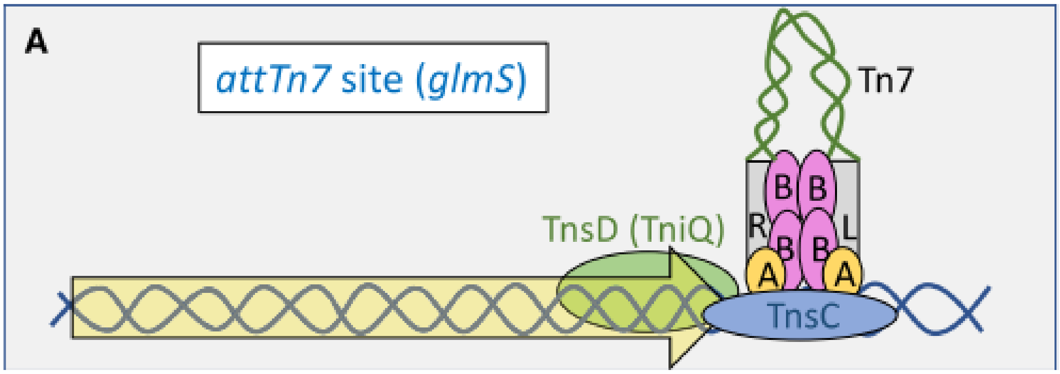


Figure 6: Tn7 transposition at *attTn7*

An ideogram of the full complex of the Tn*7* transpososome at the moment before integration just downstream of *attTn7*. The PEC is formed by the transposing Tn*7* element, multiple copies of TnsB, and multiple copies of TnsA bound to TnsD through the coordinating transposition protein, TnsC. Exact protein structures and interacting residues are currently unknown. Figure adapted from Peters, J. E. *Mol. Microbiol.* **112**, 1635–1644 (2019)[38].

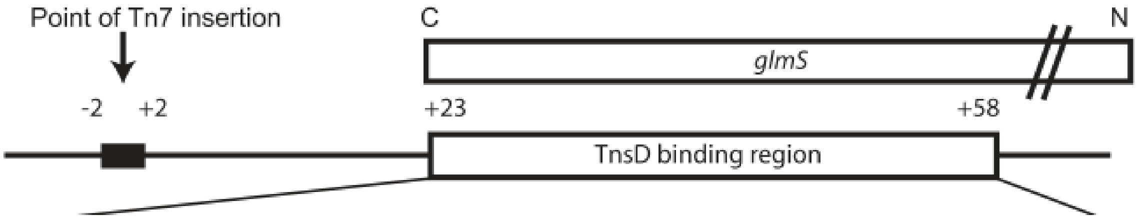


Figure 7: Detail of TnsD – *glmS* interaction at *attTn7*

TnsD contains a DNA-binding Domain of exquisite specificity. Binding to the nigh-terminal 25 bases of *glmS*, TnsD directs integration immediately downstream of the *attTn7* attachment site. The integration site is indicated, with the 5 bases that will create the TSD between the noted -2 and +2. Figure adapted from Mitra, R., *et al. Mob. DNA* **1**, 1–14 (2010)[22].

TnsE-mediated transposition

Other than the target site and the target-site binding protein, much of the pathway of TnsE-mediated transposition is similar to the TnsD-mediated pathway. The substrate for TnsE binding is a 3'-recessed end of DNA (most often found in the lagging strand during DNA replication) and the sliding clamp processivity factor DnaN, likely through TnsE's own N-terminal domain (Figure 8). Interestingly, TnsE preferentially targets conjugal plasmids, and may target replicating bacterial chromosomes, and even the filamentous phage M13[24,52]. The TnsE-mediated pathway of Tn7 transposition is sensitive to target immunity, likely due to the fact that only the target site selection has changed, the core TnsA, TnsB, and TnsC machinery is unaltered from the TnsD pathway[53].

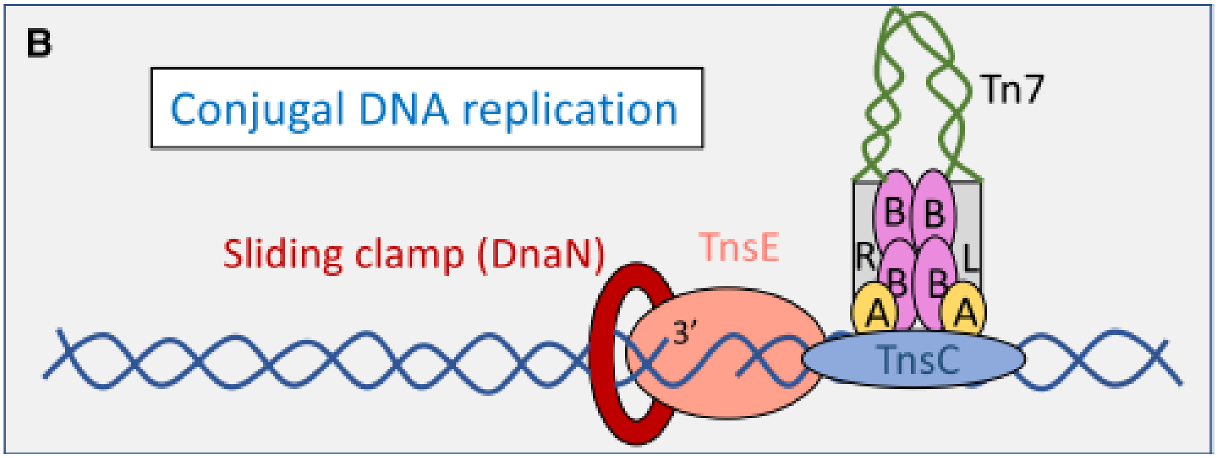


Figure 8: Tn7 transposition into a lagging strand during DNA replication

An ideogram of the full complex of the Tn*7* transpososome at the moment before integration into a growing lagging strand. TnsE is recruited to the replicating DNA through its interaction with DnaN and the recessed 3′ DNA ends. As in the TnsD pathway, TnsC coordinates integration with the PEC+TnsA/B. Exact protein structures and interacting residues are currently unknown. Figure adapted from Peters, J. E. *Mol. Microbiol.* **112**, 1635–1644 (2019)[38].

1.5: Genetic engineering with transposable elements

Due to their precise integration chemistry, transposons present exciting targets for tool development. The ability to insert a genetic payload with minimal scarring and requiring few to no host factors will always be useful, whether for use in a research lab or in a clinical setting. The existence of transposons with extreme sequence specificity requirements has emboldened scientists to attempt altering transposon insertion site sequence specificity, with the holy grail resembling an easily programmable system akin to the CRISPR-Cas systems used to generate knockouts[54,23,55,56,57]. Currently, the most widely used transposons for genetic engineering have strict, but common, target site sequence requirements, such as Sleeping Beauty's -TA- site specificity[23]. Despite the frequency of potential integration sites, transposon-based gene insertion is being used in clinical trials, notably for *ex vivo* integration of Chimeric Antigen Receptors in T-cells for immunotherapeutic approaches to malignancies[58]. There are more well-established methods for genetic payload insertion, such as lentiviral methods. However, transposons have some key advantages, notably they are less immunogenic and easier to produce than lentivirus-based methods, and as such, still hold great appeal for many scientists[59].

Chapter 2: CRISPR-Cas Biology

2.1: CRISPR-Cas systems in their native context

Phages and mobile genetic elements (MGEs) have besieged prokaryotes since their first encounter. Phages alone are estimated to lyse 30% of all bacteria daily[60]. This unending battle has forced an arms race with all members evolving new mechanisms of defense and counter defense. It is perhaps one of the oldest theories of biology, that pressure yields diversification. While certainly not speaking about phages and prokaryotes, Charles Darwin once (presciently) wrote "One may say there is a force like a hundred thousand wedges trying to force every kind of adapted structure into the gaps in the economy of nature, or rather forming gaps by thrusting out weaker ones"[61]. The diversity this arms race has yielded is a treasure trove of biological mechanisms that the scientific community has only begun to appreciate the scope of. While a sample of these systems are less obviously applicable and thus have been studied only for their own sake[62], others such as restriction enzymes from restriction modification systems, have been utilized by scientists daily for decades[63]. One such system has been thrust to the fore only recently but is poised to make an enormous contribution to the biological sciences and tangential fields: the CRISPR-Cas systems[64].

Clustered Regularly Interspaced Palindromic Repeats (CRISPR) and their CRISPR-associated (Cas) genes describe a family of adaptive immune systems in prokaryotes and archaea. There are myriad distinct mechanisms by which CRISPR-Cas systems effect their response to invading nucleic acids, but common to all are the eponymous CRISPR RNAs (crRNAs) derived from a genomic feature named the CRISPR array. Regardless of which Cas-protein or Cas-protein complex binds its cognate crRNA, the crRNA is responsible for specifying at least the initial target of the CRISPR-Cas complex[65]. Compared with rationally engineering protein structure to alter protein binding specificity, the relative ease with which crRNAs may be altered has given modern scientists unprecedented levels of control over their model organisms.

Broadly, CRISPR-Cas systems mount their defense against invasive nucleic acid species in three stages, as shown in Figure 9.

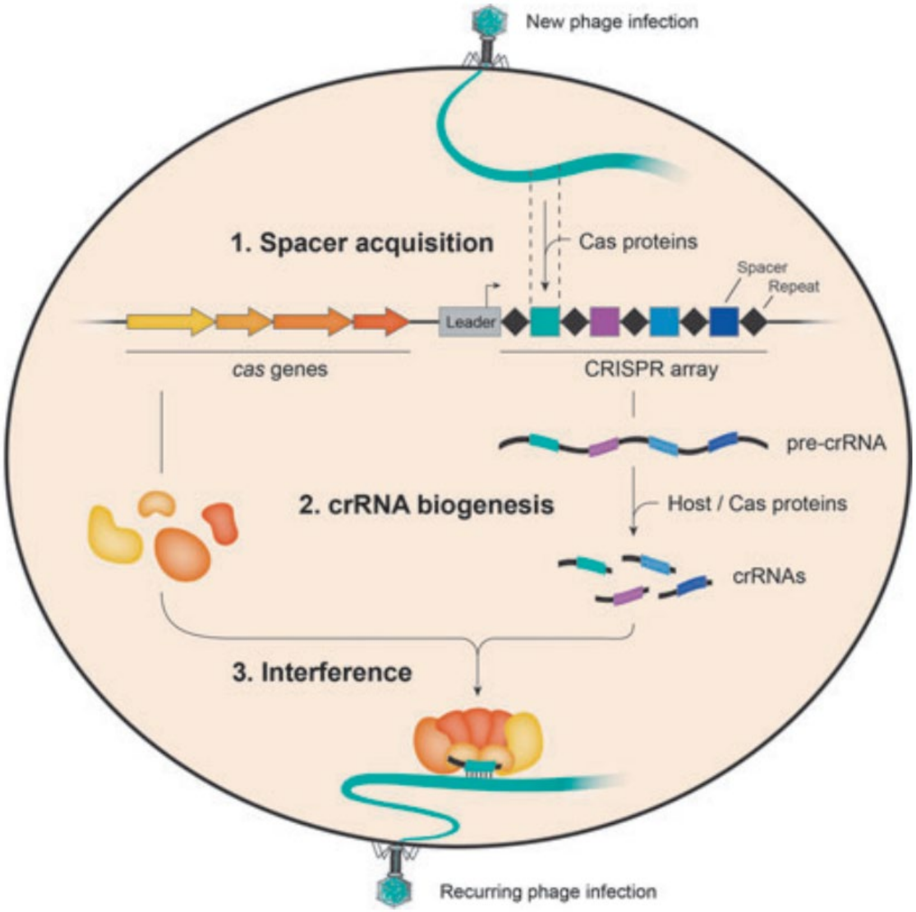


Figure 9: Stages of CRISPR-Cas defense

Host defense by a CRISPR-Cas system ensues as follows: 1, an invading phage (or other invasive nucleic acid) is processed by adaptation machinery and has a protospacer removed and incorporated into the bacterium's CRISPR array. 2, production of Cas effector proteins and pre-crRNA, maturation of the pre-crRNA into discrete crRNAs. 3, assembly of the interference module and protospacer binding based on crRNA spacer sequence, target cleavage follows shortly after. Figure adapted from Klompe, S. E. & Sternberg, S. H. *Cris. J.* **1**, 141–158 (2018)[66].

In the first stage, spacer acquisition, short sequences of an invading nucleic acid are processed by distinct Cas proteins and incorporated into the host's CRISPR array. In the second stage, expression and processing, the Cas gene expression may be upregulated and the CRISPR array is transcribed, creating the pre-crRNA. Processing of the pre-crRNA into mature crRNAs then allows complexing of the expressed Cas proteins with a single mature crRNA per complex. In the third stage, interference, Cas-crRNA Ribonucleoprotein (RNP) complexes interrogate their target nucleic acids; guided by their crRNA, the complexes bind a complementary sequence, and degrade, or signal for the degradation of, the bound nucleic acid[66]. Target search occurs in three dimensions and with the RNP complex interrogating the bound DNA for Protospacer Adjacent Motifs (PAMs)[67]. PAMs serve as a self-other identifier to prevent binding and cleavage of the spacer within the CRISPR array from whence the crRNA came[68]. Should the RNP encounter a PAM, the crRNA spacer will begin binding the protospacer (site with complementarity to the spacer) through canonical nucleic acid base pairing. Mismatches between the crRNA spacer and the target DNA are poorly tolerated in the PAM-proximal region of the crRNA, termed the seed sequence[67,69]. The end of the seed sequence is set at the nucleotide beyond which mismatches will not entirely ablate spacer-protospacer binding[70]. Complete binding of the RNP to the protospacer toggles a conformational change resulting in cleavage of the bound target, or recruitment of a nuclease, and subsequent cleavage or degradation of the RNP-bound DNA[67,71,72].

2.2: Two CRISPR-Cas classes and mechanistic differences

As a product of an ongoing arms race, CRISPR-Cas systems are staggeringly diverse in both their form and function[65,73]. To accommodate this ever-expanding diversity, CRISPR-Cas systems are classified primarily by their evolutionary relationships; focusing on Cas gene operonic composition, the architecture of CRISPR-Cas loci, sequence similarity, the phylogeny of more conserved Cas genes, and more recently, by mutual association with non-Cas genes[74,65]. The greatest distinction is between the Class 1 and Class 2 systems. Class 1 systems utilize multiple Cas proteins in complex as the interference module. Class 2 systems use a single effector protein for interference[66]. Although the Class 2 Type II Cas9 containing systems are most broadly recognized, it is the Class 1 Type I-F systems, and the Cas12-containing Class 2 Type V systems are most relevant to this thesis[9,65,66,75].

The most diverse of the Class I systems, Class I Type I systems, are typified by their use of a multiprotein complex built around a mature crRNA responsible for target site binding and R-loop formation[76]. Based on the previously mentioned classification parameters, Type I systems are further segregated into Type I-A through I-G, with the I-F subclass divided into I-F1, I-F2, and I-F3[65]. With the exception of Type I-D and Type I-F3 systems, all Type I systems use the nuclease Cas3, or a homolog, for target degradation[65,77]. Most Type I systems utilize Cas6 to mature the pre-crRNA by cleaving stem-loop structures formed within the repeats of the pre-crRNA[78,79]. The cleavage site within the repeats leaves a portion of each repeat flanking the spacer region; these partial repeats are referred to as the 5'- and 3'-handles of the crRNA[80]. The Cas6 of some Type I systems remains bound to the stem-loop of the 3'-handle post-maturation[79,81,78]. At the other end of the crRNA, Cas5 binds the 5'-handle, and is positioned adjacent to Cas8, the large subunit responsible for PAM recognition[82,83]. Across the spacer, 6 copies of Cas7 form a helical spine. Some Type I systems utilize the small subunit, Cas11, to stabilize the crRNA-DNA interaction. This assembled RNP complex is referred to as Cascade, or the CRISPR Associated Complex for Antiviral Defense[84–86]. Cascade can only bind the target DNA and possesses no inherent nuclease activity[87]. A notable feature of Type I systems is the Cas7-crRNA interaction that "flips" every sixth base in the crRNA away from potential protospacer binding[88]. Consequently, Cascade binding is tolerant of mismatches between the spacer region of the crRNA and the protospacer DNA at every 6th position[89].

In 2015, Makarova *et al* defined a putative Type V CRISPR-Cas system containing the interference module *Cpf1* (later reclassified as Cas12a), a CRISPR array, along with the adaptation module expressed from *Cas1*, *Cas2*, and *Cas4*[74]. Initial work on Cas12a established that it is an RNA-guided dsDNA endonuclease with a RuvC domain, a Nuc domain, and a ribonuclease domain[71]. Cas12a catalyzes the maturation of its own pre-crRNA to crRNA via said ribonuclease domain[90]. Unsurprisingly, Cas12a recognition and processing of the pre-crRNA requires full length repeats that maintain their stem-loop structure. However, so long as mutations in the repeat do not alter the secondary structure, Cas12a is still able to process the repeat[90]. Early work on Cas12a homologs established a 5'-TTN-3' or 5'-TTTN-3' PAM immediately upstream of the protospacer[71]. Cas12a tolerates truncated crRNAs as well as single mismatches, even when directly adjacent to the PAM[90,91]. Upon binding of a complementary protospacer, crRNA guided Cas12a initially cleaves the displaced strand in the RuvC domain, then further unwinds the target dsDNA and cleaves the crRNA-bound DNA strand an additional 5-7bp PAM-distal[92]. Although the cuts are staggered, they are in close enough proximity that a double strand break will follow successful cleavage of both the non-target and target strands of DNA[71]. By 2020, the Type V clade had drastically expanded with novel variants, displaying a diversity of CRISPR-Cas systems with varying requirements for trans-activating CRISPR-RNAs (tracrRNA, a structural RNA with partial complementarity to the crRNA, contributes to crRNA maturation), different genomic organization, and differing amounts of Cas genes. As a result of these differences, Type V systems are currently the most diverged of the Class 2 systems[65,69]. Some subgroups completely lack adaptation modules or contain a nuclease-deficient Cas12 homolog[65,93]. Currently, there are four other major types of CRISPR-Cas systems, but they fall beyond the scope of this thesis.[65]

2.3: Class I Type I-F mechanism of target degradation

Class I Type I-F systems differ from the rest of the Type I family, as they lack the small subunit Cas11 and contain a fusion Cas2/Cas3 helicase/nuclease[65,72]. The Cas2/Cas3 fusion is responsible for both spacer acquisition and degradation of DNA at the Cascade-bound target site[94]. Early work refers to the Type I-F Cascade as Csy complex, or CRISPR subtype *Yersinia pestis*[95]. Prior to recruitment by Cascade, the Cas2/Cas3 protein homodimerizes and complexes with four molecules of the Cas1 integrase, forming a Cas2/Cas3$_2$Cas1$_4$ multimer. The bound Cas1 inhibits Cas2/Cas3 activity until recruited by target-bound Cascade[94]. While they all lack Cas 11, Cascade composition differs between the three subtypes of I-F systems. The Cascade of Type I-F1 systems most closely resemble the aforementioned Type I Cascade, just lacking Cas11. The Type I-F2 systems lack Cas11 and Cas8, with PAM recognition now executed by Cas5. The Type I-F3 systems are the most reduced, lacking Cas11, and with a Cas8-Cas5 fusion protein responsible for both PAM-recognition and binding of the 5'-handle of the crRNA.

Regardless of subtype, Type I-F Cascades form a characteristic seahorse morphology consistent with other Type I Cascade complexes (Figure 10).

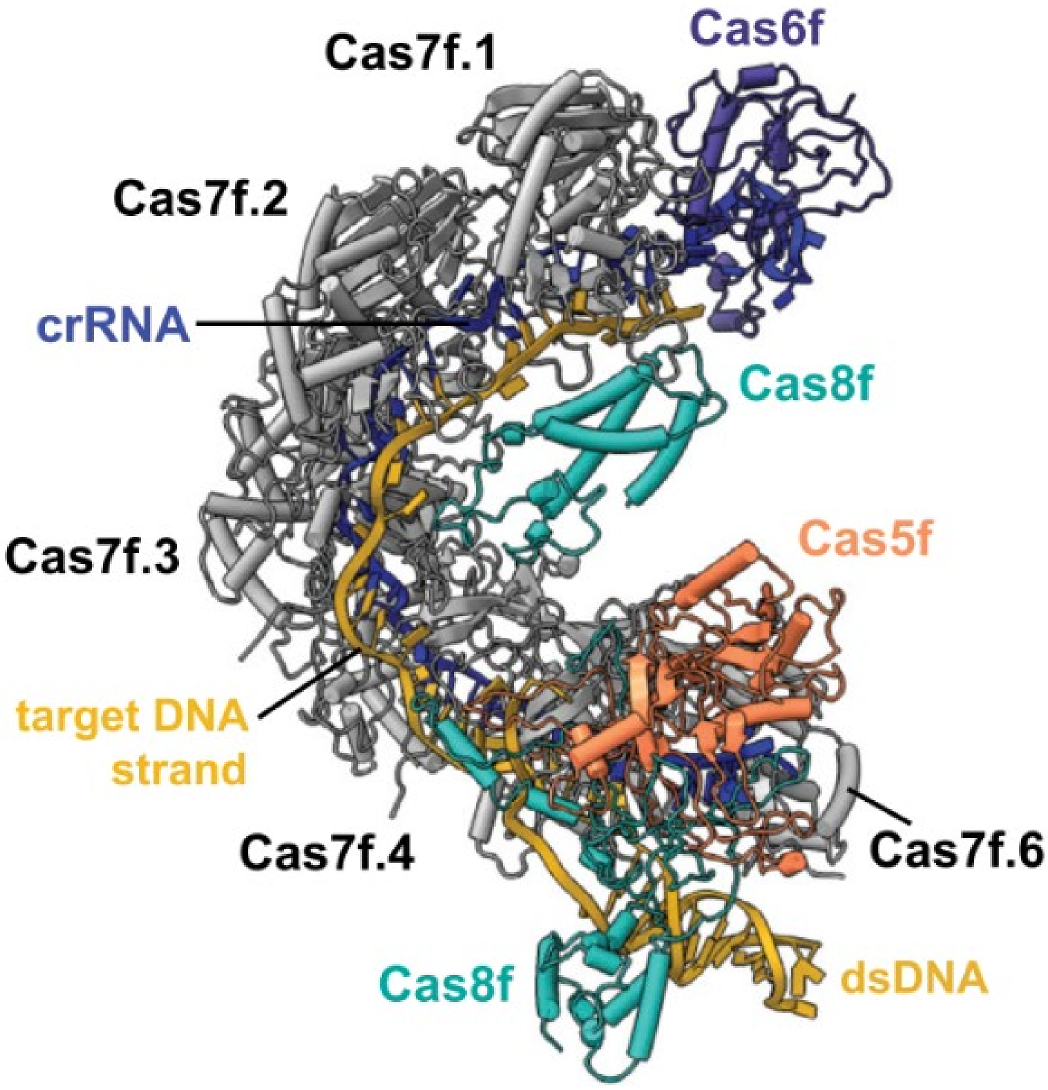


Figure 10: Atomic model of a Type I-F Cascade (Csy complex)

A 3.2 Å cryo EM model of a Type I-F Cascade from *P. aeruginosa* bound to an 80-bp dsDNA target. Identity and location of subunits are indicated in the figure. Figure adapted from Rollins, M. C. F. *et al. Mol. Cell* **74**, 132-142.e5 (2019)[72].

A curious feature of Type I-F Cascade complexes is that the number of Cas7 subunits can be regulated by the length of the crRNA, the addition or removal of sets of six bases within the spacer region of the crRNA will allow a commensurate gain or loss of additional Cas7 subunits[96,97]. Complete binding of the Cascade complex to a fully complementary target site triggers a conformational change of the DNA bound complex: The gross morphology differs little between the DNA-bound and unbound Cascade, but the bound complex has undergone three notable conformational changes. First, with Cas8 and Cas7.6, Cascade has "clamped" the dsDNA at the PAM. Second, Cas5 has rotated away from the Cas6 end of the complex, elongating the Cas7 spine, and thus the entire complex. Third, a helical bundle in Cas8 rotates 180° so that it now contacts Cas 7.2 and Cas7.3. The rotation of Cas8 is dependent on R-loop formation and is the trigger to recruit and dissociate the Cas2/$Cas3_2Cas1_4$ multimer; freeing Cas2/Cas3 from inhibition by Cas1 (Figure 11)[72]. Liberated from Cas1, Cas2/Cas3 degrades both the bound and displaced strands of DNA at the Cascade-bound target. The displaced strand is degraded at a measurably quicker rate[94].

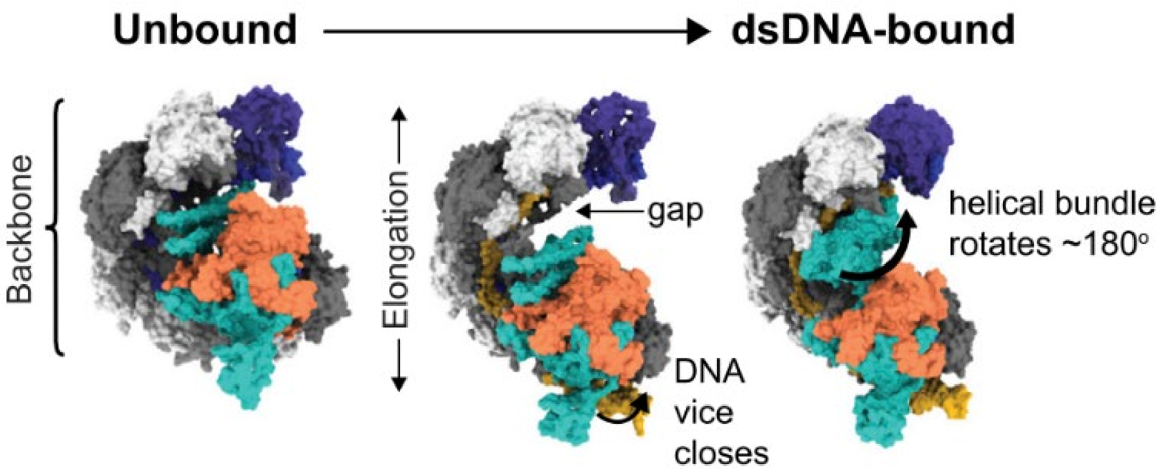


Figure 11: Conformational change of I-F Cascade as it binds a dsDNA target

Representation of the conformational change I-F Cascade undergoes when bound to a dsDNA target with complete R-loop formation. Left, unbound complex. Middle, bound to a partially duplexed dsDNA. Right, dsDNA bound complex with full R-loop formation, poised for Cas2/Cas3 recruitment. Figure adapted Rollins, M. C. F. *et al. Mol. Cell* **74**, 132-142.e5 (2019)[72].

2.4: Heterologous genetic engineering with Type I-F or Type V CRISPR-Cas systems

As with most CRISPR-Cas systems, there is extensive interest in the development of Type I and Type V systems as tools in heterologous systems, most notably, mammalian cells. Three reports from 2019 establish the use of Type I-E systems in mammalian cells. Dolan *et al* describe long range deletions of up to 20kb of genomic DNA in human embryonic stem cells[98]. Cameron *et al* fused a non-sequence-specific FokI nuclease domain to Cas8 and assembled paired FokI-Cascade complexes for precise cleavage between the two protospacers[99]. Lastly, Pickar-Oliver *et al* fused transcriptional activators to Cascade components and noted specific upregulation of transcripts from genes immediately downstream of the crRNA target[100]. This strategy, initially developed with a catalytically inactive Cas9 (dCas9) fused to transcriptional activators is named CRISPR activation, or CRISPRa[101]. Yang *et al* followed Pickar-Oliver, but instead of a I-E system, repurposed a I-F system by fusing transcriptional activators to Cas7, resulting in 6 copies of the transcriptional activator, VPR, per Cascade to effect CRISPRa where targeted[102]. To date, no I-F system has been used for DNA cleavage in mammalian cells.

In the very first paper demonstrating Cas12 (then Cpf1) function, Zetsche *et al* screened Cas12 orthologs for activity in human cells, with two candidates showing clear editing[90]. A later report from the same lab mutated Cas12b orthologs to increase the efficiency of their dsDNA cleavage at temperatures lower than their optimal reaction temperatures[103]. More recently, Kleinstiver et al subjected a Cas12a to rampant, albeit rational, mutation to yield a novel Cas12a nuclease with reduced off-target effects and increased multiplex gene editing efficacy[91].

2.5: Transposon-encoded CRISPR-Cas systems

Within prokaryotic genomes, CRISPR-Cas systems and other defense genes tend to cluster in defense islands, either on or near MGEs[104]. As Transposable Elements carry defense genes, and are themselves MGEs, these defense islands provide plausible intersections of CRISPR-Cas systems and TEs. Although CRISPR-Cas systems have broad-ranging mechanisms, common to all is their programmability and specificity, suggesting this shared characteristic provides a fitness advantage. Given that: 1, the proximity of CRISPR-Cas systems and TEs within the highly mutable defense islands; 2, that CRISPR-Cas systems display extreme target specificity; and 3, that some TEs autoregulate their transposition through severe site-specificity of integration, it seemed plausible that CRISPR-Cas systems and TEs may have previously co-opted functional modules from one another[104]. To this end, the identification of persistent partial CRISPR-Cas systems suggested they may be put to another use. The identification of such a partial system, a minimal I-F system (now called I-F3), lacking Cas1, the Cas2/Cas3 fusion, and containing a fusion of Cas8 and Cas5 initiated an *in silico* screen of minimal I-F3 systems and their surrounding loci[74]. Construction of phylogenies based on I-F3 Cas7 homologs and the local gene neighborhood yielded a monophyletic branch with the minimal I-F3 system and homologs of the Tn7 transposition proteins TnsA, and TnsD or its homolog, TniQ[32]. Similar dendrograms created with either a TnsA seed or a TniQ/TnsD seed both produced clades with the minimal I-F3 systems (Figure 12)[32]. These phylogenetic analyses demonstrate an unusually common co-occurrence of a minimal Type I-F3 system with Tn*7* elements. Subsequent analysis of these clades showed no intact minimal I-F3 systems beyond the +/-10kb sampled from the Tn*7* elements, indicating that the minimal I-F3 systems may only be functional when proximal to Tn*7* elements[32]. Sequences that contained intact minimal I-F3 systems near Tn*7* elements were further analyzed to determine whether the I-F3 system was contained between the Left and Right ends of the Tn*7* elements[32]. Analysis of the spacers within the minimal I-F3 CRISPR array mapped mostly to plasmids and phages associated with the host genera[32]. But, in two cases, the spacers mapped to a sequence directly upstream of the Right end of the Tn7-like transposon, suggesting these spacers may have been used to direct transposition[32]. Taken in concert, these data pointed towards a model of I-F3 CRISPR-Cas-mediated transposition of Tn*7*-like transposons.

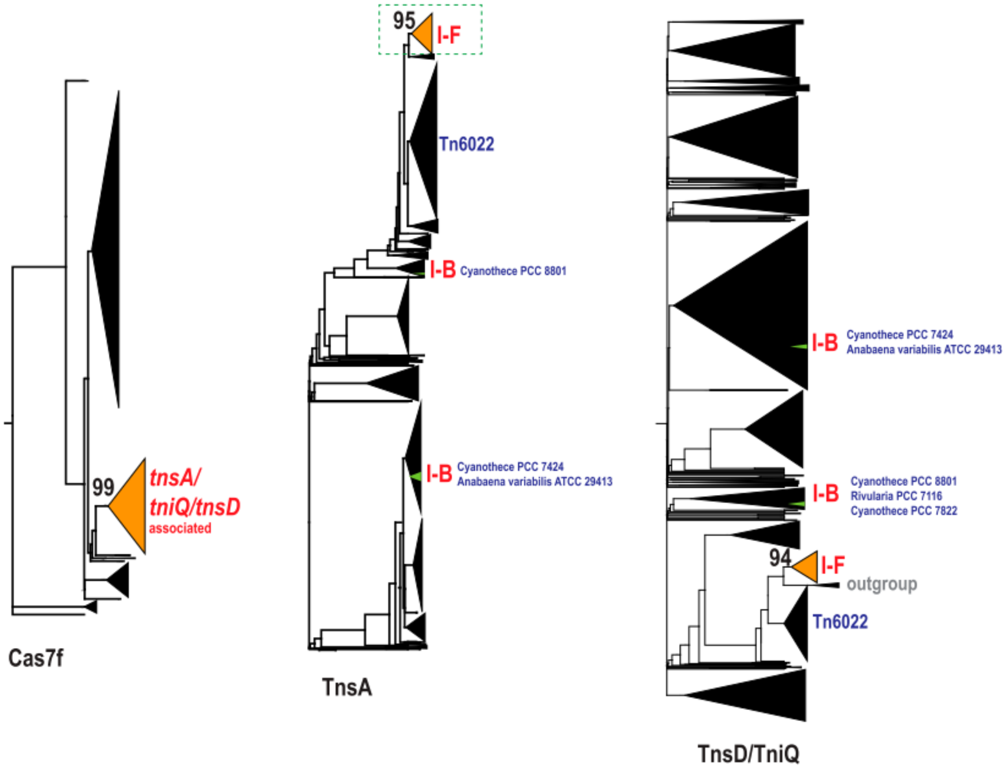


Figure 12: Dendrograms constructed from I-F Cas7, TnsA, and TnsD/TniQ protein families

The gene indicated below each dendrogram was used as the seed of a PSI BLAST. After obtaining a number of orthologs, the sequences were mapped back onto their respective genomes and a window of 20 kb was centered on the seed, genes within that window were annotated, and used to construct the above phylogenetic trees. Left, sampling of 2,905 minimal I F Cas7 proteins from a PSI BLAST and their neighboring genes produced the large orange triangular group labeled *tnsA*/*tniQ*/*tnsD*, indicating minimal I F systems are often proximal to Tn7 family genes. The same process was repeated with TnsA (middle, 7,203 variants) and Tnsd/TniQ (right, 7,963 variants). Figure adapted from Peters, J. E *et al*, *Proc. Natl. Acad. Sci.* **114**, E7358–E7366 (2017)[32].

Shortly after the previous findings regarding the rate of co-occurrence of minimal I-F3 systems and Tn*7* elements were released, a paper doing the same for Type V-K systems followed[105]. The Class 2 Type V-K system appeared to contain Cas12 homologs with an inactivated RuvC domain. Similar to the aforementioned minimal Type I-F3 system, this V-K family was predicted to be unable to carry out any nuclease function and would therefore lack any defense capability[106,107].

Chapter 3: Transposon-Encoded CRISPR-Cas Systems Direct RNA-Guided DNA Integration

This chapter has been adapted from:

Klompe, S.E., Vo, P.L.H., Halpin-Healy, T.S. & Sternberg, S. H. Transposon-encoded CRISPR–Cas systems direct RNA-guided DNA integration. *Nature* 571, 219–225 (2019).

Appendix A contains the paper as published.

Contributions:

S.E.K. and S.H.S. conceived of and designed the project. S.E.K. performed most transposition experiments, generated NGS libraries, and analyzed the data. P.L.H.V. helped with cloning and transposition experiments, and performed computational analyses. T.S.H.-H. performed biochemical experiments. S.H.S., S.E.K. and all other authors discussed the data and wrote the manuscript.

P.L.H.V. & T.S.H.-H. contributed equally.

3.1: Abstract

Conventional CRISPR–Cas systems maintain genomic integrity by leveraging guide RNAs for the nuclease-dependent degradation of mobile genetic elements, including plasmids and viruses. Here we describe a notable inversion of this paradigm, in which bacterial *Tn7*-like transposons have co-opted nuclease-deficient CRISPR–Cas systems to catalyze RNA-guided integration of mobile genetic elements into the genome. Programmable transposition of *Vibrio cholerae* Tn*6677* in *Escherichia coli* requires CRISPR and transposon-associated molecular machineries, including a co-complex between the DNA-targeting complex Cascade and the transposition protein TniQ. Integration of donor DNA occurs in one of two possible orientations at a fixed distance downstream of target DNA sequences, and can accommodate variable length genetic payloads. Deep-sequencing experiments reveal highly specific, genome-wide DNA insertion across dozens of unique target sites. This discovery of a fully programmable, RNA-guided integrase lays the foundation for genomic manipulations that obviate the requirements for double-strand breaks and homology-directed repair.

3.2: Introduction

Horizontal gene transfer, a process that allows genetic information to be transmitted between phylogenetically unrelated species, is a major driver of genome evolution across the three domains of life[7,108,109]. Mobile genetic elements that facilitate horizontal gene transfer are especially pervasive in bacteria and archaea, in which viruses, plasmids and transposons constitute the vast prokaryotic mobilome[110]. In response to the ceaseless assault of genetic parasites, bacteria have evolved numerous innate and adaptive defense strategies for protection, including RNA-guided immune systems encoded by clustered regularly interspaced short palindromic repeats

(CRISPR) and CRISPR-associated (Cas) genes[64,111,112]. Remarkably, the evolution of CRISPR–Cas is intimately linked to the large reservoir of genes provided by mobile genetic elements, with core enzymatic machineries involved in both new spacer acquisition (Cas1) and RNA-guided DNA targeting (Cas9 and Cas12) derived from transposable elements[113–118]. These examples support a 'guns-for-hire' model, in which the rampant shuffling of genes between offensive and defensive roles results from the perennial arms race between bacteria and mobile genetic elements.

We set out to uncover examples of functional associations between defense systems and mobile genetic elements. In this regard, we were inspired by a recent report that described a class of bacterial *Tn7*-like transposons encoding evolutionarily linked CRISPR–Cas systems and proposed a functional relationship between RNA-guided DNA targeting and transposition[32]. The well-studied *E. coli Tn7* transposon is unique in that it mobilizes via two mutually exclusive pathways—one that involves non-sequence-specific integration into the lagging-strand template during replication, and a second that involves site-specific integration downstream of a conserved genomic sequence[119]. Notably, those *Tn7*-like transposons that specifically associate with CRISPR–Cas systems lack a key gene involved in DNA targeting, and the CRISPR–Cas systems that they encode lack a key gene involved in DNA degradation. We therefore hypothesized that transposon-encoded CRISPR–Cas systems have been repurposed for a role other than adaptive immunity, in which RNA-guided DNA targeting is leveraged for a novel mode of transposon mobilization.

Here we demonstrate that a CRISPR–Cas effector complex from *V. cholerae* directs an accompanying transposase to integrate DNA downstream of a genomic target site complementary to a guide RNA, representing the discovery of a programmable integrase. Beyond revealing an elegant mechanism by which mobile genetic elements have hijacked RNA-guided DNA targeting for their evolutionary success, our work highlights an opportunity for facile, site-specific DNA insertion without requiring homologous recombination.

3.3: Cascade directs site-specific DNA integration

We set out to develop assays for monitoring transposition from a plasmid-encoded donor into the genome, first using *E. coli Tn7*, a well-studied cut-and-paste DNA transposon[120] (Figure 18a). The *Tn7* transposon contains characteristic left- and right-end sequences and encodes five tns genes, tnsA–tnsE[119], which collectively encode a heteromeric transposase: TnsA and TnsB are catalytic enzymes that excise the transposon donor via coordinated double-strand breaks; TnsB, a member of the retroviral integrase superfamily, catalyzes DNA integration; TnsD and TnsE constitute mutually exclusive targeting factors that specify DNA insertion sites; and TnsC is an ATPase that communicates between TnsAB and TnsD or TnsE. Previous studies have shown that *E. coli* TnsD (EcoTnsD) mediates site-specific *Tn7* transposition into a conserved *Tn7* attachment site (att*Tn7*) downstream of the glmS gene in *E. coli*[121,122], whereas EcoTnsE mediates random transposition into the lagging-strand template during replication[24]. We recapitulated TnsD-mediated transposition by transforming *E. coli* BL21(DE3) cells with pEcoTnsABCD and pEcoDonor, and detecting genomic transposon insertion events by PCR and Sanger sequencing (Figure 18).

To test the hypothesis that CRISPR-associated targeting complexes direct transposons to genomic sites complementary to a guide RNA (Figure 18a), we selected a representative transposon from *V. cholerae* strain HE-45, Tn*6677*, which encodes a variant type I-F CRISPR–Cas system[123,124] (Figure 18f, 28-34). This transposon is bounded by left- and right-end sequences, distinguishable by their TnsB-binding sites, and includes a terminal operon that comprises the tnsA, tnsB and tnsC genes. Notably, the tniQ gene, a homolog of *E. coli* tnsD, is encoded within the cas rather than the tns operon, whereas tnsE is absent entirely. Like other such transposon-encoded CRISPR–Cas systems[32], the cas1 and cas2 genes responsible for spacer acquisition are conspicuously absent, as is the cas3 gene responsible for target DNA degradation. The putative DNA-targeting complex Cascade (also known as Csy complex[64]) is encoded by three genes: cas6, cas7 and a natural cas8–cas5 fusion[124] (hereafter referred to simply as cas8). The native CRISPR array, comprising four repeat and three spacer sequences, encodes mature CRISPR RNAs (crRNAs) that we also refer to as guide RNAs.

We transformed *E. coli* with plasmids that encode components of the *V. cholerae* transposon, including a mini-transposon donor (pDonor), the tnsA-tnsB-tnsC operon (pTnsABC), and the tniQ-cas8-cas7-cas6 operon alongside a synthetic CRISPR array (pQCascade) (Figure 13b). The CRISPR array was designed to produce a non-targeting crRNA or crRNA-1, which targets a genomic site downstream of glmS flanked by a 5′-CC-3′ protospacer adjacent motif (PAM)[125]. Notably, we observed PCR products from cellular lysate between a genome-specific primer and either of two transposon-specific primers in experiments containing pTnsABC, pDonor and pQCascade expressing crRNA-1, but not with a non-targeting crRNA or any empty vector controls (Figure 13c, d).

Because parallel reactions with oppositely oriented transposon primers revealed integration events within the same biological sample, we hypothesized that, unlike *E. coli Tn7*, RNA-guided transposition might occur in either orientation. We tested this by performing additional PCRs, by adding a downstream genomic primer, and by targeting an additional site with crRNA-2 found in the same genomic locus but on the opposite strand. For both crRNA-1 and crRNA-2, transposition products in both orientations were present, although with distinct orientation preferences based on relative band intensities (Figure 13e). Given the presence of discrete bands, it appeared that integration was occurring at a set distance from the target site, and Sanger and next-generation sequencing (NGS) analyzes revealed that more than 95% of integration events for crRNA-1 occurred 49 base pairs (bp) from the 3′ edge of the target site. The observed pattern with crRNA-2 was more complex, with integration clearly favoring distances of 48 and 50 bp over 49 bp. Both sequencing approaches also revealed the expected 5-bp target-site duplication that is a hallmark feature of *Tn7* transposition products[119] (Figure 13f, g).

The *V. cholerae* Tn*6677* transposon is not naturally present downstream of glmS, and we saw no evidence of site-specific transposition within this locus when we omitted the crRNA (Figure 13d). Nevertheless, we wanted to ensure that integration specificity was solely guided by the crRNA sequence, and not by any intrinsic preference for the *glmS* locus. We therefore cloned and tested crRNA-3 and crRNA-4, which target opposite strands within the *lacZ* coding sequence. We again observed bidirectional integration 48–50 bp downstream of both target sites, and were able to isolate clonally integrated, lacZ-knockout strains after performing blue–white colony screening on X-gal-containing LB-agar plates (Figure 13h, i and Figure 19). Collectively, these experiments demonstrate transposon integration downstream of genomic target sites complementary to guide RNAs.

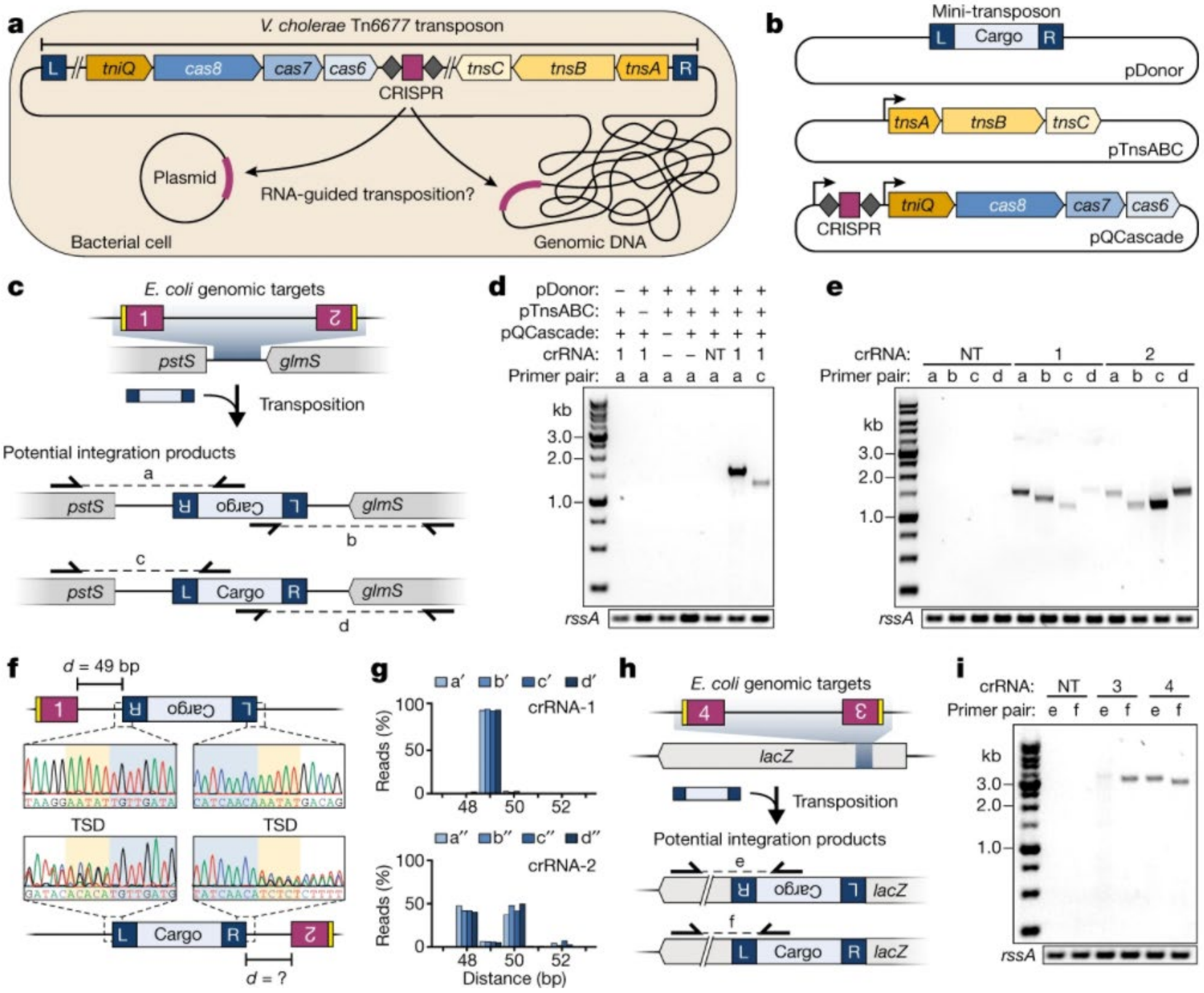


Figure 13: RNA-guided DNA integration with a *V. cholerae* transposon

a, Hypothetical scenario for Tn*6677* transposition into plasmid or genomic target sites complementary to a crRNA. **b**, Plasmid schematics for transposition experiments in which a mini-transposon on pDonor is mobilized in *trans*. The CRISPR array comprises two repeats (grey diamonds) and a single spacer (maroon rectangle). **c**, Genomic locus targeted by crRNA-1 and crRNA-2, two potential transposition products, and the PCR primer pairs to selectively amplify them. The PAMs and target sites are in yellow and maroon, respectively. **d**, PCR analysis of transposition with a non-targeting crRNA (crRNA-NT) and crRNA-1, resolved by agarose gel electrophoresis. **e**, PCR analysis of transposition with crRNA-NT, crRNA-1 and crRNA-2 using four distinct primer pairs, resolved by agarose gel electrophoresis. **f**, Sanger sequencing chromatograms for upstream and downstream junctions of genomically integrated transposons from experiments with crRNA-1 and crRNA-2. Overlapping peaks for crRNA-2 suggest the presence of multiple integration sites. The distance between the 3′ end of the target site and the first base of the transposon sequence is designated ‘*d*’. TSD, target-site duplication. **g**, NGS analysis of the distance between the Cascade target site and transposon integration site, determined for crRNA-1 and crRNA-2 with four primer pairs. **h**, Genomic locus targeted by crRNA-3 and crRNA-4. **i**, PCR analysis of transposition with crRNA-NT, crRNA-3 and crRNA-4, resolved by agarose gel electrophoresis. For **d**, **e** and **i**, amplification of *rssA* serves as a loading control.

3.4: Protein requirements of RNA-guided DNA integration

To confirm the involvement of transposon- and CRISPR-associated proteins in catalyzing RNA-guided DNA integration, we cloned and tested a series of plasmids in which each individual tns and cas gene was deleted, or in which the active site of each individual enzyme was mutated. Removal of any protein component abrogated transposition activity, as did mutations in the active site of the TnsB transposase, which catalyzes DNA integration[126], the TnsC ATPase, which regulates target site selection[49]vv, and the Cas6 RNase, which catalyzes pre-crRNA processing[81] (Figure 14a). A TnsA mutant that is catalytically impaired still facilitated RNA-guided DNA integration. On the basis of previous studies of *E. coli Tn7*, this variant system is expected to mobilize via replicative transposition as opposed to cut-and-paste transposition[29].

In *E. coli*, site-specific transposition requires att*Tn7* binding by EcoTnsD, followed by interactions with the EcoTnsC regulator protein to directly recruit the EcoTnsA-TnsB-donor DNA[36]. Given the essential nature of tniQ (a tnsD homolog) in RNA-guided transposition, and its location within the cas8-cas7-cas6 operon, we envisioned that the Cascade complex might directly bind TniQ and thereby deliver it to genomic target sites. We tested this hypothesis by recombinantly expressing CRISPR RNA and the *V. cholerae* tniQ-cas8-cas7-cas6 operon containing an N-terminal His10 tag on the TniQ subunit (Figure 20.a). TniQ co-purified with Cas8, Cas7 and Cas6, as shown by SDS–PAGE and mass spectrometry analysis, and the relative band intensities for each Cas protein were similar to TniQ-free Cascade and consistent with the 1:6:1

Cas8:Cas7:Cas6 stoichiometry expected for a I-F variant Cascade complex[80] (Figure 14b and Figure 20b). The complex migrated through a gel filtration column with an apparent molecular mass of roughly 440 kDa, in good agreement with its approximate expected mass, and both Cascade and TniQ–Cascade co-purified with a 60-nucleotide RNA species, which we confirmed was a mature crRNA by deep sequencing (Figure 14c, d and Figure 2.c, d). To validate the interaction between Cascade and TniQ further, we incubated separately purified samples in vitro and demonstrated complex formation by size-exclusion chromatography (Figure 20e). Together, these results reveal the existence of a novel TniQ–Cascade co-complex, highlighting a direct functional link between a CRISPR RNA-guided effector complex and a transposition protein.

To determine whether specific TniQ–Cascade interactions are required, or whether TniQ could direct transposition adjacent to generic R-loop structures or via artificial recruitment to DNA, we used Streptococcus pyogenes Cas9 (SpyCas9)[69] and Pseudomonas aeruginosa Cascade (PaeCascade)[80] as orthogonal RNA-guided DNA-targeting systems. After generating protein–RNA expression plasmids and programming both effector complexes with crRNAs that target the same lacZ sites as our earlier transposition experiments, we first validated DNA targeting by demonstrating efficient cell killing in the presence of an active Cas9 nuclease or the PaeCascade-dependent Cas2-3 nuclease (Figure 21a, b). When we transformed strains containing pTnsABCQ and pDonor with a plasmid encoding either catalytically deactivated Cas9-sgRNA (dCas9-sgRNA) or PaeCascade and performed PCR analysis of the resulting cell lysate, we found no evidence of site-specific transposition (Figure 14e), indicating that a genomic R-loop is insufficient for site-specific integration. We also failed to detect transposition when TniQ was directly fused to either terminus of dCas9, or to the Cas8 or Cas6 subunit of PaeCascade (Figure 14e), at least for the linker sequences tested. Notably, however, a similar fusion of TniQ to the Cas6 subunit of *V.*

cholerae Cascade, but not to the Cas8 subunit, restored RNA-guided transposition activity (Figure 14e and Figure 21c).

Together with our biochemical results, we conclude that TniQ forms essential interactions with Cascade, possibly via the Cas6 subunit, which could account for our finding that RNA-guided DNA insertion occurs downstream of the PAM-distal end of the target site where Cas6 is bound[84,127]v (Figure 14f). Because TniQ is required for transposition, we propose that it serves as an important connection between the CRISPR- and transposon-associated machineries during DNA targeting and integration, although further biochemical and structural studies will be required to define these mechanistic steps in greater detail.

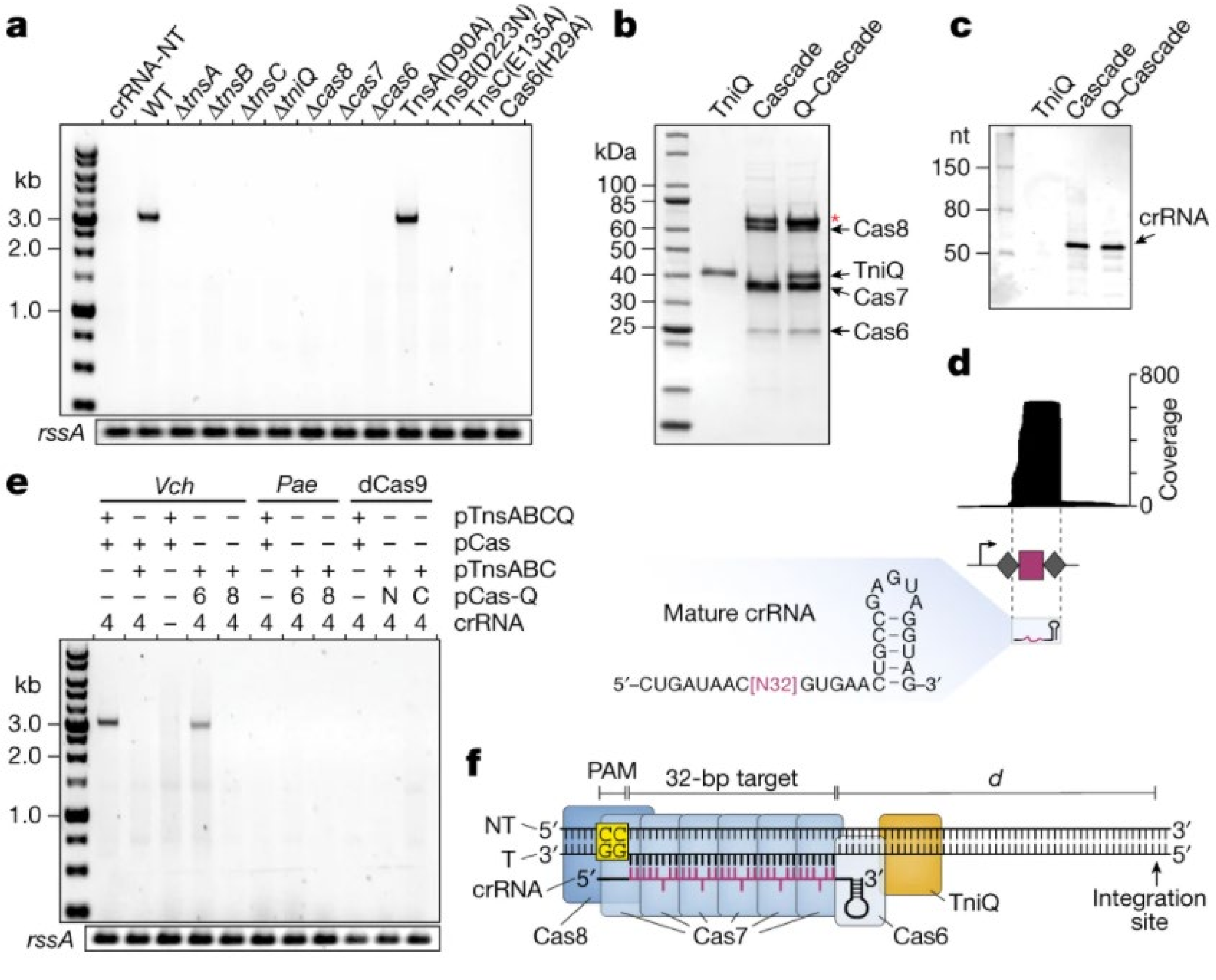


Figure 14: TniQ forms a complex with Cascade and is necessary for RNA-guided DNA integration.

a, PCR analysis of transposition with crRNA-4 and a panel of gene deletions or point mutations, resolved by agarose gel electrophoresis. b, SDS–PAGE analysis of purified TniQ, Cascade and a TniQ–Cascade (Q–Cascade) co-complex. Asterisk denotes an HtpG contaminant. c, Denaturing urea–PAGE analysis of co-purifying nucleic acids. nt, nucleotides. d, Top, RNA sequencing analysis of RNA co-purifying with Cascade. Bottom, reads mapping to the CRISPR array reveal the mature crRNA sequence. e, PCR analysis of transposition experiments testing whether generic R-loop formation or artificial TniQ tethering can direct targeted integration. The *V. cholerae* transposon and TnsA-TnsB-TnsC were combined with DNA-targeting components that comprise *V. cholerae* (Vch) Cascade, *P. aeruginosa* (Pae) Cascade, or *S. pyogenes* dCas9-RNA (dCas9). TniQ was expressed either on its own from pTnsABCQ or as a fusion to the targeting complex (pCas-Q) at the Cas6 C terminus (6), Cas8 N terminus (8), or dCas9 N or C terminus. f, Schematic of the R-loop formed upon target DNA binding by Cascade, with the approximate position of each protein subunit denoted. The putative TniQ-binding site and the distance to the primary integration site are indicated. NT, non-target strand; T, target strand. For a and e, amplification of rssA serves as a loading control.

3.5: Donor requirements of RNA-guided DNA integration

To determine the minimal donor requirements for RNA-guided DNA integration, as well as the effects of truncating the transposon ends and altering the cargo size, we first developed a quantitative PCR (qPCR) method for scoring transposition efficiency that could accurately and sensitively measure genomic integration events in both orientations (Figure 22). Analysis of cell lysates from transposition experiments using lacZ-targeting crRNA-3 and crRNA-4 yielded overall integration efficiencies of 62% and 42% without selection, respectively. The preference for integrating the 'right' versus the 'left' transposon end proximal to the genomic site targeted by Cascade was 39-to-1 for crRNA-3 and 1-to-1 for crRNA-4, suggesting the existence of additional sequence determinants that regulate integration orientation (Figure 15a, b).

With a quantitative assay in place, we were curious to investigate the effect of transposon size on RNA-guided integration efficiency and determine possible size constraints. When we progressively shortened or lengthened the DNA cargo in between the donor ends, beginning with our original mini-transposon donor plasmid (977 bp), we found that integration efficiency with our three-plasmid expression system was maximal with a 775-bp transposon and decayed with both the shorter and longer cargos tested (Figure 15c). Interestingly, naturally occurring *Tn7*-like transposons that encode CRISPR–Cas systems range from 20 to more than 100 kb in size[32], although their capacity for active mobility is unknown.

We next separately truncated both ends of the transposon. We found that around 105 bp of the left end and 47 bp of the right end were absolutely crucial for efficient RNA-guided DNA integration, corresponding to three and two intact putative TnsB-binding sites, respectively (Figure 23). Shorter transposons containing right-end truncations were integrated more efficiently, accompanied by a notable change in the orientation bias.

These experiments reveal crucial parameters for the development of programmable DNA integration technology. Future efforts will be required to explore how transposition is affected by vector design, to what extent transposon end mutations are tolerated, and whether rational engineering allows for integration of larger cargos and/or greater control over integration orientation.

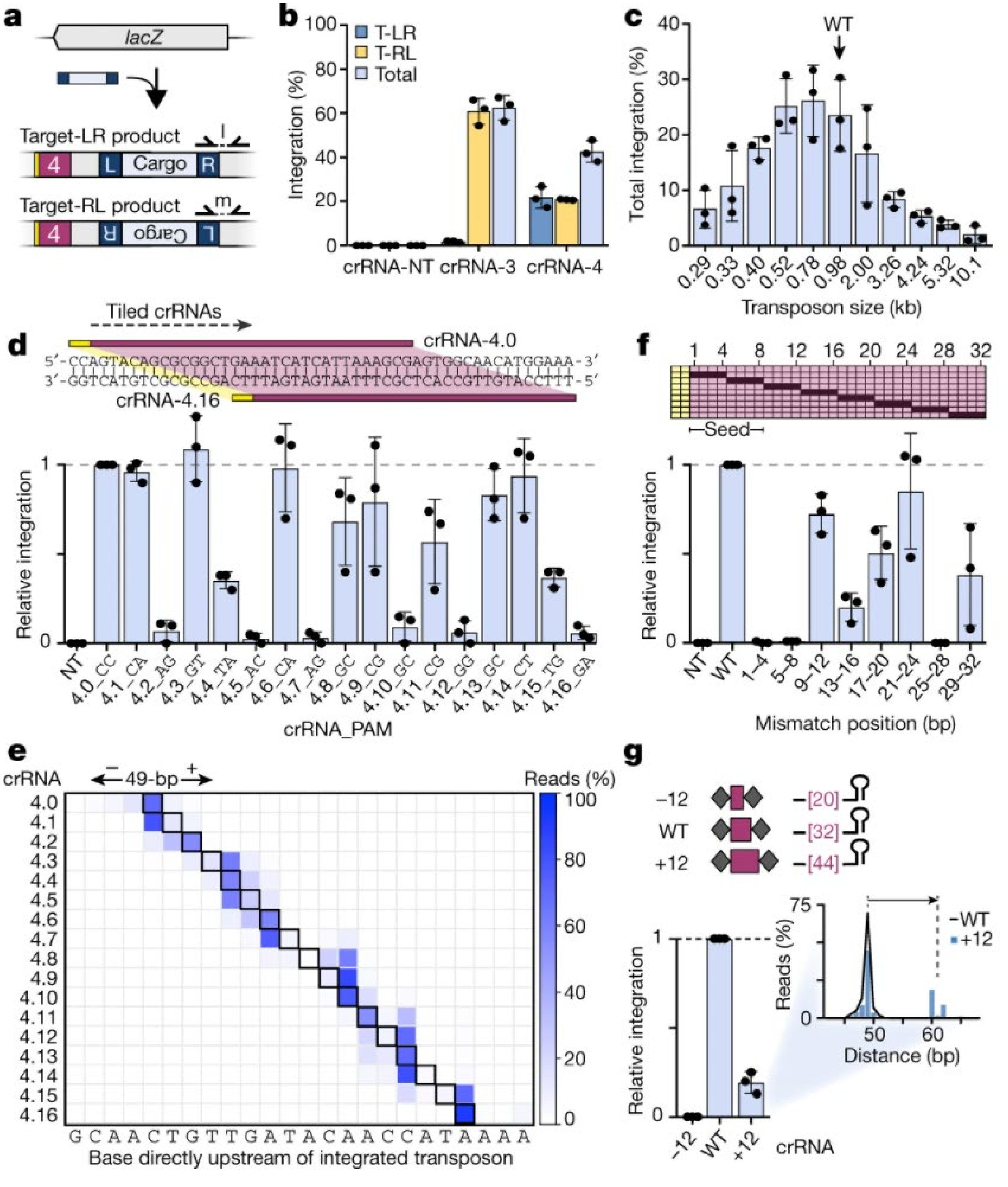


Figure 15: Influence of cargo size, PAM sequence, and crRNA mismatches on RNA-guided DNA integration.

a, Schematic of alternative integration orientations and the primer pairs to selectively detect them by qPCR. b, qPCR-based quantification of transposition efficiency in both orientations with crRNA-NT, crRNA-3 and crRNA-4. T-LR and T-RL denote transposition products in which the transposon left end and right end are proximal to the target site, respectively. c, Total integration efficiency with crRNA-4 as a function of transposon size. The arrow denotes the wild-type (WT) pDonor used in most assays throughout this study. d, crRNAs were tiled along the lacZ gene in 1-bp increments relative to crRNA-4 (4.0) (top), and the resulting integration efficiencies were determined by qPCR (bottom). Data are normalized to crRNA-4.0, and the 2-nucleotide PAM for each crRNA is shown. e, Heat map showing the integration site distribution (x axis) for each of the tiled crRNAs (y axis) in d, determined by NGS. The 49-bp distance for each crRNA is denoted by a black box. f, crRNAs were mutated in 4-nucleotide blocks to introduce crRNA-target DNA mismatches (black, top), and the resulting integration efficiencies were determined by qPCR (bottom). Data are normalized to crRNA-4. g, The crRNA-4 spacer length was shortened or lengthened by 12 nucleotides (top), and the resulting integration efficiencies were determined by qPCR (bottom). Data are normalized to crRNA-4 (WT). The inset shows a comparison of integration site distributions for crRNA-4 and crRNA-4.+12, determined by NGS. Data in b–d, f and g are shown as mean ± s.d. for n = 3 biologically independent samples.

3.6: Guide RNA and target DNA requirements

The Tn*6677*-encoded CRISPR–Cas system is most closely related to the I-F subtype, in which DNA target recognition by Cascade requires a consensus 5′-CC-3′ PAM[125], a high degree of sequence complementarity within a PAM-proximal seed sequence[80], and additional base-pairing across the entire 32-bp protospacer[128]. To determine sequence determinants of RNA-guided DNA integration, we first tested 12 dinucleotide PAMs by sliding the guide sequence in 1-bp increments along the lacZ gene relative to crRNA-4 (Figure 15d). In total, 8 distinct dinucleotide PAMs supported transposition at levels that were more than 25% of the 5′-CC-3′ PAM, and transposition occurred at over 1% total efficiency across the entire set of PAMs tested (Figure 15d). Additional deep sequencing revealed that the distance between the Cascade target site and primary transposon insertion site remained fixed at approximately 47–51 bp across the panel of crRNAs tested, although interesting patterns emerged, suggesting an additional layer of insertion site preference that requires further investigation (Figure 15e and Figure 24a).

Nevertheless, these experiments highlight how PAM recognition plasticity can be harnessed to direct a high degree of insertion flexibility and specificity at base-pair resolution.

To probe the sensitivity of transposition to RNA–DNA mismatches, we tested consecutive blocks of 4-nucleotide mismatches along the guide portion of crRNA-4 (Figure 15f). As expected from previous studies with Cascade homologs[129–131], mismatches within the 8-nucleotide seed sequence severely reduced transposition, probably owing to the inability to form a stable R-loop. Unexpectedly, however, our results highlighted a second region of mismatches at positions 25–28 that abrogated DNA integration, despite previous studies demonstrating that the stability of DNA binding is largely insensitive to mismatches in this region[129–131]. For the terminal mismatch block, which retained 17% integration activity, the distribution of observed insertion sites was markedly skewed to shorter distances from the target site relative to crRNA-4 (Figure 24b), which we hypothesize is the result of R-loop conformational heterogeneity.

Our emerging model for RNA-guided DNA integration involves Cascade-mediated recruitment of TniQ to target DNA. In the absence of any structural data, we realized that we could investigate whether TniQ may be positioned near the PAM-distal end of the R-loop by testing engineered crRNAs that contain spacers of variable lengths. Previous work with *E. coli* Cascade has demonstrated that crRNAs with extended spacers form complexes that contain additional Cas7 subunits[132], which would increase the distance between the PAM-bound Cas8 and the Cas6 at the other end of the R-loop. We therefore cloned and tested modified crRNAs containing spacers that were either shortened or lengthened in 6-nucleotide increments from the 3′ end. crRNAs with truncated spacers showed little or no activity, whereas extended spacers facilitated targeted integration, albeit at reduced levels with increasing length (Figure 24c, d). The +12-nucleotide crRNA directed transposition to two distinct regions: one approximately 49 bp from the 3′ end of the wild-type 32-nucleotide spacer, and an additional region shifted 11–13 bp away, in agreement with the expected increase in the length of the R-loop measured from the PAM (Figure 15g). Although more experiments are required to deduce the underlying mechanisms that explain this bimodal distribution, as well as the insertion site distribution observed for other extended crRNAs, these data, together with the mismatch panel, provide further evidence that TniQ is tethered to the PAM-distal end of the R-loop structure.

3.7: Programmability and genome-wide specificity

We lastly sought to examine both the programmability and the genome-wide specificity of our RNA-guided DNA integration system. First, we cloned and tested a series of crRNAs targeting additional genomic sites flanked by 5′-CC-3′ PAMs within the lac operon. Using the same primer pair for each resulting cellular lysate, we showed by PCR analysis that transposition was predictably repositioned with each distinct crRNA (Figure 16a). Our experiments thus far specifically interrogated genomic loci containing the anticipated integration products, and it therefore remained possible that non-specific integration was simultaneously occurring elsewhere, either at off-target genomic sites bound by Cascade, or independently of Cascade targeting. We thus adopted a transposon insertion sequencing (Tn-seq) pipeline previously developed for mariner transposons[133,134], in which all integration sites genome-wide are revealed by NGS (Figure 16b, Figure 25a, b and Methods). We first applied Tn-seq to a plasmid-encoded mariner transposon and found that our pipeline successfully recapitulated the genome-wide integration landscape previously observed with the Himar1c9 transposase[133,135] (Figure 16c, d and Figure 25c, d).

When we performed the same analysis for the RNA-guided *V. cholerae* transposon programmed with crRNA-4, we observed exquisite selectivity for lacZ-specific DNA integration

(Figure 16c). The observed integration site, which accounted for 99.0% of all Tn-seq reads that passed our filtering criteria (Methods), precisely matched the site observed by previous PCR amplicon NGS analysis (Figure 16e), and we did not observe reproducible off-target integration events elsewhere in the genome across three biological replicates (Figure 25e, f). Our Tn-seq data furthermore yielded diagnostic read pile-ups that highlighted the 5-bp target-site duplication and corroborated our previous measurements of transposon insertion orientation bias (Figure 16f). Tn-seq libraries from *E. coli* strains expressing pQCascade programmed with the non-targeting crRNA, or from strains lacking Cascade altogether (but still containing pDonor and pTnsABCQ), yielded far fewer genome-mapping reads, and no integration sites were consistently observed across several biological replicates (Figure 16c, Figure 136g, h).

In addition to performing Tn-seq with the crRNAs targeting glmS and lacZ genomic loci (Figure 26a), we cloned and tested an additional 16 crRNAs targeting the *E. coli* genome at 8 arbitrary locations spaced equidistantly around the circular chromosome. Beyond requiring that target sites were unique, were flanked by a 5′-CC-3′ PAM, and would direct DNA insertion to intergenic regions, we applied no further design rules or empirical selection criteria. Remarkably, when we analyzed the resulting Tn-seq data, we found that 16 out of 16 crRNAs directed highly precise RNA-guided DNA integration 46–55 bp downstream of the Cascade target, with around 95% of all filtered Tn-seq reads mapping to the on-target insertion site (Figure 16g and Figure 26b, c). These experiments highlight the high degree of intrinsic programmability and genome-wide integration specificity directed by transposon-encoded CRISPR–Cas systems.

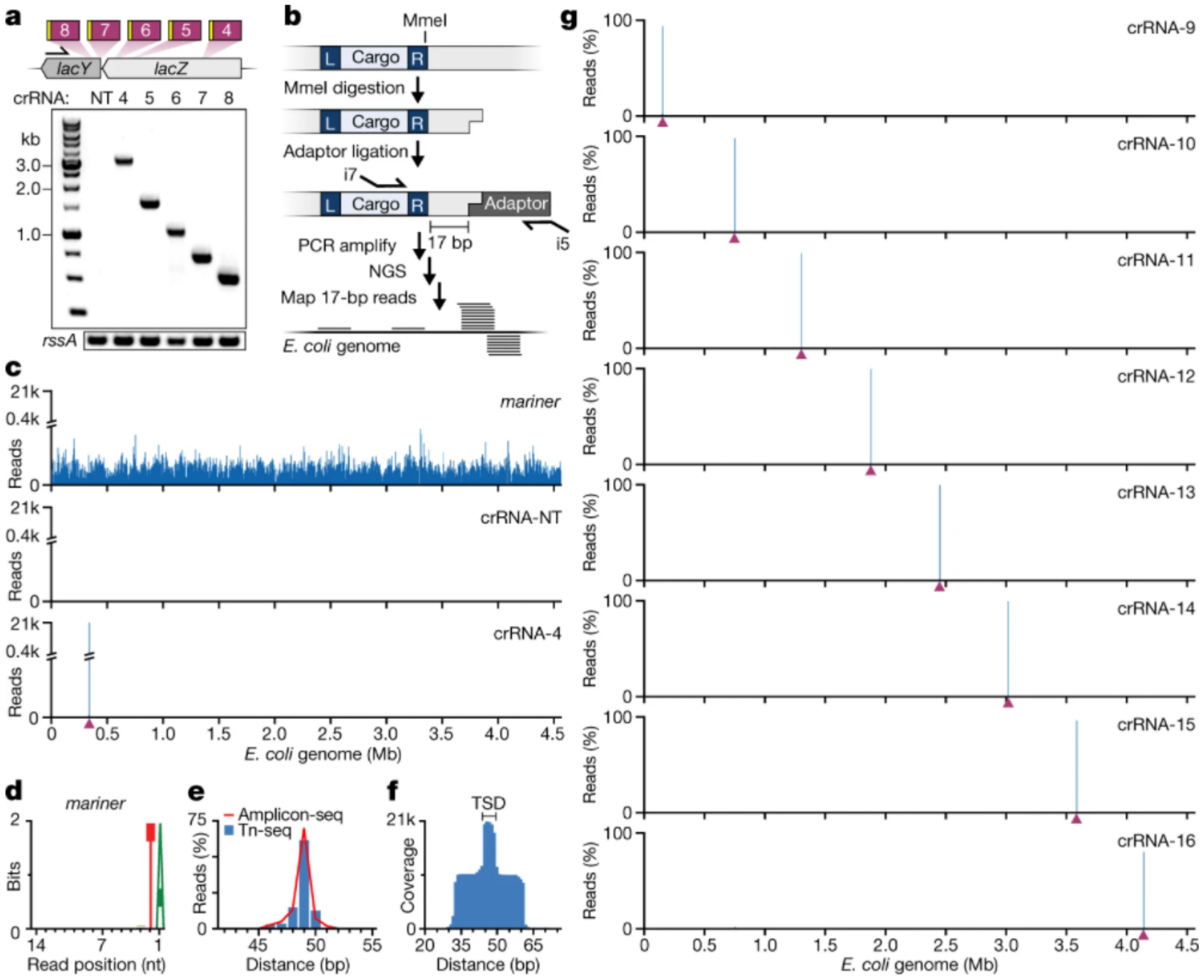


Figure 16: Genome-wide analysis of programmable RNA-guided DNA integration.

a, Genomic locus targeted by crRNAs 4–8 (top), and PCR analysis of transposition resolved by agarose gel electrophoresis (bottom). Amplification of rssA serves as a loading control. Tn-seq workflow for deep sequencing of genome-wide transposition events. c, Mapped Tn-seq reads from transposition experiments with the mariner transposon, and with the *V. cholerae* transposon programmed with either crRNA-NT or crRNA-4. The crRNA-4 target site is denoted by a maroon triangle. d, Sequence logo of all mariner Tn-seq reads, highlighting the TA dinucleotide target-site preference. e, Comparison of integration site distributions for crRNA-4 determined by PCR amplicon sequencing and Tn-seq, for the T-RL product; the distance between the Cascade target site and transposon integration site is plotted. f, Zoomed-in view of Tn-seq read coverage at the primary integration site for experiments with crRNA-4, highlighting the 5-bp target-site duplication (TSD); the distance from the Cascade target site is plotted. g, Genome-wide distribution of genome-mapping Tn-seq reads from transposition experiments with crRNAs 9–16 for the *V. cholerae* transposon. The location of each target site is denoted by a maroon triangle.

3.8: Discussion

Transposases and integrases are generally thought to mobilize their specific genetic payloads either by integrating randomly, with a low degree of sequence specificity, or by targeting specialized genomic loci through inflexible, sequence-specific homing mechanisms[8]. We have discovered a fully programmable integrase, in which the DNA insertion activity of a heteromeric transposase from *V. cholerae* is directed by an RNA-guided complex known as Cascade, the DNA-targeting specificity of which can be easily tuned. Beyond defining fundamental parameters that govern this activity, our work also reveals a complex between Cascade and TniQ that mechanistically connects the transposon- and CRISPR-associated machineries. On the basis of our results, and of previous studies of *Tn7* transposition[119], we propose a model for the RNA-guided mobilization of *Tn7*-like transposons encoding CRISPR–Cas systems (Figure 17). Because integration does not disrupt the Cascade-binding site, an important question for future investigation is whether the *V. cholerae* transposon exhibits a similar mode of target immunity as *E. coli Tn7*[42], in which repeated transposition into the same genomic locus is prevented.

Almost all type I-F CRISPR–Cas systems within the Vibrionaceae family are associated with mobile genetic elements, and those found within *Tn7*-like transposons frequently co-occur with restriction-modification and type three secretion systems[32,123]. It is therefore tempting to speculate that RNA-guided DNA integration may facilitate sharing of innate immune systems and virulence mechanisms via horizontal gene transfer, particularly within marine environments[136]. Interestingly, we and others[137,138] recently observed a unique clade of type V CRISPR–Cas systems that also reside within bacterial transposons, which bear many of the same features as *V. cholerae* Tn*6677*: the presence of the tniQ gene, the lack of predicted DNA cleavage activity by the RNA-guided effector complex[117], and cargo genes that frequently include other innate immune systems

(Figure 27). Although future experiments will be necessary to determine whether these systems also possess RNA-guided DNA integration activity, the bioinformatic evidence points to a more pervasive functional coupling between CRISPR–Cas systems and transposable elements than previously appreciated.

Many biotechnology products require genomic integration of large genetic payloads, including gene therapies[139], engineered crops[140] and biopharmaceuticals[141], and the advent of CRISPR-based genome editing has increased the need for effective knock-in methods. Yet current genome engineering solutions are limited by a lack of specificity, as with viral transduction[142], randomly integrating transposases[143] and non-homologous end joining[144] approaches, or by a lack of efficiency and cell-type versatility, as with homology-directed repair[145,146]. The ability to INsert Transposable Elements by Guide RNA-Assisted TargEting (INTEGRATE) offers an opportunity for site-specific DNA integration that would obviate the need for double-strand breaks in the target DNA, homology arms in the donor DNA, and host DNA repair factors. By virtue of its facile programmability, this technology could furthermore be leveraged for multiplexing and large-scale screening using guide RNA libraries. Together with other recent studies[107,147–149], our work highlights the far-reaching possibilities for genetic manipulation that continue to emerge from the diverse functions of CRISPR–Cas systems.

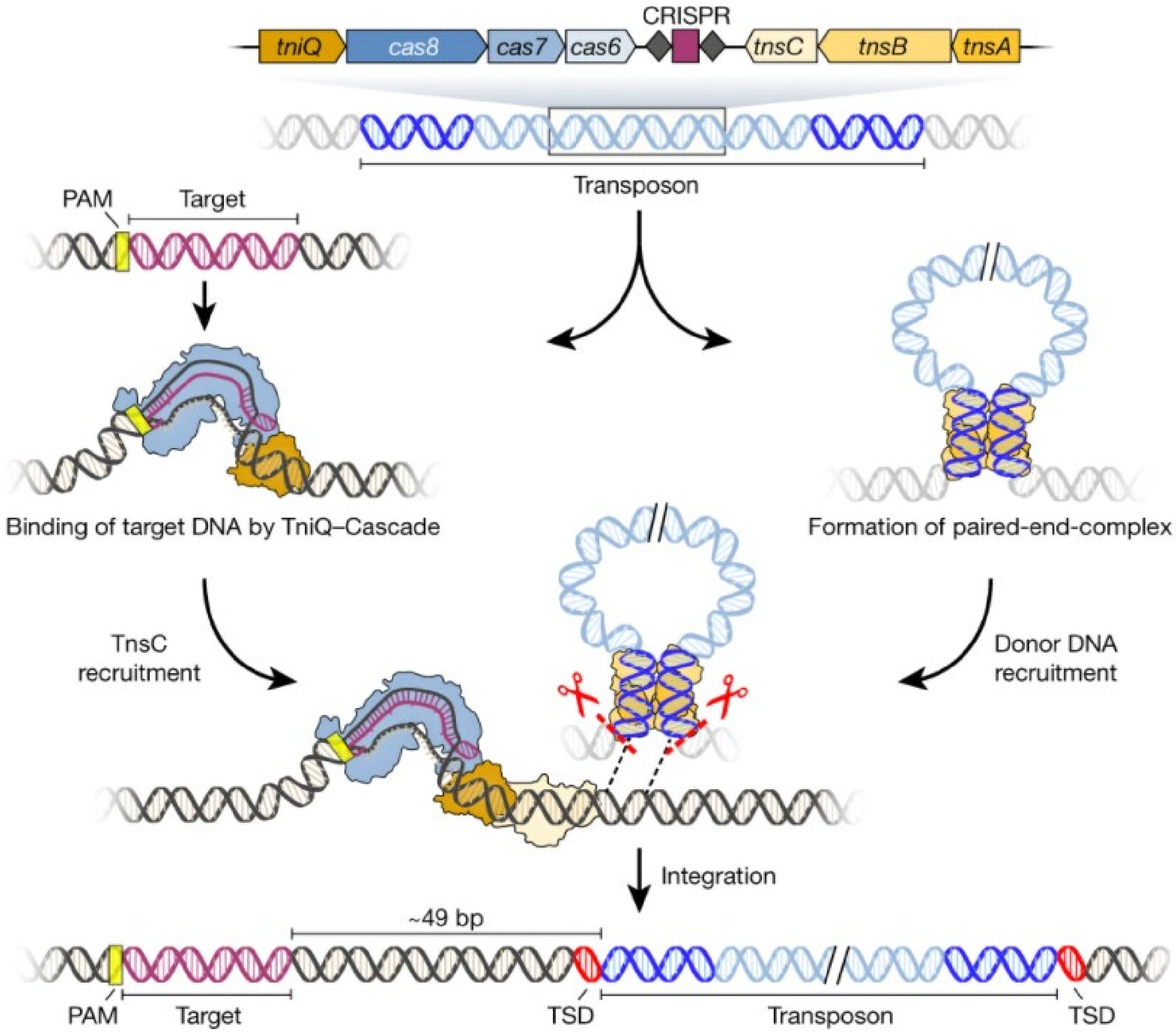


Figure 17: Proposed model for RNA-guided DNA integration by *Tn7*-like transposons encoding CRISPR-Cas systems.

The *V. cholerae* Tn*6677* transposon encodes a programmable RNA-guided DNA-binding complex called Cascade, which we have shown forms a co-complex with TniQ. We propose that TniQ–Cascade complexes survey the cell for matching DNA target sites, which may be found on the host chromosome or mobile genetic elements. After target binding and R-loop formation, TniQ presumably recruits the non-sequence-specific DNA-binding protein TnsC, based on previous studies of *E. coli Tn7* (reviewed in ref.106). The transposon itself is bound at the left and right ends by TnsA and TnsB, forming a so-called paired-end complex that is recruited to the target DNA by TnsC. Excision of the transposon from its donor site allows for targeted integration at a fixed distance downstream of DNA-bound TniQ–Cascade, resulting in a 5-bp target-site duplication.

3.9: Materials and Methods

Data Reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Plasmid Construction

All plasmids used in this study are described in an appended Supplementary Table, and a subset is available from Addgene. In brief, genes encoding *V. cholerae* strain HE-45 TnsA-TnsB-TnsC and TniQ-Cas8-Cas7-Cas6 (Figures 28-34) were synthesized by GenScript and cloned into pCOLADuet-1 and pCDFDuet-1, respectively, yielding pTnsABC and pQCascadeΔCRISPR. A pQCascade entry vector (pQCascade_entry) was generated by inserting tandem BsaI restriction sites flanked by two CRISPR repeats downstream of the first T7 promoter, and specific spacers (Supplementary Table 3) were subsequently cloned by oligoduplex ligation, yielding pQCascade. To generate pDonor, a gene fragment (GenScript) encoding both transposon ends was cloned into pUC19, and a chloramphenicol-resistance gene was subsequently inserted within the mini-transposon. Further derivatives of these plasmids were cloned using a combination of methods, including Gibson assembly, restriction digestion-ligation, ligation of hybridized oligonucleotides, and around-the-horn PCR. Plasmids were cloned and propagated in NEB Turbo cells (NEB), purified using Miniprep Kits (Qiagen), and verified by Sanger sequencing (GENEWIZ).

For transposition experiments involving the *E. coli Tn7* transposon, pEcoDonor was generated similarly to pDonor, and pEcoTnsABCD was subcloned from pCW4 (a gift from N. Craig, Addgene plasmid 8484). For transposition and cell killing experiments involving the I-F system from *P. aeruginosa*, genes encoding Cas8-Cas5-Cas7-Cas6 (also known as Csy1-Csy2-Csy3-Csy4) were subcloned from pBW64 (a gift from B. Wiedenheft), and the gene encoding the natural Cas2-3 fusion protein was subcloned from pCas1_Cas2/3 (a gift from B. Wiedenheft, Addgene plasmid 89240). For transposition and cell killing experiments involving the II-A system from *S. pyogenes*, the gene encoding Cas9 was subcloned from a vector in-house. For control Tn-seq experiments using the mariner transposon and Himar1C9 transposase, the relevant portions were subcloned from pSAM_Ec (a gift from M. Mulvey, Addgene plasmid 102939).

Expression plasmids for protein purification were subcloned from pQCascade into p2CT-10 (a gift from the QB3 MacroLab, Addgene plasmid 55209), and the crRNA expression construct was cloned into pACYCDuet-1.

Multiple sequence alignments (Figures 28-34) were performed using Clustal Omega with default parameters and visualized with ESPript 3.0[150]. Analysis of spacers from C2c5 CRISPR arrays (Figure 27) was performed using CRISPRTarget.

Transposition Experiments

All transposition experiments were performed in *E. coli* BL21(DE3) cells (NEB). For experiments including pDonor, pTnsABC and pQCascade (or variants thereof), chemically competent cells were first co-transformed with pDonor and pTnsABC, pDonor and pQCascade, or pTnsABC and pQCascade, and transformants were isolated by selective plating on double antibiotic LB-agar plates. Liquid cultures were then inoculated from single colonies, and the resulting strains were made chemically competent using standard methods, aliquoted and snap frozen. The third plasmid was introduced in a new transformation reaction by heat shock, and after recovering cells in fresh LB medium at 37 °C for 1 h, cells were plated on triple antibiotic LB-agar plates containing 100 µg ml−1 carbenicillin, 50 µg ml−1 kanamycin, and 50 µg ml−1 spectinomycin. After overnight growth at 37 °C for 16 h, hundreds of colonies were scraped from the plates, and a portion was resuspended in fresh LB medium before being re-plated on triple antibiotic LB-agar plates as before, this time supplemented with 0.1 mM IPTG to induce protein expression. Solid media culturing was chosen over liquid culturing in order to avoid growth competition and population bottlenecks. Cells were incubated an additional 24 h at 37 °C and typically grew as densely spaced colonies, before being scraped, resuspended in LB medium, and prepared for subsequent analysis. Control experiments lacking one or more molecular components were performed using empty vectors and the exact same protocol as above. Experiments investigating the effect of induction level on transposition efficiency contained variable IPTG concentrations in the media (Figure 22d). To isolate clonal, lacZ-integrated strains via blue-white colony screening, cells were re-plated on triple antibiotic LB-agar plates supplemented with 1 mM IPTG and 100 µg ml−1 X-gal (GoldBio), and grown overnight at 37 °C before colony PCR analysis.

PCR and Sanger Sequencing Analysis of Transposition Products

Optical density measurements at 600 nm were taken of scraped colonies that had been resuspended in LB medium, and approximately 3.2×10^8 cells (the equivalent of 200 µl of OD600 = 2.0) were transferred to a 96-well plate. Cells were pelleted by centrifugation at 4,000g for 5 min and resuspended in 80 µl of H2O, before being lysed by incubating at 95 °C for 10 min in a thermal cycler. The cell debris was pelleted by centrifugation at 4,000g for 5 min, and 10 µl of lysate supernatant was removed and serially diluted with 90 µl of H2O to generate 10- and 100-fold lysate dilutions for qPCR and PCR analysis, respectively.

PCR products were generated with Q5 Hot Start High-Fidelity DNA Polymerase (NEB) using 5 µl of 100-fold diluted lysate per 12.5 µl reaction volume serving as template. Reactions contained 200 µM dNTPs and 0.5 µM primers, and were generally subjected to 30 thermal cycles with an annealing temperature of 66 °C. Primer pairs contained one genome-specific primer and one transposon-specific primer, and were varied such that all possible integration orientations could be detected both upstream and downstream of the target site (see Supplementary Tables for selected oligonucleotides used in this study). Colony PCRs (Figure 19b) were performed by inoculating overnight cultures with individual colonies and performing PCR analysis as described above. PCR amplicons were resolved by 1–2% agarose gel electrophoresis and visualized by staining with SYBR Safe (Thermo Scientific). Negative control samples were always analyzed in parallel with experimental samples to identify mispriming products, some of which presumably result from the analysis being performed on crude cell lysates that still contain the high-copy pDonor. PCRs were initially performed with different DNA polymerases, variable cycling conditions, and different sample preparation methods. We note that higher concentrations of the crude lysate appeared to inhibit successful amplification of the integrated transposition product.

To map integration sites by Sanger sequencing, bands were excised after separation by gel electrophoresis, DNA was isolated by Gel Extraction Kit (Qiagen), and samples were submitted to and analyzed by GENEWIZ.

Integration Site Distribution Analysis by NGS of PCR Amplicons

PCR-1 products were generated as described above, except that primers contained universal Illumina adaptors as 5′ overhangs (Supplementary Tables) and the cycle number was reduced to 20. These products were then diluted 20-fold into a fresh polymerase chain reaction

(PCR-2) containing indexed p5/p7 primers and subjected to 10 additional thermal cycles using an annealing temperature of 65 °C. After verifying amplification by analytical gel electrophoresis, barcoded reactions were pooled and resolved by 2% agarose gel electrophoresis, DNA was isolated by Gel Extraction Kit (Qiagen), and NGS libraries were quantified by qPCR using the NEBNext Library Quant Kit (NEB). Illumina sequencing was performed using a NextSeq mid output kit with 150-cycle single-end reads and automated demultiplexing and adaptor trimming (Illumina). Individual bases with Phred quality scores under 20 (corresponding to a base miscalling rate of >1%) were changed to 'N,' and only reads with at least half the called bases above Q20 were retained for subsequent analysis.

To determine the integration site distribution for a given sample, the following steps were performed using custom Python scripts. First, reads were filtered based on the requirement that they contain 20 bp of perfectly matching transposon end sequence. Fifteen base pairs of sequence immediately flanking the transposon were then extracted and aligned to a 1-kb window of the *E. coli* BL21(DE3) genome (GenBank accession CP001509) surrounding the crRNA-matching genomic target site. The distance between the nearest transposon–genome junction and the PAM-distal edge of the 32-bp target site was determined. Histograms were plotted after compiling these distances across all the reads within a given library (Supplementary Tables for NGS statistics).

Cell Killing Experiments

For experiments with Cas9, 40 µl chemically competent BL21(DE3) cells were transformed with 100 ng Cas9-sgRNA expression plasmid encoding either sgRNA-3 or sgRNA-4, which target equivalent lacZ sites as *V. cholerae* crRNA-3 or crRNA-4 but on opposite strands, or a truncated/non-functional sgRNA derived from the BsaI-containing entry vector

(Supplementary Tables). After a one-hour recovery at 37 °C, variable dilutions of cells were plated on LB-agar plates containing 100 µg ml−1 carbenicillin and 0.1 mM IPTG and grown an additional 16 h at 37 °C. The number of resulting colonies was quantified across three biological replicates, and the data were plotted as colony-forming units per microgram of plasmid DNA. Additional control experiments used an expression plasmid encoding Cas9 nuclease-inactivating D10A and H840A mutations (dCas9).

For experiments with Cascade and Cas2-3 from *P. aeruginosa*, BL21(DE3) cells were first transformed with a Cas2-3 expression vector, and the resulting strains were made chemically competent. Forty microlitres of these cells were then transformed with 100 ng PaeCascade expression plasmid encoding either crRNA-Pae3 or crRNA-Pae4, which target equivalent lacZ sites as *V. cholerae* crRNA-3 or crRNA-4, or a truncated/non-functional crRNA derived from the BsaI-containing entry vector (Supplementary Tables). After a one-hour recovery at 37 °C, variable dilutions of cells were plated on LB-agar plates containing 100 µg ml−1 carbenicillin and 50 µg ml−1 kanamycin and grown an additional 16 h at 37 °C. The number of resulting colonies was quantified across three biological replicates, and the data were plotted as colony-forming units per microgram of plasmid DNA. We found that even low concentrations of IPTG led to crRNA-independent toxicity in these experiments, whereas crRNA-dependent cell killing was readily observed in the absence of induction, presumably from leaky expression by T7 RNAP. We therefore omitted IPTG from experiments using PaeCascade and Cas2-3.

qPCR analysis of transposition efficiency

For both crRNA-3 and crRNA-4, pairs of transposon- and genome-specific primers were designed to amplify an approximately 140–240-bp fragment resulting from RNA-guided DNA integration at the expected lacZ locus in either orientation. A separate pair of genome-specific primers was designed to amplify an *E. coli* reference gene (rssA) for normalization purposes (Supplementary Tables). qPCR reactions (10 µl) contained 5 µl of SsoAdvanced Universal SYBR Green Supermix (BioRad), 1 µl H2O, 2 µl of 2.5 µM primers, and 2 µl of tenfold diluted lysate prepared from scraped colonies, as described for the PCR analysis above. Reactions were prepared in 384-well clear/white PCR plates (BioRad), and measurements were performed on a CFX384 Real-Time PCR Detection System (BioRad) using the following thermal cycling parameters: polymerase activation and DNA denaturation (98 °C for 2.5 min), 40 cycles of amplification (98 °C for 10 s, 62 °C for 20 s), and terminal melt-curve analysis (65–95 °C in 0.5 °C per 5 s increments).

We first prepared lysates from a control BL21(DE3) strain containing pDonor and both empty expression vectors (pCOLADuet-1 and pCDFDuet-1), and from strains that underwent clonal integration into the lacZ locus downstream of both crRNA-3 and crRNA-4 target sites in both orientations. By testing our primer pairs with each of these samples diluted across five orders of magnitude, and then determining the resulting Cq values and PCR efficiencies, we verified that our experimental and reference amplicons were amplified with similar efficiencies, and that our primer pairs selectively amplified the intended transposition product (Figure 22a, b). We next simulated variable transposition efficiencies across five orders of magnitude (ranging from 0.002 to 100%) by mixing control lysates and clonally-integrated lysates in various ratios, and showed that we could accurately and reproducibly detect transposition products at both target sites, in either orientation, at levels >0.01% (Figure 22.b). Finally, we simulated variable integration orientation biases by mixing clonally-integrated lysates together in varying ratios together with control lysates, and showed that these could also be accurately measured (Figure 22c).

In our final qPCR analysis protocol, each biological sample is analyzed in three parallel reactions: one reaction contains a primer pair for the *E. coli* reference gene, a second reaction contains a primer pair for one of the two possible integration orientations, and a third reaction contains a primer pair for the other possible integration orientation. Transposition efficiency for each orientation is then calculated as 2ΔCq, in which ΔCq is the Cq difference between the experimental reaction and the control reaction. Total transposition efficiency for a given experiment is calculated as the sum of transposition efficiencies for both orientations. All measurements presented in the text and figures were determined from three independent biological replicates.

We note that experiments with pDonor variants were performed by delivering pDonor in the final transformation step, whereas most other experiments were performed by delivering pQCascade in the final transformation step. Integration efficiencies between samples from these two experiments appeared to differ slightly as a result (compare Figure 15b with Figure 15c). Additionally, because we did not want to bias our qPCR analysis of the donor end truncation samples by successively shortening the PCR amplicon, different primer pairs were used for these samples. Within the left and right end truncation panel (Figure 23b, c), the transposon end that was not being perturbed was selectively amplified during qPCR analysis.

Recombinant protein expression and purification

The protein components for Cascade, TniQ and TniQ–Cascade were expressed from a pET-derivative vector containing an N-terminal His10-MBP-TEVsite fusion on Cas8, TniQ and TniQ, respectively (see Figure 20a). The crRNAs for Cascade and TniQ–Cascade were expressed separately from a pACYC-derivative vector (Supplementary Tables). *E. coli* BL21(DE3) cells containing one or both plasmids were grown in 2xYT medium with the appropriate antibiotic(s) at 37 °C to OD600 = 0.5–0.7, at which point IPTG was added to a final concentration of 0.5 mM and growth was allowed to continue at 16 °C for an additional 12–16 h. Cells were harvested by centrifugation at 4,000g for 20 min at 4 °C.

Cascade and TniQ–Cascade were purified as follows. Cell pellets were resuspended in Cascade lysis buffer (50 mM Tris-Cl, pH 7.5, 100 mM NaCl, 0.5 mM PMSF, EDTA-free Protease Inhibitor Cocktail tablets (Roche), 1 mM dithiothreitol (DTT), 5% glycerol) and lysed by sonication with a sonic dismembrator (Fisher) set to 40% amplitude and 12 min total process time (cycles of 10 s on and 20 s off, for a total of 4 min on and 8 min off). Lysates were clarified by centrifugation at 15,000g for 30 min at 4 °C. Initial purification was performed by immobilized metal-ion affinity chromatography with NiNTA Agarose (Qiagen) using NiNTA wash buffer (50 mM Tris-Cl, pH 7.5, 100 mM NaCl, 10 mM imidazole, 1 mM DTT, 5% glycerol) and NiNTA elution buffer (50 mM Tris-Cl pH 7.5, 100 mM NaCl, 300 mM imidazole, 1 mM DTT, 5% glycerol). The His10-MBP fusion was removed by incubation with TEV protease overnight at 4 °C in NiNTA elution buffer, and complexes were further purified by anion exchange chromatography on an AKTApure system (GE Healthcare) using a 5 ml HiTrap Q HP Column (GE Healthcare) with a linear gradient from 100% buffer A (20 mM Tris-Cl, pH 7.5, 100 mM NaCl, 1 mM DTT, 5% glycerol) to 100% buffer B (20 mM Tris-Cl, pH 7.5, 1 M NaCl, 1 mM DTT, 5% glycerol) over 20 column volumes. Pooled fractions were identified by SDS–PAGE analysis and concentrated, and the sample was further refined by size exclusion chromatography over one or two tandem Superose 6 Increase 10/300 columns (GE Healthcare) equilibrated with Cascade storage buffer (20 mM Tris-Cl, pH 7.5, 200 mM NaCl, 1 mM DTT, 5% glycerol). Fractions were pooled, concentrated, snap frozen in liquid nitrogen, and stored at −80 °C.

TniQ was purified similarly, except the lysis, NiNTA wash, and NiNTA elution buffers contained 500 mM NaCl instead of 100 mM NaCl. Separation by ion exchange chromatography was performed on a 5 ml HiTrap SP HP Column (GE Healthcare) using the same buffer A and buffer B as above, and the final size-exclusion chromatography step was performed on a HiLoad Superdex 75 16/600 column (GE Healthcare) in Cascade storage buffer. The TniQ protein used in TniQ–Cascade binding experiments (Figure 20e) contained an N-terminal StrepII tag (Supplementary Tables).

Mass spectrometry analysis

Total protein (0.5–5 µg) was separated on 4–20% gradient SDS–PAGE and stained with Imperial Protein Stain (Thermo Scientific). In-gel digestion was performed essentially as described, with minor modifications. Protein gel slices were excised, washed with 1:1 acetonitrile:100 mM ammonium bicarbonate (v/v) for 30 min, dehydrated with 100% acetonitrile for 10 min, and dried in a speed-vac for 10 min without heat. Gel slices were reduced with 5 mM DTT for 30 min at 56 °C and then alkylated with 11 mM iodoacetamide for 30 min at room temperature in the dark. Gel slices were washed with 100 mM ammonium bicarbonate and 100% acetonitrile for 10 min each, and excess acetonitrile was removed by drying in a speed-vac for 10 min without heat. Gel slices were then rehydrated in a solution of 25 ng µl−1 trypsin in 50 mM ammonium bicarbonate for 30 min on ice, and trypsin digestions were performed overnight at 37 °C. Digested peptides were collected and further extracted from gel slices in mass spectrometry (MS) extraction buffer (1:2 5% formic acid:acetonitrile (v/v)) with high-speed shaking. Supernatants were dried down in a speed-vac, and peptides were dissolved in a solution containing 3% acetonitrile and 0.1% formic acid.

Desalted peptides were injected onto an EASY-Spray PepMap RSLC C18 50 cm × 75 μm column (Thermo Scientific), which was coupled to the Orbitrap Fusion Tribrid mass spectrometer (Thermo Scientific). Peptides were eluted with a nonlinear 100-min gradient of 5–30% mass spectrometry buffer B (MS buffer A: 0.1% (v/v) formic acid in water; MS buffer B: 0.1% (v/v) formic acid in acetonitrile) at a flow rate of 250 nl min−1. Survey scans of peptide precursors were performed from 400 to 1,575 m/z at 120K full width at half-maximum resolution (at 200 m/z) with a 2 × 10^5 ion count target and a maximum injection time of 50 ms. The instrument was set to run in top speed mode with 3-s cycles for the survey and the tandem mass spectrometry (MS/MS) scans. After a survey scan, tandem mass spectrometry was performed on the most abundant precursors exhibiting a charge state from 2 to 6 of greater than 5 × 10^3 intensity by isolating them in the quadrupole at 1.6 Th. CID fragmentation was applied with 35% collision energy, and resulting fragments were detected using the rapid scan rate in the ion trap. The AGC target for MS/MS was set to 1 × 10^4 and the maximum injection time limited to 35 ms. The dynamic exclusion was set to 45 s with a 10 ppm mass tolerance around the precursor and its isotopes. Monoisotopic precursor selection was enabled.

Raw mass spectrometric data were processed and searched using the Sequest HT search engine within the Proteome Discoverer 2.2 software (Thermo Scientific) with custom sequences and the reference *E. coli* BL21(DE3) strain database downloaded from Uniprot. The default search settings used for protein identification were as follows: two mis-cleavages for full trypsin, with fixed carbamidomethyl modification of cysteine and oxidation of methionine; deamidation of asparagine and glutamine and acetylation on protein N termini were used as variable modifications. Identified peptides were filtered for a maximum 1% false discovery rate using the

Percolator algorithm, and the PD2.2 output combined folder was uploaded in Scaffold (Proteome Software) for data visualization. Spectral counting was used for analysis to compare the samples.

crRNA analysis and RNA sequencing

To analyze the nucleic acid component co-purifying with Cascade and TniQ–Cascade, nucleic acids were isolated by phenol-chloroform extraction, resolved by 10% denaturing urea–PAGE, and visualized by staining with SYBR Gold (Thermo Scientific). Analytical RNase and DNase digestions were performed in 10 µl reactions with approximately 4 pmol nucleic acid and either 10 µg RNase A (Thermo Scientific) or 2 U DNase I (NEB), and were analyzed by 10% denaturing urea–PAGE and SYBR Gold staining.

RNA sequencing was performed generally as previously described[151]. In brief, RNA was isolated from Cascade and TniQ–Cascade complexes by phenol-chloroform extraction, ethanol precipitated, and 5′-phosphorylated/3′-dephosphorylated using T4 polynucleotide kinase (NEB), followed by clean-up using the ssDNA/RNA Clean & Concentrator Kit (Zymo Research). A ssDNA universal Illumina adaptor containing 5′-adenylation and 3′-dideoxycytidine modifications (Supplementary Tables) was ligated to the 3′ end with T4 RNA Ligase 1 (NEB), followed by hybridization of a ssDNA reverse transcriptase primer and ligation of ssRNA universal Illumina adaptor to the 5′ end with T4 RNA Ligase 1 (NEB). cDNA was synthesized using Maxima H Minus Reverse Transcriptase (Thermo Scientific), followed by PCR amplification using indexed p5/p7 primers. Illumina sequencing was performed using a NextSeq mid output kit with 150-cycle single-end reads and automated demultiplexing and adaptor trimming (Illumina). Individual bases with Phred quality scores under 20 (corresponding to a base miscalling rate of >1%) were changed to 'N,' and only reads with at least half the called bases above Q20 were retained for subsequent analysis. Reads were aligned to the crRNA expression plasmid used for recombinant Cascade and TniQ–Cascade expression and purification.

TniQ-Cascade binding experiments

Binding reactions (120 µl) contained 1 µM Cascade and 5 µM StrepII-tagged TniQ, and were prepared in Cascade storage buffer and incubated at room temperature for 30 min, before being loaded into a 100 µl sample loop on an AKTApure system (GE Healthcare). Reactions were resolved by size exclusion chromatography over a Superose 6 Increase 10/300 column (GE Healthcare) in Cascade storage buffer, and proteins in each peak fraction were acetone precipitated and analyzed by SDS–PAGE. Control reactions lacked either Cascade or TniQ.

Tn-seq experiments

Transposition experiments were performed as described above, except pDonor contained two point mutations in the transposon right end that introduced an MmeI restriction site (Supplementary Tables and Figure 25a, b). Colonies from triple antibiotic LB-agar plates containing IPTG (typically numbering in the range of 10^2–10^3) were resuspended in 4 ml fresh LB medium, and 0.5 ml (corresponding to around 2×10^9 cells) was used for genomic DNA (gDNA) extraction with the Wizard Genomic DNA Purification Kit (Promega). This procedure typically yielded 50 µl of 0.5–1.5 µg µl−1 gDNA, which is a mixture of the *E. coli* circular chromosome (4.6 Mb, copy number of 1), pDonor (3.6 kb, copy number 100+), pTnsABC (6.9 kb, copy number ~20–40), and pQCascade (8.4 kb, copy number ~20–40).

NGS libraries were prepared in parallel on 96-well plates, as follows. First, 1 µg of gDNA was digested with 4 U of MmeI (NEB) for 12 h at 37 °C in a 50 µl reaction containing 50 µM Sadenosyl methionine and 1× CutSmart Buffer, before heat inactivation at 65 °C for 20 min. MmeI cleaves the transposon 17–19 nucleotides outside of the terminal repeat, leaving 2-nucleotide 3′-overhangs. Reactions were cleaned up using 1.8× Mag-Bind TotalPure NGS magnetic beads (Omega) according to the manufacturer's instructions, and elutions were performed using 30 µl of 10 mM Tris-Cl, pH 7.0. MmeI-digested gDNA was ligated to a double-stranded i5 universal adaptor containing a 3′-terminal NN overhang (Supplementary Tables) in a 20 µl ligation reaction containing 16.86 µl of MmeI-digested gDNA, 280 nM adaptor, 400 U T4 DNA ligase (NEB), and 1× T4 DNA ligase buffer. Reactions were incubated at room temperature for 30 min before being cleaned up with magnetic beads as before. To reduce the degree of pDonor contamination within our NGS libraries, since pDonor also contains the full-length transposon with an MmeI site, we took advantage of the presence of a unique HindIII restriction site just outside the transposon right end within pDonor. The entirety of the adaptor-ligated gDNA sample was thus digested with 20 Units of HindIII (NEB) in a 34.4 µl reaction for 1 h at 37 °C, before a heat inactivation step at 65 °C for 20 min. Magnetic bead-based DNA clean-up was performed as before.

Adaptor-ligated transposons were enriched in a PCR-1 step using a universal i5 adaptor primer and a transposon-specific primer containing a universal i7 adaptor as 5′ overhang. Reactions were 25 µl in volume and contained 16.75 µl of HindIII-digested gDNA, 200 µM dNTPs, 0.5 µM primers, 1× Q5 reaction buffer, and 0.5 U Q5 Hot Start High-Fidelity DNA Polymerase (NEB). Amplification was allowed to proceed for 25 cycles, with an annealing temperature of 66 °C. Reaction products were then diluted 20-fold into a second 20 µl polymerase chain reaction (PCR-2) containing indexed p5/p7 primers, and this was subjected to 10 additional thermal cycles using an annealing temperature of 65 °C. After verifying amplification for select libraries by analytical gel electrophoresis, barcoded reactions were pooled and resolved by 2%

agarose gel electrophoresis, DNA was isolated by Gel Extraction Kit (Qiagen), and NGS libraries were quantified by qPCR using the NEBNext Library Quant Kit (NEB). Illumina sequencing was performed using a NextSeq mid output kit with 150-cycle single-end reads and automated demultiplexing and adaptor trimming (Illumina). Individual bases with Phred quality scores under 20 (corresponding to a base miscalling rate of >1%) were changed to ‘N,’ and only reads with at least half the called bases above Q20 were retained for subsequent analysis.

Tn-seq libraries with the mariner transposon were prepared as for the *V. cholerae* transposon, but with the following changes. Transformation reactions contained BL21(DE3) cells and a single pDonor plasmid, which encodes a KanR-containing mariner transposon with MmeI restriction sites on both ends, and a separate expression cassette for the Himar1C9 transposase controlled by a lac promoter. Transformed cells were recovered at 37 °C for 1 h before being plated on bioassay dishes containing 100 µg ml−1 carbenicillin, yielding on the order of 5×10^4 colonies. Cells were resuspended in 20 ml fresh LB medium after a single 16-h overnight growth, and the equivalent of 2×10^9 cells were used for genomic DNA (gDNA) extraction. NGS libraries were prepared as described above, except the restriction enzyme digestion reactions to deplete pDonor contained 20 U of BamHI and KpnI instead of HindIII.

Tn-seq data visualization and analysis

The software application Geneious Prime was used to further filter reads based on three criteria: that read lengths correspond to the expected products resulting from MmeI cleavage and adaptor ligation to genomically integrated transposons (112–113 bp for the *V. cholerae* transposon and 87–88 bp for mariner); that each read contain the expected transposon end sequence (allowing for one mismatch); and that the transposon-flanking sequence (trimmed to 17 bp for the *V. cholerae* transposon and 14 bp for mariner) map perfectly to the reference genome. Mapping to the *E. coli* BL21(DE3) genome (GenBank accession CP001509) was done using the function 'Map to reference' and the following settings: Mapper: Geneious; Fine tuning: None (fast / read mapping); Word length: 17; Maximum mismatches: 0%; Maximum Ambiguity: 1. The 'Map multiple best matches' setting was set to either 'none,' effectively excluding any reads except those that map uniquely to a single site (which we will refer to as 'uniquely mapping reads'), or to 'all,' which allows reads to map to one or multiple sites on the *E. coli* genome (which we will refer to as 'processed mapping reads'). Both sets of reads were exported as fastq files and used for downstream analysis using custom Python scripts. We note that many reads removed in this process perfectly mapped to the donor plasmid (Supplementary Table 4), revealing that HindIII or BamHI/KpnI cleavage was insufficient to completely remove contaminating pDonor-derived sequences. Coverage data for 'processed mapping reads' were exported to generate Figure 16f.

To visualize the genome-wide integration site distribution for a given sample, 'uniquely mapping reads' were mapped to the same *E. coli* reference genome with custom Python scripts. We define the integration site for each read as the genomic coordinate (with respect to the reference genome) corresponding to the 3′ edge of the mapped read. For visualization purposes, integration events within 5-kb bins were computed and plotted as genome-wide histograms in Figure 16c, g and Figure 3.101.14.a, b. Plots were generated using the Matplotlib graphical library. The sequence logo in Figure 16.d was generated using WebLogo 3.

Plots comparing integration sites among biological replicates (Figure 25d-h) were generated by binning the genome-wide histograms based on gene annotations (mariner) using GenBank accession CP001509, or into 100-bp bins (*V. cholerae* transposon). For the *V. cholerae* transposon, the bins were shifted so that the 3′ end of the Cascade target site for each sample would correspond to the start of its corresponding 100-bp bin. Linear regression and bivariate analysis for the mariner plot (Figure 25d) was performed using the SciPy statistical package.

To analyze the primary integration site for each sample, custom Python scripts were used to map ‘processed mapping reads’ to a 600-bp genomic window surrounding the corresponding genomic target site. For reads mapping to the opposite strand as the target (that is, for the T-LR orientation, in which integration places the ‘left’ transposon end closest to the Cascade-binding site), the integration site was shifted 5 bp from the 3′ edge of the target site in order to account for the 5-bp target-site duplication. We define the primary integration site within this 600-bp window by the largest number of mapped reads, while we arbitrarily designate 100 bp centered at the primary integration site as the ‘on-target’ window. The percentage of on-target integration for each sample is calculated as the number of reads resulting from transposition within the 100-bp window, divided by the total number of reads mapping to the genome. We also determined the ratio of integration in one orientation versus the other; this parameter only utilizes on-target reads, and is calculated as the number of reads resulting from integration of the transposon ‘right’ end closest to the Cascade-binding site (T-RL), divided by the number reads resulting from integration of the transposon left end closest to the Cascade target site (T-LR). The distribution of integration around the primary site was plotted for both orientations for each sample, and was used to generate Figure 16e and Figure 26c. We note that these analyzes are susceptible to potential biases from differential efficiencies in the ligation of 3′-terminal NN overhang adaptors, which are not taken into account in our analyzes.

Statistics and reproducibility

Analytical PCRs resolved by agarose gel electrophoresis gave similar results in three independent replicates (Figures 13d, e, i, 14a, and 16a) or were analyzed by gel electrophoresis once (Figure 14e and Figures 18d, 19b, d, and f) but verified with qPCR for three independent replicates (Figure 14e). Sanger sequencing and next-generation sequencing of PCR amplicons was performed once (Figures 13f, g, 15e, 16e and Figures 18e, 19a, e, and ref. 132). SDS–PAGE experiments were performed for two or more different preparations of the same protein complexes and yielded similar results (Figure 14b and Figure 20b). Protein binding reactions were performed and analyzed by SDS–PAGE once (Figure 20e). Nucleic acid extraction from purified protein preparations and urea–PAGE analysis of samples with and without RNase or DNase treatment was performed twice, with similar results (Figure 14c and Figure 3.101.8.d). RNA sequencing was performed once (Figure 14d).

3.10: Figures

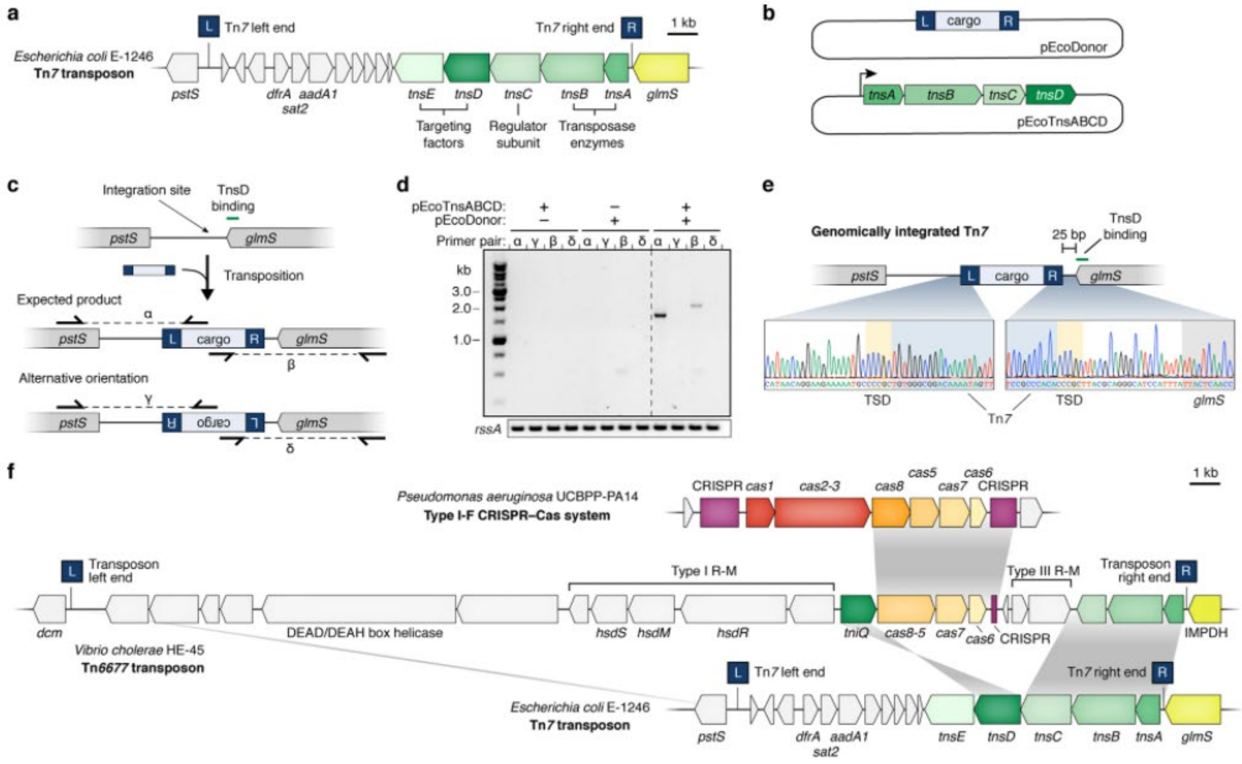


Figure 18: Transposition of the *E. coli Tn7* transposon and genetic architecture of the Tn*6677* transposon from *V. cholerae*

a, Genomic organization of the native *E. coli Tn7* transposon adjacent to its known attachment site (att*Tn7*) within the glmS gene. b, Expression plasmid and donor plasmid for *Tn7* transposition experiments. c, Genomic locus containing the conserved TnsD-binding site (att*Tn7*), including the expected and alternative orientation *Tn7* transposition products and PCR primer pairs to selectively amplify them. d, PCR analysis of *Tn7* transposition, resolved by agarose gel electrophoresis. Amplification of rssA serves as a loading control. e, Sanger sequencing chromatograms of both upstream and downstream junctions of genomically integrated *Tn7*. f, Genomic organization of the native *V. cholerae* strain HE-45 Tn*6677* transposon. Genes that are conserved between Tn*6677* and the *E. coli Tn7* transposon, and between Tn*6677* and a canonical type I-F CRISPR–Cas system from *P. aeruginosa*28, are highlighted. The cas1 and cas2-3 genes, which mediate spacer acquisition and DNA degradation during the adaptation and interference stages of adaptive immunity, respectively, are missing from CRISPR–Cas systems encoded by *Tn7*-like transposons. Similarly, the tnsE gene, which facilitates non-sequence-specific transposition, is absent. The *V. cholerae* HE-45 genome contains another *Tn7*-like transposon (located within GenBank accession ALED01000025.1), which lacks an encoded CRISPR–Cas system and exhibits low sequence similarity to the Tn*6677* transposon investigated in this study.

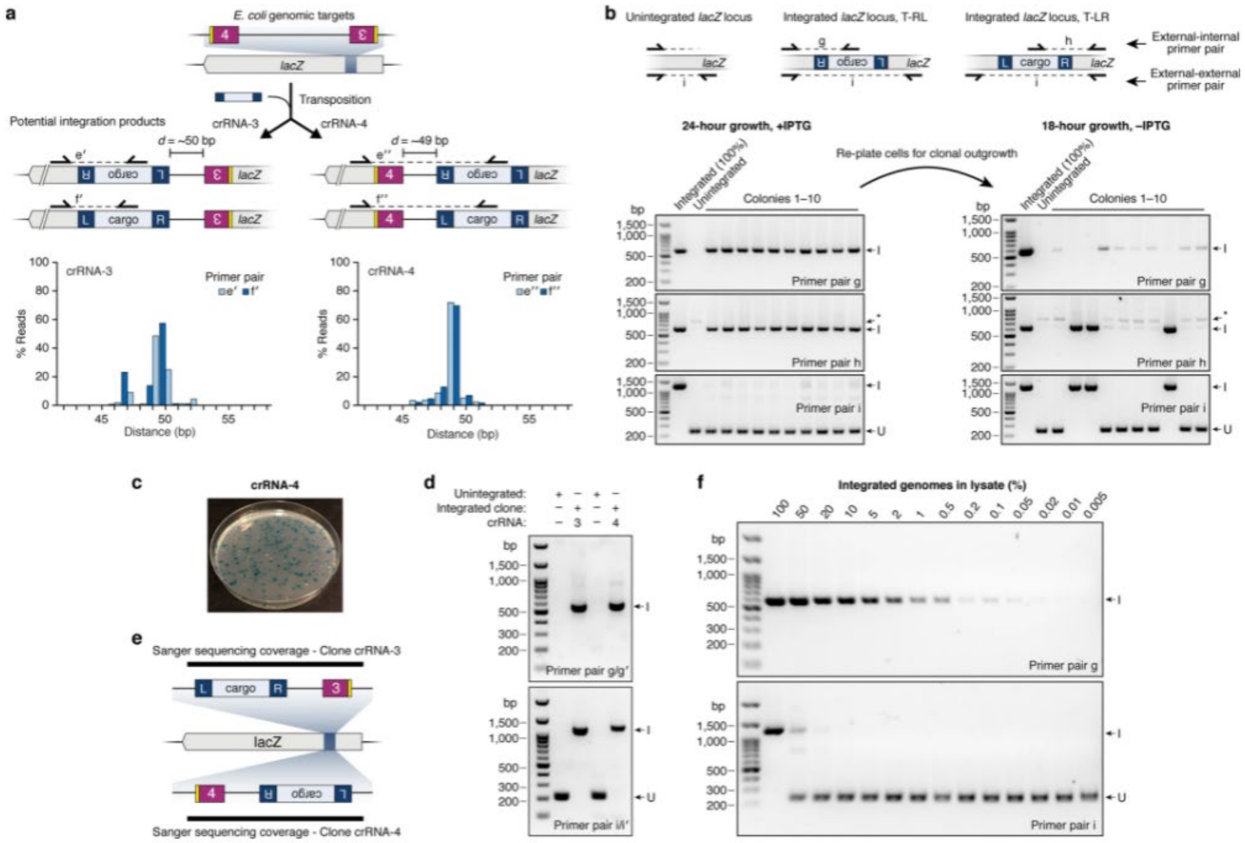
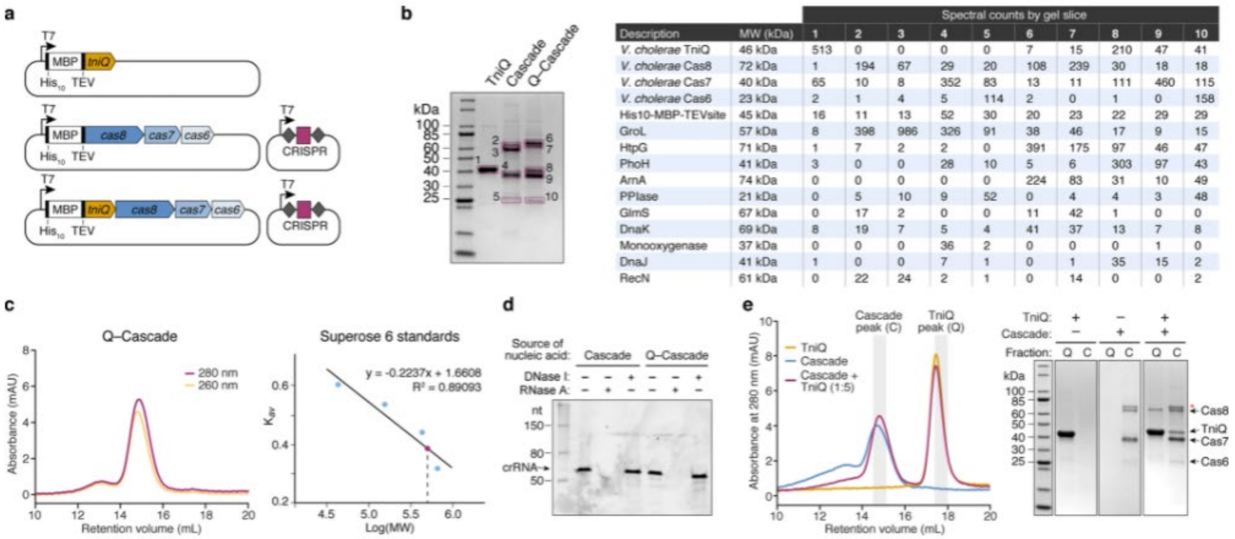


Figure 19: Analysis of *E. coli* cultures and strain isolates containing *lacZ*-integrated transposons

a, Top, genomic locus targeted by crRNA-3 and crRNA-4, including both potential transposition products and the PCR primer pairs to selectively amplify them. Bottom, NGS analysis of the distance between the Cascade target site and transposon insertion site for crRNA-3 (left) and crRNA-4 (right), determined with two alternative primer pairs. b, Top, schematic of the lacZ locus with or without integrated transposon after transposition experiments with crRNA-4. T-LR and T-RL denote transposition products in which the transposon left end and right end are proximal to the target site, respectively. Primer pairs g and h (external–internal) selectively amplify the integrated locus, whereas primer pair i (external–external) amplifies both unintegrated and integrated loci. Bottom, PCR analysis of 10 colonies after 24-h growth on +IPTG plates (left) indicates that all colonies contain integration events in both orientations (primer pairs g and h), but with efficiencies sufficiently low that the unintegrated product predominates after amplification with primer pair i. After resuspending cells, allowing for an additional 18 h of clonal growth on −IPTG plates, and performing the same PCR analysis on 10 colonies (right), 3 out of 10 colonies now exhibit clonal integration in the T-LR orientation (compare primer pairs h and i). The remaining colonies show low-level integration in both orientations, which presumably occurred during the additional 18-h growth owing to leaky expression. These analyses indicate that colonies are genetically heterogeneous after growth on +IPTG plates, and that RNA-guided DNA integration only occurs in a proportion of cells within growing colonies. I, integrated product; U, unintegrated product. Asterisk denotes mispriming product also present in the negative (unintegrated) control. c, Photograph of LB-agar plate used for blue–white colony screening. Cells from IPTG-containing plates were replated on X-gal-containing plates, and white colonies expected to contain lacZ-inactivating transposon insertions were selected for further characterization. d, PCR analysis of *E. coli* strains identified by blue–white colony screening that contain clonally integrated transposons, as in b. e, Schematic of Sanger sequencing coverage across the lacZ locus for strains shown in d. f, PCR analysis of transposition experiment with crRNA-4 after serially diluting lysate from a clonally integrated strain with lysate from a control strain to simulate variable integration efficiencies, as in b. These experiments demonstrate that transposition products can be reliably detected by PCR with an external–internal primer pair at efficiencies above 0.5%, but that PCR bias leads to preferential amplification of the unintegrated product using the external-external primer pair at any efficiency substantially below 100%.



		Spectral counts by gel slice									
Description	MW (kDa)	1	2	3	4	5	6	7	8	9	10
V. cholerae TniQ	46 kDa	513	0	0	0	0	7	15	210	47	41
V. cholerae Cas8	72 kDa	1	194	67	29	20	108	239	30	18	18
V. cholerae Cas7	40 kDa	65	10	8	352	83	13	11	111	460	115
V. cholerae Cas6	23 kDa	2	1	4	5	114	2	0	1	0	158
His10-MBP-TEVsite	45 kDa	16	11	13	52	30	20	23	22	29	29
GroL	57 kDa	8	398	986	326	91	38	46	17	9	15
HtpG	71 kDa	1	7	2	2	0	391	175	97	46	47
PhoH	41 kDa	3	0	0	28	10	5	6	303	97	43
ArnA	74 kDa	0	0	0	0	0	224	83	31	10	49
PPIase	21 kDa	0	5	10	9	52	0	4	4	3	48
GlmS	67 kDa	0	17	2	0	0	11	42	1	0	0
DnaK	69 kDa	8	19	7	5	4	41	37	13	7	8
Monooxygenase	37 kDa	0	0	0	36	2	0	0	0	1	0
DnaJ	41 kDa	1	0	0	7	1	0	1	35	15	2
RecN	61 kDa	0	22	24	2	1	0	14	0	0	2

Figure 20: Analysis of *V. cholerae* Cascade and TniQ-Cascade complexes.

a, Expression vectors for recombinant protein or ribonucleoprotein complex purification. b, Left, SDS–PAGE analysis of purified TniQ, Cascade and TniQ–Cascade complexes, highlighting protein bands excised for in-gel trypsin digestion and mass spectrometry analysis. Right, table listing *E. coli* and recombinant proteins identified from these data, and spectral counts of their associated peptides. Note that Cascade and TniQ–Cascade samples used for this analysis are distinct from the samples presented in Fig. 2. c, Size-exclusion chromatogram of the TniQ–Cascade co-complex on a Superose 6 10/300 column (left), and a calibration curve generated using protein standards (right). The measured retention time of TniQ–Cascade (maroon) is consistent with a complex having a molecular mass of approximately 440 kDa. d, RNase A and DNase I sensitivity of nucleic acids that co-purified with Cascade and TniQ–Cascade, resolved by denaturing urea–PAGE. e, TniQ, Cascade and a Cascade + TniQ binding reaction were resolved by size-exclusion chromatography (left), and indicated fractions were analysed by SDS–PAGE (right). Asterisk denotes an HtpG contaminant.

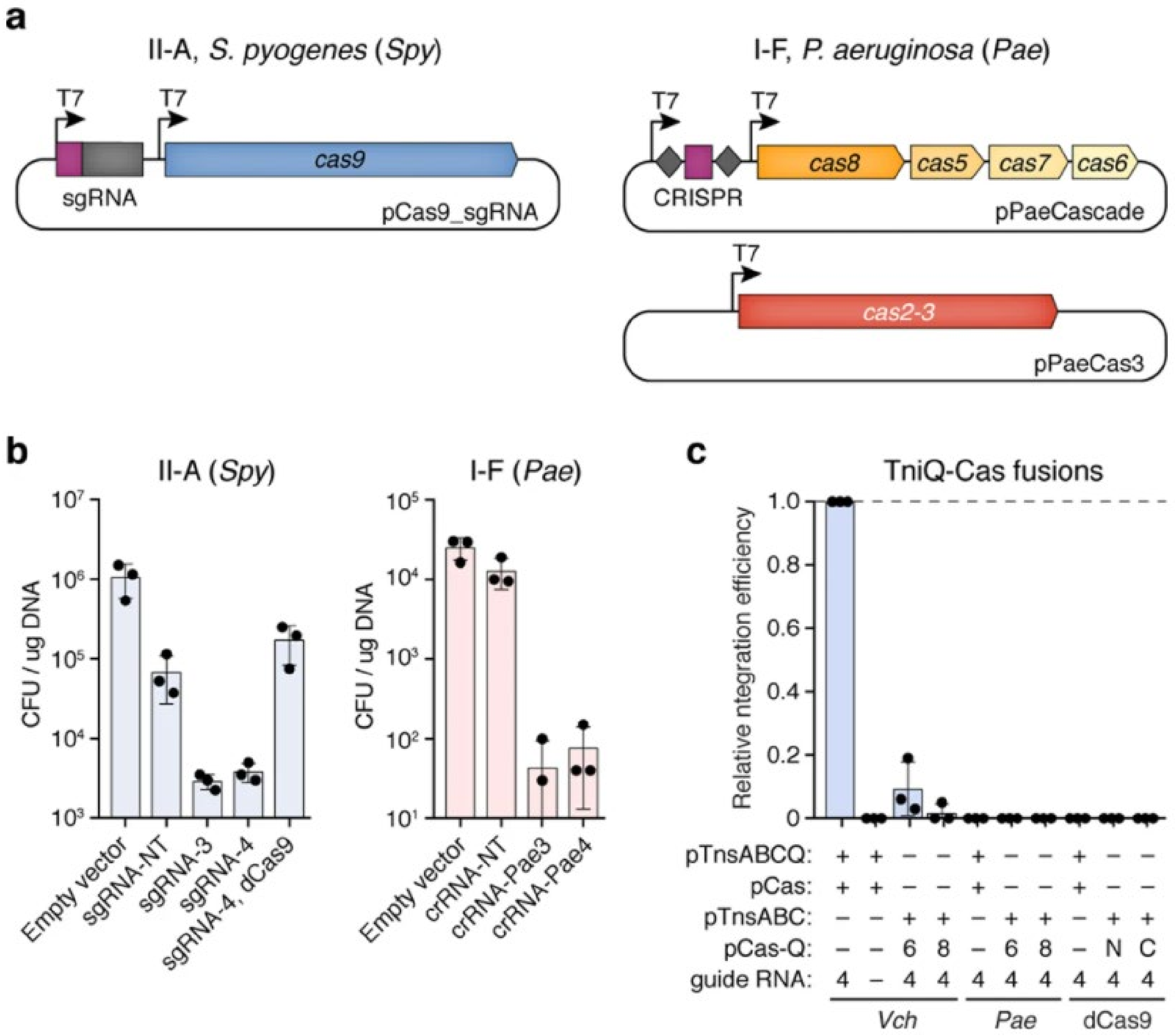


Figure 21: Control experiments demonstrating efficient DNA targeting with Cas9 and *P. aeruginosa* Cascade.

a, Plasmid expression system for *S. pyogenes* (Spy) Cas9-sgRNA (type II-A, left) and *P. aeruginosa* Cascade (PaeCascade) and Cas2-3 (type I-F, right). The Cas2-3 expression plasmid was omitted from experiments described in Figure 14.e, b, Cell killing experiments using *S. pyogenes* Cas9-sgRNA (left) or PaeCascade and Cas2-3 (right), monitored by determining colony-forming units (CFU) after plasmid transformation. Complexes were programmed with guide RNAs that target the same genomic lacZ sites as with *V. cholerae* crRNA-3 and crRNA-4, such that efficient DNA targeting and degradation results in lethality and thus a drop in transformation efficiency. c, qPCR-based quantification of transposition efficiency from experiments using the *V. cholerae* transposon donor and TnsA-TnsB-TnsC, together with DNA targeting components comprising *V. cholerae* Cascade (Vch), *P. aeruginosa* Cascade (Pae) or *S. pyogenes* dCas9–RNA (dCas9). TniQ was expressed either on its own from pTnsABCQ or as a fusion to the targeting complex (pCas-Q) at the Cas6 C terminus (6), Cas8 N terminus (8), or dCas9 N or C terminus. The same sample lysates as in Figure 14.e were used. Data in b and c are shown as mean ± s.d. for n = 3 biologically independent samples.

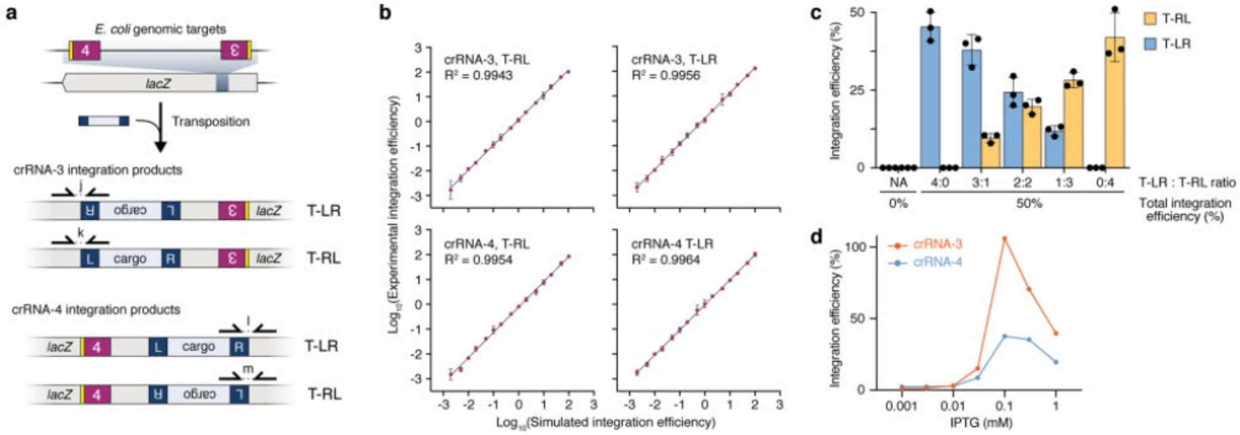


Figure 22: qPCR-based quantification of RNA-guided DNA integration efficiencies

a, Potential lacZ transposition products in either orientation for both crRNA-3 and crRNA-4, and qPCR primer pairs to selectively amplify them. b, Comparison of simulated integration efficiencies for T-LR and T-RL orientations, generated by mixing clonally integrated and unintegrated lysates in known ratios, versus experimentally determined integration efficiencies measured by qPCR. c, Comparison of simulated mixtures of bidirectional integration efficiencies for crRNA-4, generated by mixing clonally integrated and unintegrated lysates in known ratios, versus experimentally determined integration efficiencies measured by qPCR. d, RNA-guided DNA integration efficiency as a function of IPTG concentration for crRNA-3 and crRNA-4, measured by qPCR. Data in b and c are shown as mean ± s.d. for n = 3 biologically independent samples.

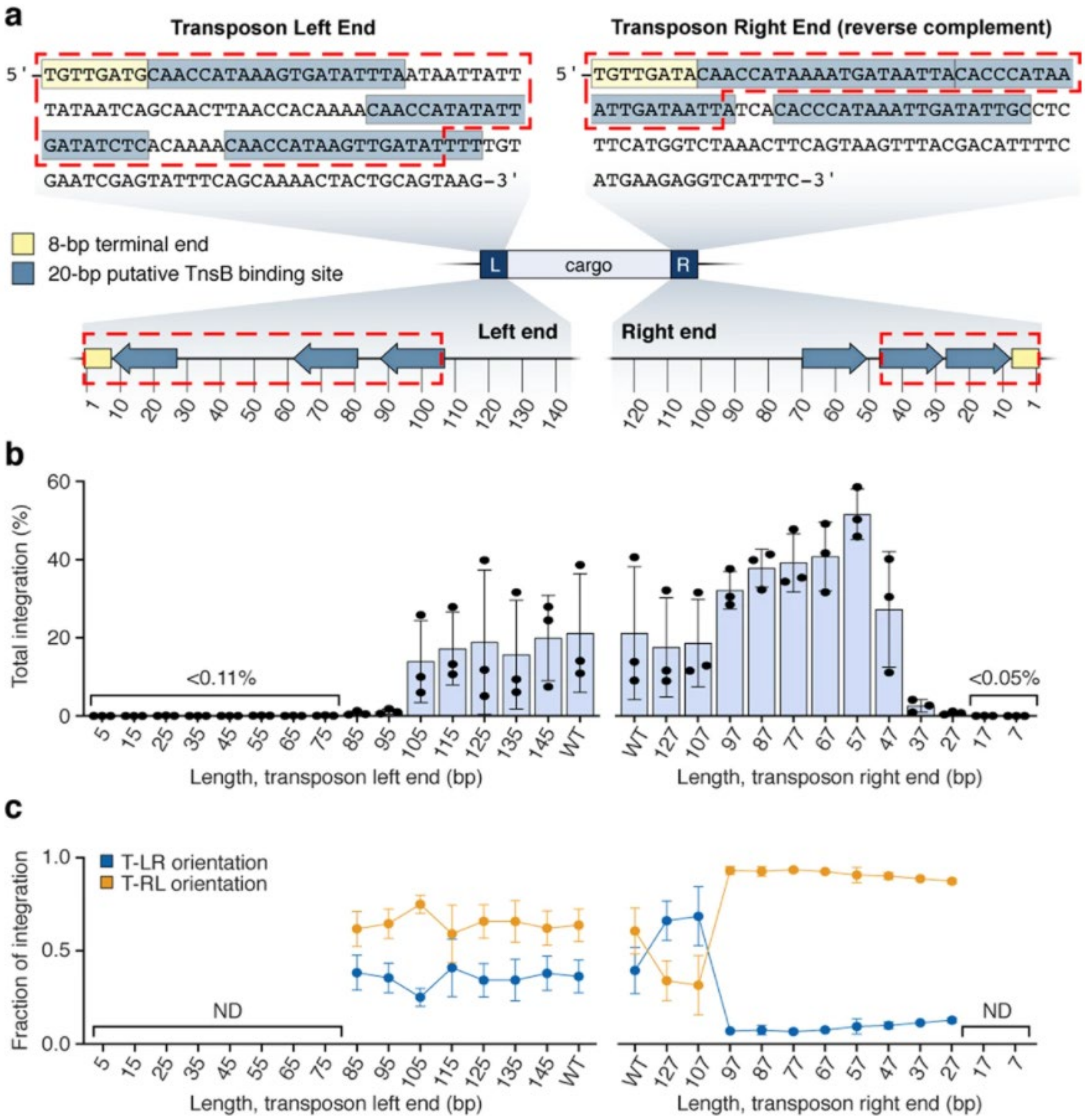


Figure 23: Influence of transposon end sequences on RNA-guided DNA integration

a, Sequence (top) and schematic (bottom) of *V. cholerae* Tn*6677* left- and right-end sequences. The putative TnsB-binding sites (blue) were determined based on sequence similarity to the TnsBbinding sites previously described in 28. The 8-bp terminal ends are shown in yellow, and the empirically determined minimum end sequences required for transposition are denoted by red dashed boxes. b, Integration efficiency with crRNA-4 as a function of transposon end length, as determined by qPCR. c, The relative fraction of both integration orientations as a function of transposon end length, determined by qPCR. ND, not determined. Data in b and c are shown as mean ± s.d. for n = 3 biologically independent samples.

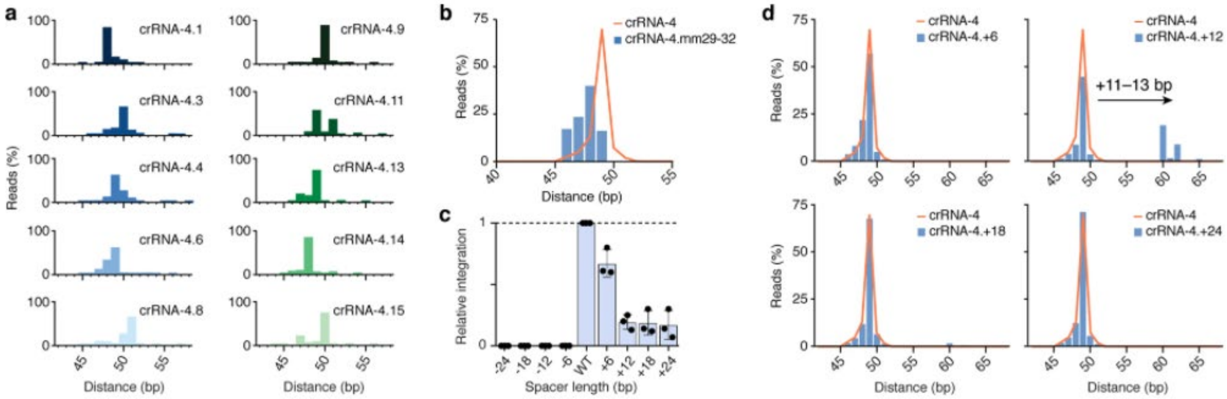


Figure 24: Analysis of RNA-guided DNA integration for PAM-tiled crRNAs and extended spacer length crRNAs

a, Integration site distribution for all crRNAs described in Figure 18.d & Figure 18.e having a normalized transposition efficiency more than 20%, determined by NGS. b, Integration site distribution for a crRNA containing mismatches at positions 29–32, compared with the distribution with crRNA-4, determined by NGS. c, The crRNA-4 spacer length was shortened or lengthened by 6-nucleotide increments, and the resulting integration efficiencies were determined by qPCR. Data are normalized to crRNA-4 and are shown as mean ± s.d. for n = 3 biologically independent samples. d, Integration site distribution for extended length crRNAs compared with the distribution with crRNA-4, determined by NGS.

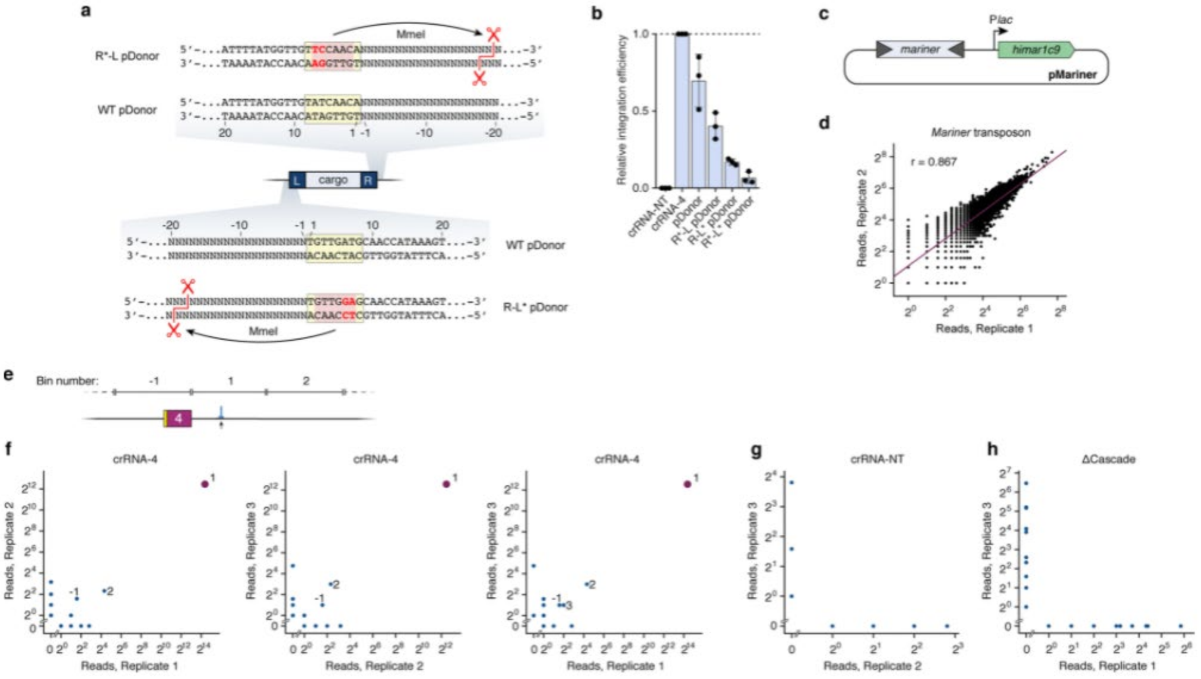


Figure 25: Development and analysis of Tn-seq.

a, Schematic of the *V. cholerae* transposon end sequences. The 8-bp terminal sequence of the transposon is boxed and highlighted in light yellow. Mutations generated to introduce MmeI recognition sites are shown in red letters, and the resulting recognition site is highlighted in red. Cleavage by MmeI occurs 17–19 bp away from the transposon end, generating a 2-bp overhang. b, Comparison of integration efficiencies for the wild-type and MmeI-containing transposon donors, determined by qPCR. Labels on the x axis denote which plasmid was transformed last; we reproducibly observed higher integration efficiencies when pQCascade was transformed last (crRNA-4) than when pDonor was transformed last. The transposon containing an MmeI site in the transposon 'right' end (R∗-L pDonor) was used for all Tn-seq experiments. Data are mean ± s.d. for n = 3 biologically independent samples. c, Plasmid expression system for Himar1C9 and the mariner transposon. d, Scatter plot showing correlation between two biological replicates of Tn-seq experiments with the mariner transposon. Reads were binned by *E. coli* gene annotations, and a linear regression fit and Pearson linear correlation coefficient (r) are shown. e, Schematic of 100-bp binning approach used for Tn-seq analysis of transposition experiments with the *V. cholerae* transposon, in which bin 1 is defined as the first 100 bp immediately downstream (PAM-distal) of the Cascade target site. f, Scatter plots showing correlation between biological replicates of Tn-seq experiments with the *V. cholerae* transposon programmed with crRNA-4. All highly sampled reads fall within bin 1, but we also observed low-level but reproducible, long-range integration into 100-bp bins just upstream and downstream of the primary integration site (bins 1, 2 and 3). g, Scatter plot showing correlation between biological replicates of Tn-seq experiments with the *V. cholerae* transposon programmed with a non-targeting crRNA (crRNA-NT). h, Scatter plot showing correlation between biological replicates of Tn-seq experiments with the *V. cholerae* transposon expressing TnsA-TnsB-TnsC-TniQ but not Cascade. For f–h, bins are only plotted when they contain at least one read in either dataset.

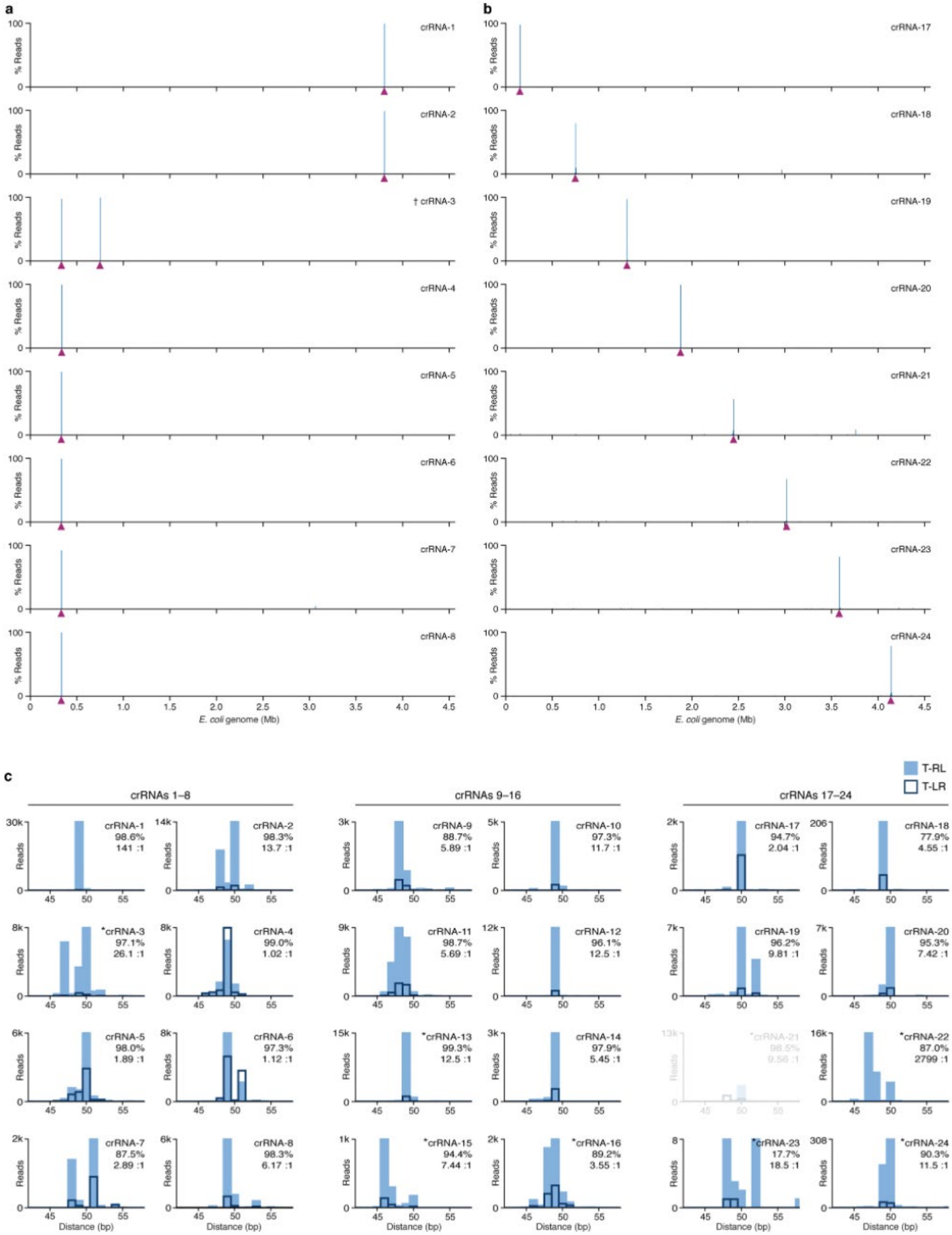

a
b
% Reads
100
0
crRNA-1
crRNA-2
† crRNA-3
crRNA-4
crRNA-5
crRNA-6
crRNA-7
crRNA-8
crRNA-17
crRNA-18
crRNA-19
crRNA-20
crRNA-21
crRNA-22
crRNA-23
crRNA-24
0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
E. coli genome (Mb)
c
T-RL
T-LR
crRNAs 1–8
crRNAs 9–16
crRNAs 17–24
Reads
Distance (bp)
45
50
55
30k
crRNA-1
98.6%
141 :1
14k
crRNA-2
98.3%
13.7 :1
8k
*crRNA-3
97.1%
26.1 :1
8k
crRNA-4
99.0%
1.02 :1
6k
crRNA-5
98.0%
1.89 :1
8k
crRNA-6
97.3%
1.12 :1
2k
crRNA-7
87.5%
2.89 :1
6k
crRNA-8
98.3%
6.17 :1
3k
crRNA-9
88.7%
5.89 :1
5k
crRNA-10
97.3%
11.7 :1
9k
crRNA-11
98.7%
5.69 :1
12k
crRNA-12
96.1%
12.5 :1
15k
*crRNA-13
99.3%
12.5 :1
3k
crRNA-14
97.9%
5.45 :1
1k
*crRNA-15
94.4%
7.44 :1
2k
*crRNA-16
89.2%
3.55 :1
2k
crRNA-17
94.7%
2.04 :1
206
crRNA-18
77.9%
4.55 :1
7k
crRNA-19
96.2%
9.81 :1
7k
crRNA-20
95.3%
7.42 :1
13k
*crRNA-21
98.5%
9.56 :1
16k
*crRNA-22
87.0%
2799 :1
8
*crRNA-23
17.7%
18.5 :1
308
*crRNA-24
90.3%
11.5 :1

Figure 26: Tn-seq data for additional crRNAs tested.

a, b, Genome-wide distribution of genome-mapping Tn-seq reads from transposition experiments with the *V. cholerae* transposon programmed with crRNAs 1–8 (a) and crRNAs 17–24 (b). The location of each target site is denoted by a maroon triangle. Dagger symbol indicates that the lacZ target site for crRNA-3 is duplicated within the λ DE3 prophage, as is the transposon integration site; Tn-seq reads for this dataset were mapped to both genomic loci for visualization purposes only, although we are unable to determine from which locus they derive. c, Analysis of integration site distributions for crRNAs 1–24 determined from the Tn-seq data; the distance between the Cascade target site and transposon insertion site is shown. Data for both integration orientations are superimposed, with filled blue bars representing the T-RL orientation and the dark outlines representing the T-LR orientation. Values in the top-right corner of each graph give the on-target specificity (%), calculated as the percentage of reads resulting from integration within 100 bp of the primary integration site, as compared with the total number of reads aligning to the genome; and the orientation bias (X:Y), calculated as the ratio of reads for the T-RL orientation to reads for the T-LR orientation. Most crRNAs favor integration in the T-RL orientation 49–50 bp downstream of the Cascade target site. crRNA-21 is greyed out because the expected primary integration site is present in a repetitive stretch of DNA that does not allow us to map the reads confidently. Asterisks denote samples for which more than 1% of the genome-mapping reads could not be uniquely mapped.

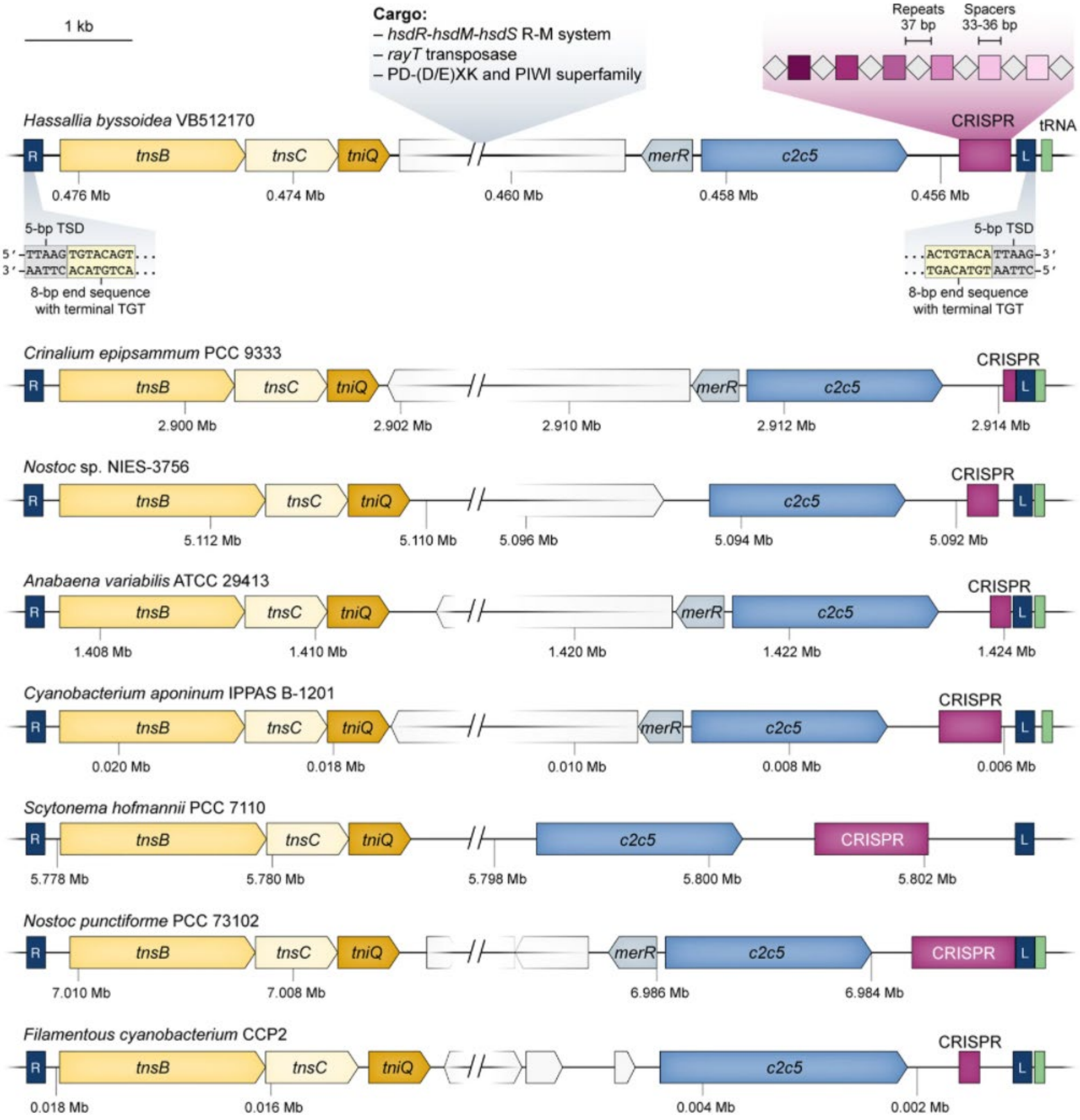


Figure 27: Bacterial transposons also contain type V-U5 CRISPR-Cas systems encoding C2c5.

Representative genomic loci from various bacterial species containing identifiable transposon left and right ends (blue boxes, L and R), genes with homology to tnsB-tnsC-tniQ (shades of yellow), CRISPR arrays (maroon), and the CRISPR-associated gene c2c5 (blue). The example from Hassallia byssoidea (top) highlights the target-site duplication and terminal repeats, as well as genes found within the cargo portion of the transposon. As with the type I CRISPR–Cas system-containing *Tn7*-like transposons, type V CRISPR–Cas system-containing transposons appear to preferentially contain genes associated with innate immune system functions, such as restriction-modification systems. c2c5 genes are frequently flanked by the predicted transcriptional regulator, merR (light blue), and the C2c5-containing transposons appear to usually fall just upstream of tRNA genes (green), a phenomenon that has also been observed for other prokaryotic integrative elements[152,153]. Analysis of 50 spacers from the 8 CRISPR arrays shown with CRISPRTarget[154] revealed 6 spacers with imperfectly matching targets (average of 6 mismatches), none of which mapped to bacteriophages, plasmids, or to the same bacterial genome containing the transposon itself. Whether C2c5 also mediates RNA-guided DNA integration awaits future experimentation.

3.11: Supplementary Figures

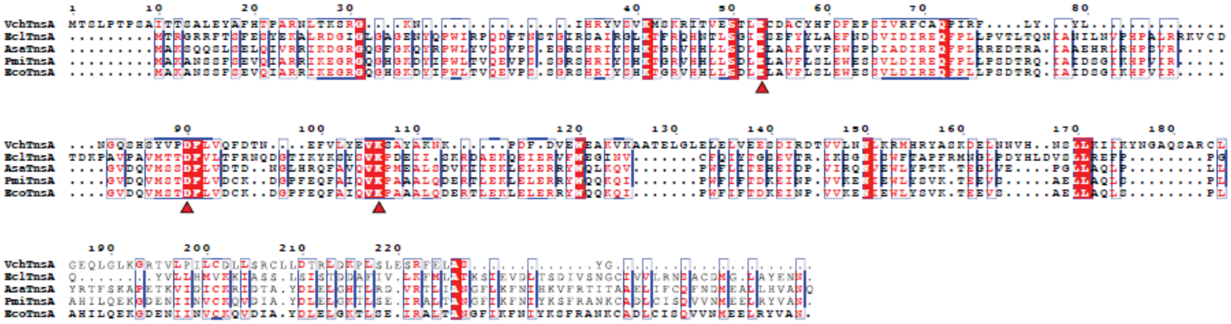


Figure 28: Multiple sequence alignment of TnsA.

Conserved catalytic residues are indicated with red triangles. Vch, *Vibrio cholerae*; Ecl, *Enterobacter cloacae*; Asa, *Aeromonas Salmonicida*; Pmi, *Proteus Mirabilis*; Eco, *Escherichia coli*.

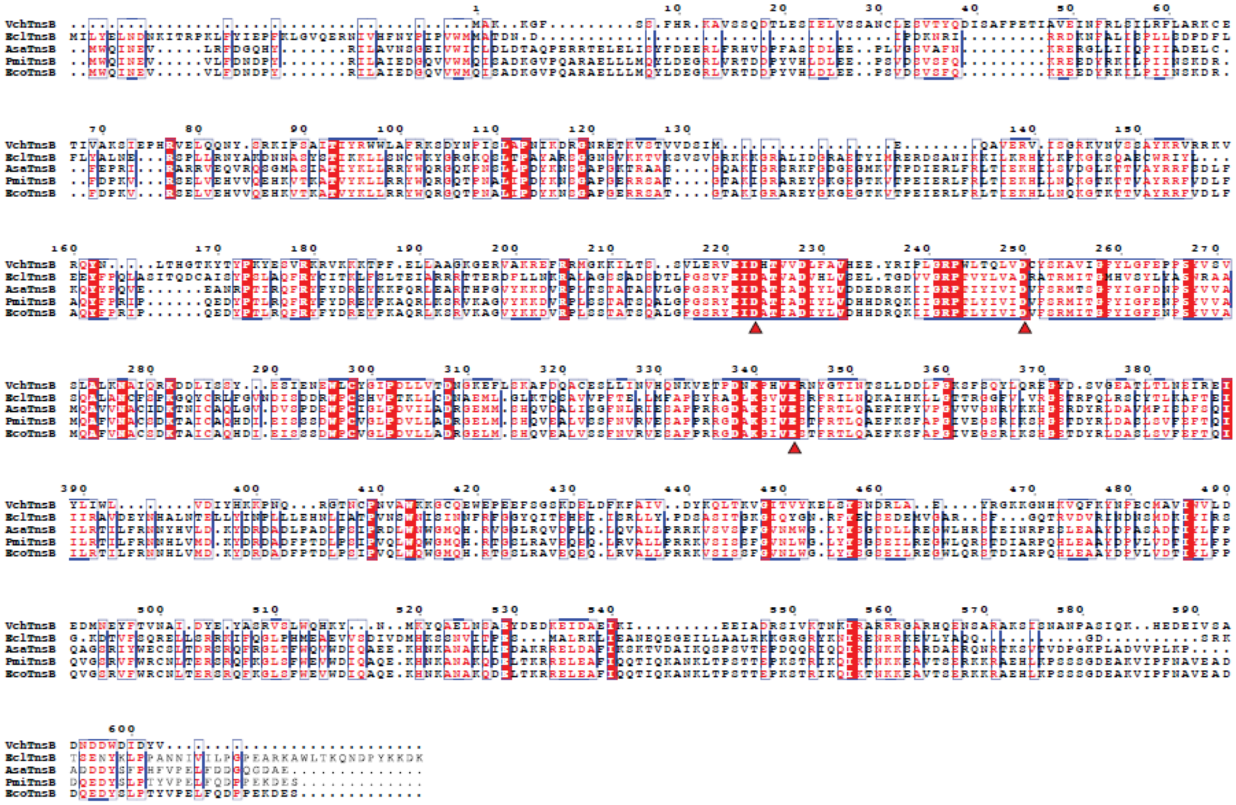


Figure 29: Multiple sequence alignment of TnsB.

Conserved catalytic residues of the DDE motif are indicated with red triangles. Vch, *Vibrio cholerae*; Ecl, *Enterobacter cloacae*; Asa, *Aeromonas Salmonicida*; Pmi, *Proteus Mirabilis*; Eco, *Escherichia coli*.

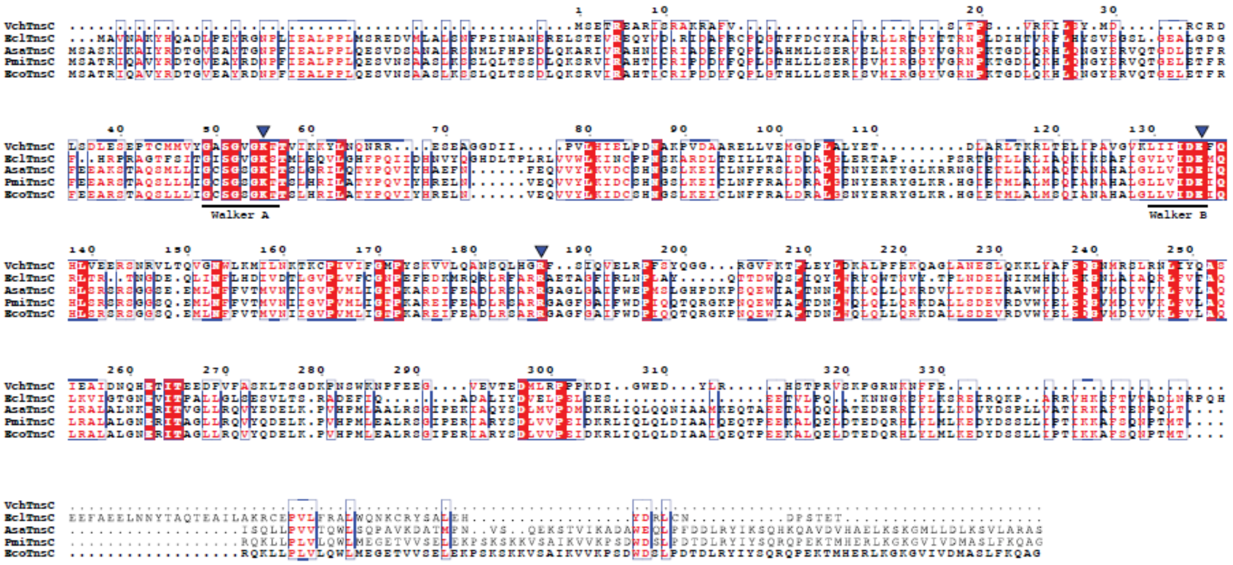


Figure 30: Multiple sequence alignment of TnsC.

Walker A and Walker B motifs characteristic of AAA+ ATPases are indicated, and active site residues involved in ATPase activity are indicated with blue triangles. Some TnsC homologs are annotated as TniB. Vch, *Vibrio cholerae*; Ecl, *Enterobacter cloacae*; Asa, *Aeromonas Salmonicida*; Pmi, *Proteus Mirabilis*; Eco, *Escherichia coli*.

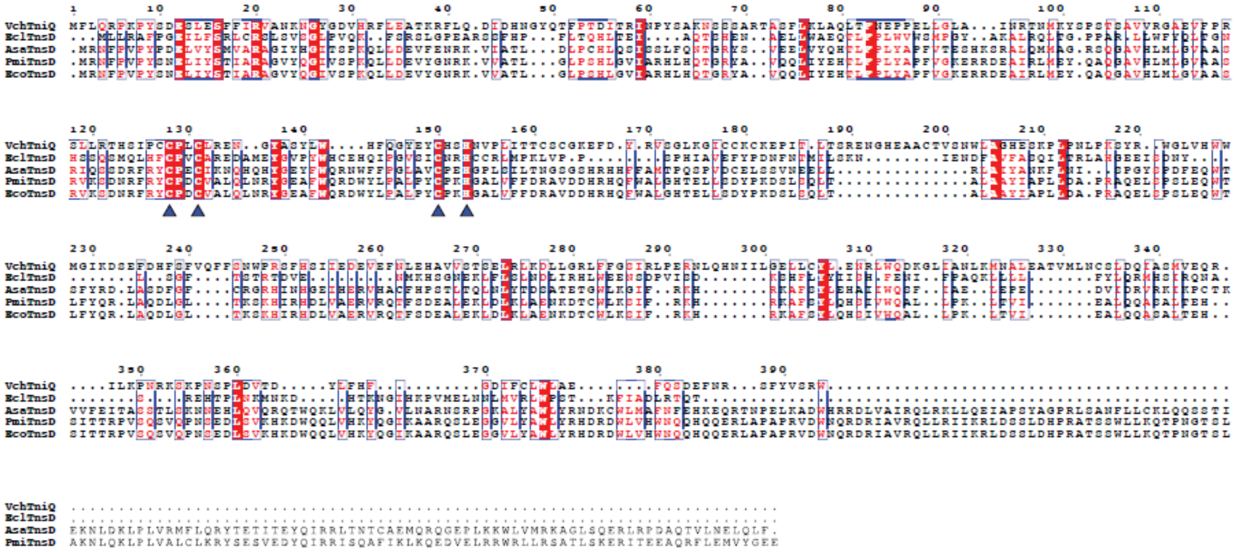


Figure 31: Multiple sequence alignment of TniQ/TnsD.

VchTniQ is aligned to members of the TniQ/TnsD family. Conserved zinc finger motif residues are indicated with blue arrows. Vch, *Vibrio cholerae*; Ecl, *Enterobacter cloacae*; Asa, *Aeromonas Salmonicida*; Pmi, *Proteus Mirabilis*; Eco, *Escherichia coli*.

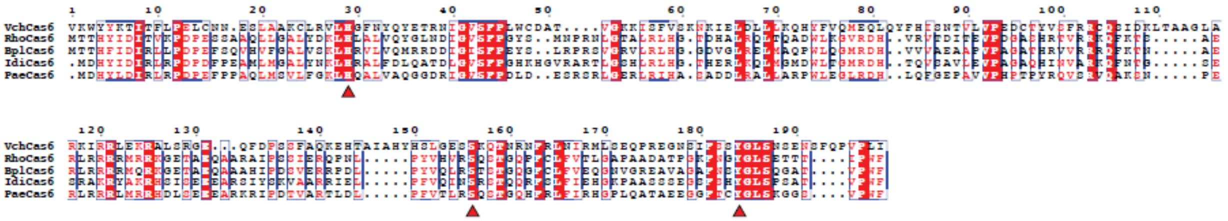


Figure 32: Multiple sequence alignment of Cas6.

VchCas6 is aligned to other I-F Cas6 proteins, which are often annotated as Cas6f or Csy4. Conserved catalytic residues are indicated with red arrows. Vch, *Vibrio cholerae*; Rho, *Rhodanobacter* sp; Bpl, *Burkholderia plantarii*; Idi, *Idiomarina* sp. H105; Pae, *Pseudomonas aeruginosa*.

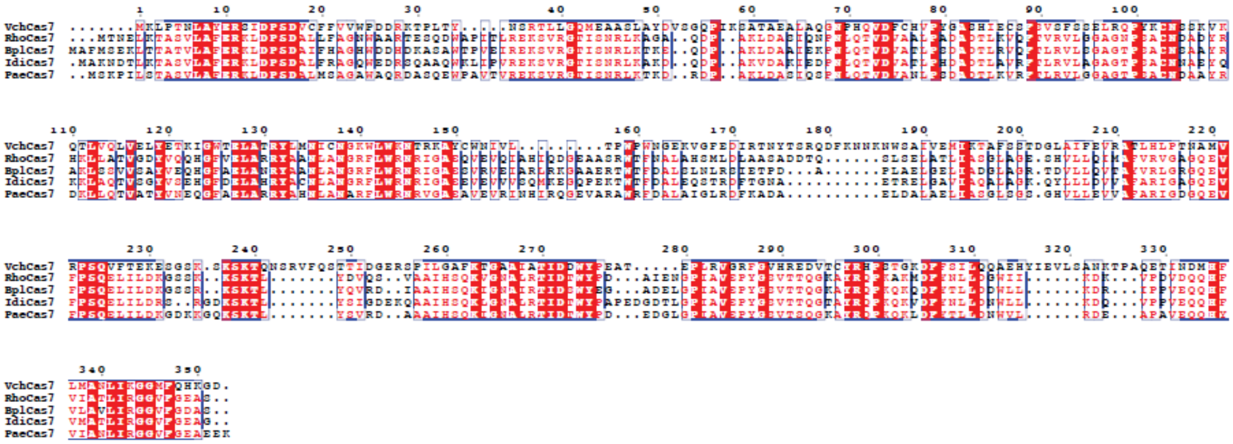


Figure 33: Multiple sequence alignment of Cas7.

VchCas6 is aligned to other I-F Cas7 proteins, which are often annotated as Csy3. Conserved catalytic residues are indicated with red arrows. Vch, *Vibrio cholerae*; Rho, *Rhodanobacter* sp; Bpl, *Burkholderia plantarii*; Idi, *Idiomarina* sp. H105; Pae, *Pseudomonas aeruginosa*.

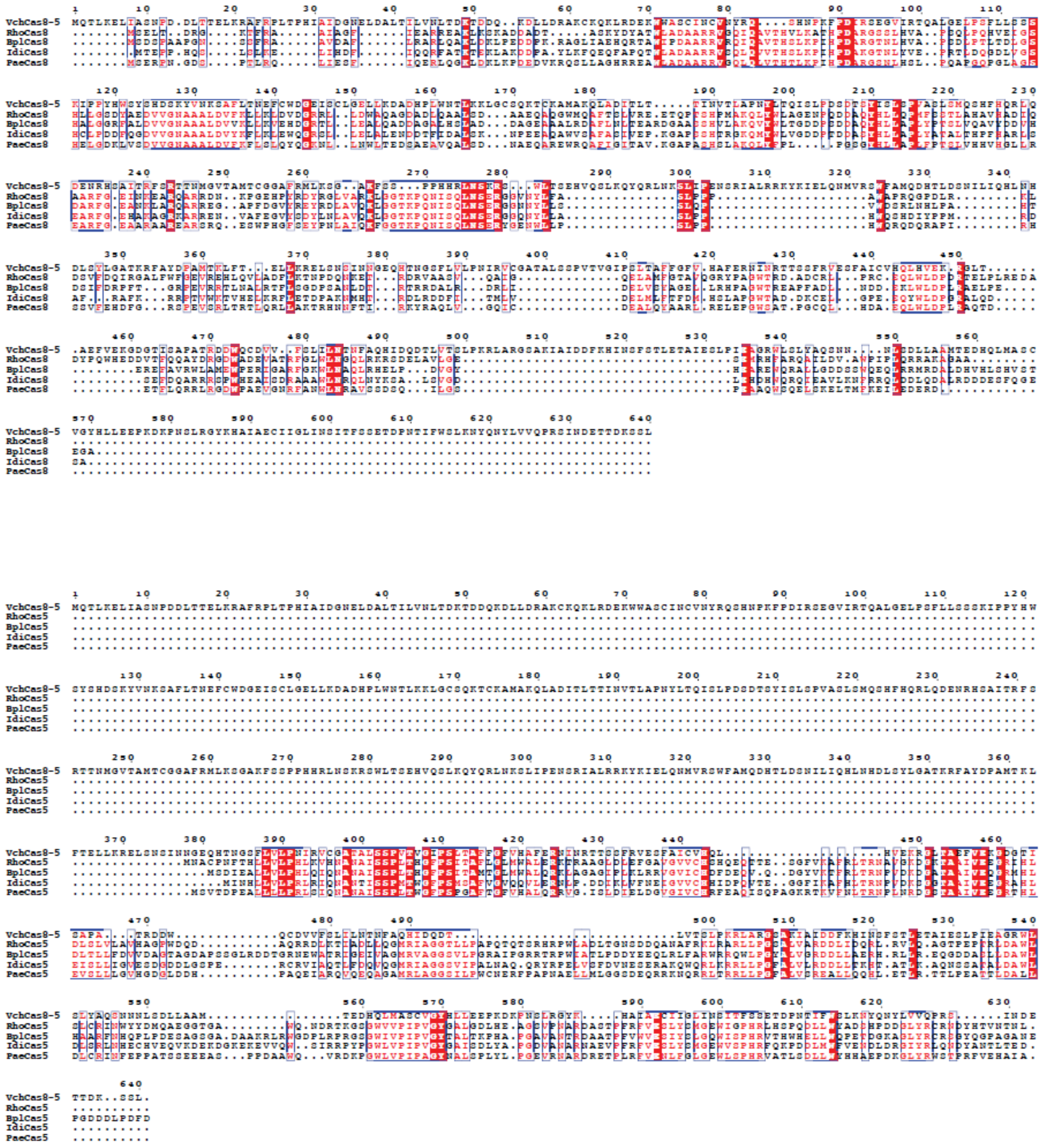


Figure 34: Multiple sequence alignment of Cas8 and Cas5.

VchCas8, a natural Cas8-Cas5 fusion protein, is aligned to other I-F Cas8 proteins (top, which are often annotated as Csy1, and to other I-F Cas5 proteins (bottom_, which are often annotated as Csy2. Vch, *Vibrio cholerae*; Rho, *Rhodanobacter* sp; Bpl, *Burkholderia plantarii*; Idi, *Idiomarina* sp. H105; Pae, *Pseudomonas aeruginosa*.

Chapter 4: Structural Basis of DNA Targeting by a Transposon-Encoded CRISPR-Cas System

This chapter has been adapted from:

Halpin-Healy, T.S., Klompe, S.E., Sternberg, S.H. & Fernandez, I.F. Structural basis of DNA targeting by a transposon-encoded CRISPR–Cas system. *Nature* 577, 271–274 (2020).

Appendix B contains the paper as published.

Contributions:

All authors conceived and designed the project. T.S.H.-H. purified ribonucleoprotein complexes and assisted in cryo-EM data acquisition. I.S.F. collected EM data and solved the structures. I.S.F., S.H.S. and the other authors discussed the data and wrote the manuscript.

4.1: Abstract

Bacteria use adaptive immune systems encoded by CRISPR and Cas genes to maintain genomic integrity when challenged by pathogens and mobile genetic elements[64,111,112]. Type I CRISPR–Cas systems typically target foreign DNA for degradation via joint action of the ribonucleoprotein complex Cascade and the helicase–nuclease Cas3[155,156], but nuclease-deficient type I systems lacking Cas3 have been repurposed for RNA- guided transposition by bacterial *Tn7*-like transposons[32,157]. How CRISPR- and transposon-associated machineries collaborate during DNA targeting and insertion remains unknown. Here we describe structures of a TniQ–Cascade complex encoded by the Vibrio cholerae Tn*6677* transposon using cryo-electron microscopy, revealing the mechanistic basis of this functional coupling. The cryo-electron microscopy maps enabled de novo modelling and refinement of the transposition protein TniQ, which binds to the Cascade complex as a dimer in a head-to-tail configuration, at the interface formed by Cas6 and Cas7 near the 3′ end of the CRISPR RNA (crRNA). The natural Cas8–Cas5 fusion protein binds the 5′ crRNA handle and contacts the TniQ dimer via a flexible insertion domain. A target DNA-bound structure reveals critical interactions necessary for protospacer-adjacent motif recognition and R-loop formation. This work lays the foundation for a structural understanding of how DNA targeting by TniQ–Cascade leads to downstream recruitment of additional transposase proteins, and will guide protein engineering efforts to leverage this system for programmable DNA insertions in genome-engineering applications.

4.2: Introduction

We previously demonstrated that a transposon derived from *V. cholerae* Tn*6677* undergoes programmable transposition in Escherichia coli directed by a crRNA, and that this activity requires four transposon- and three CRISPR-associated genes in addition to a CRISPR array[157] (Figure 35a). Whereas TnsA, TnsB and TnsC exhibit functions that are consistent with their homologs from the related and well-studied cut-and-paste DNA transposon *E. coli Tn7*[119], we showed that TniQ, a homolog of *E. coli* TnsD, forms a co-complex with the Cascade ribonucleoprotein complex encoded by the type I-F variant CRISPR–Cas system. This finding suggested an alternative role for TniQ, compared with the role of *E. coli* TnsD in identifying target sites during *Tn7* transposition. We proposed that RNA-guided DNA targeting by Cascade could deliver TniQ to DNA in a manner compatible with downstream transpososome formation, and that TniQ might interact with Cascade near the 3′ end of the crRNA, consistent with RNA-guided DNA insertion occurring around 49 bp down-stream of the protospacer-adjacent motif (PAM)-distal edge of the target site. To determine this unambiguously, we purified the *V. cholerae* TniQ– Cascade complex loaded with a native crRNA and determined its structure by cryo-electron microscopy (cryo-EM) (Supplementary Table 1)

4.3: Cryo-EM Structure of TniQ-Cascade Complex

The overall complex adopts a helical architecture with protuberances at both ends (Figure 35, Figure 39, and 40). The global architecture is similar to previously determined structures of Cascade from I-E and I-F systems[89,158–160] (Figure 41), with the exception of a large mass of additional density attributable to TniQ (Figure 35c). Maximum likelihood classification methods implemented in Relion3[161] enabled us to identify marked dynamics in the entire complex, which appears to ‘breathe’, widening and narrowing the distance between the two protuberances (Figure 39d). The large subunit encoded by a natural Cas8–Cas5 fusion protein (hereafter referred to simply as Cas8) forms one protuberance and recognizes the 5′ end of the crRNA via base- and backbone-specific contacts (Figure 42, 43a–c, 44a), similar to the canonical roles of Cas8 and Cas5 (Figure 41). Cas8 contains two primary subdomains formed mainly by α-helices and a third domain of approximately 100 residues (residues 277 to 385) that is predicted to form three α-helices but could not be built in our maps owing to its intrinsic flexibility (Figure 35c). However, low-pass-filtered maps revealed that this flexible domain connects with the TniQ protuberance at the opposite end of the crescent-shaped complex (Figure 40e). Additionally, there seemed to be a loose coupling between the Cas8 flexible domain and overall breathing of the complex, as stronger density for that domain could be observed in the closed state (Figure 40d).

Six Cas7 subunits protect much of the crRNA by forming a helical filament along its length (Figure 35b, d), similar to other type I Cascade complexes[88,89,127,158] (Figure 41). A 'finger' motif in Cas7 clamps the crRNA at regular intervals, causing every sixth nucleotide (nt) of the 32-nt spacer to flip out while leaving the flanking nucleotides available for DNA recognition (Figure 42f, 44). These bases are pre-ordered in short helical segments, with a conserved phenylalanine stacking below the first base of every segment. Cas7.1, the monomer furthest away from Cas8, interacts with Cas6 (also known as Csy4), which is the RNase responsible for processing of the precursor RNA transcript derived from the CRISPR locus. The Cas6–Cas7.1 interaction is mediated by a β-sheet formed by the contribution of a β-strand from Cas6 and the two β-strands that form the finger of Cas7.1 (Figure 43f). Cas6 also forms extensive interactions with the conserved stem-loop in the repeat-derived 3′ crRNA handle (Figure 35, Figure 43d, e), with an arginine-rich α-helix (residues 110 to 128) docked in the major groove, positioning multiple basic residues within interaction distance of the negatively charged RNA backbone.

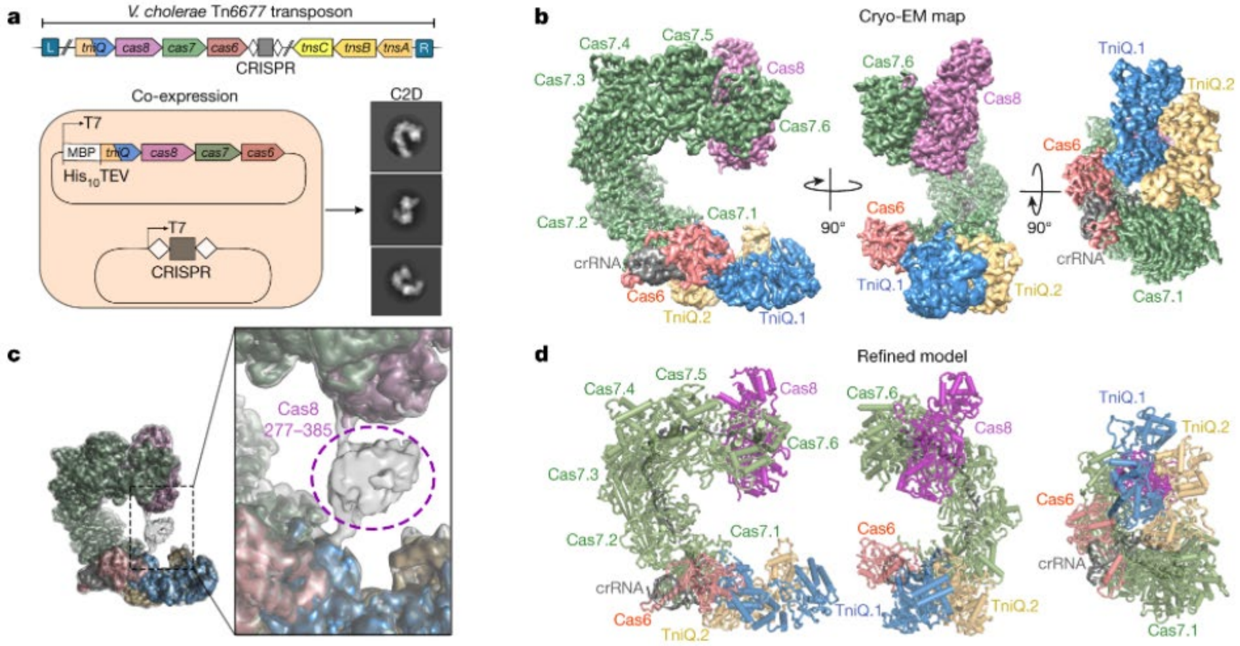


Figure 35: Overall architecture of the *V. cholerae* TniQ-Cascade complex

a, Genetic architecture of the Tn*6677* transposon (top), and plasmid constructs used to express and purify the TniQ–Cascade complex. Right, selected cryo-EM reference-free two-dimensional class averages in multiple orientations. **b**, Orthogonal views of the cryo-EM map of the TniQ–Cascade complex, showing Cas8 (purple), six Cas7 monomers (green), Cas6 (salmon), crRNA (grey) and TniQ monomers (blue, yellow). The complex adopts a helical architecture with protuberances at both ends. **c**, A flexible domain in Cas8 comprising residues 277–385 (grey) could only be visualized in low-pass-filtered maps. The unsharpened map is shown as semi-transparent, grey map overlaid on the post-processed map segmented and colored as in a. **d**, Refined model for the TniQ–Cascade complex derived from the cryo-EM maps shown in **b**.

The interaction established between Cas6 and Cas7.1 forms a continuous surface on which TniQ is docked, forming the other protuberance of the crescent. The intrinsic flexibility of the complex resulted in lower local resolutions in this area of the maps, which we overcame using local alignments masking the area comprising TniQ, Cas6, Cas7.1 and the crRNA handle (Figure 45). The enhanced maps enabled de novo modelling and refinement of TniQ, for which no previous structure or homology model has been reported, to our knowledge (Figure 36). Notably, TniQ binds to Cascade as a dimer with head-to-tail configuration (Figure 36), a surprising result given the expectation that *E. coli* TnsD functions as a monomer during *Tn7* transposition[162].

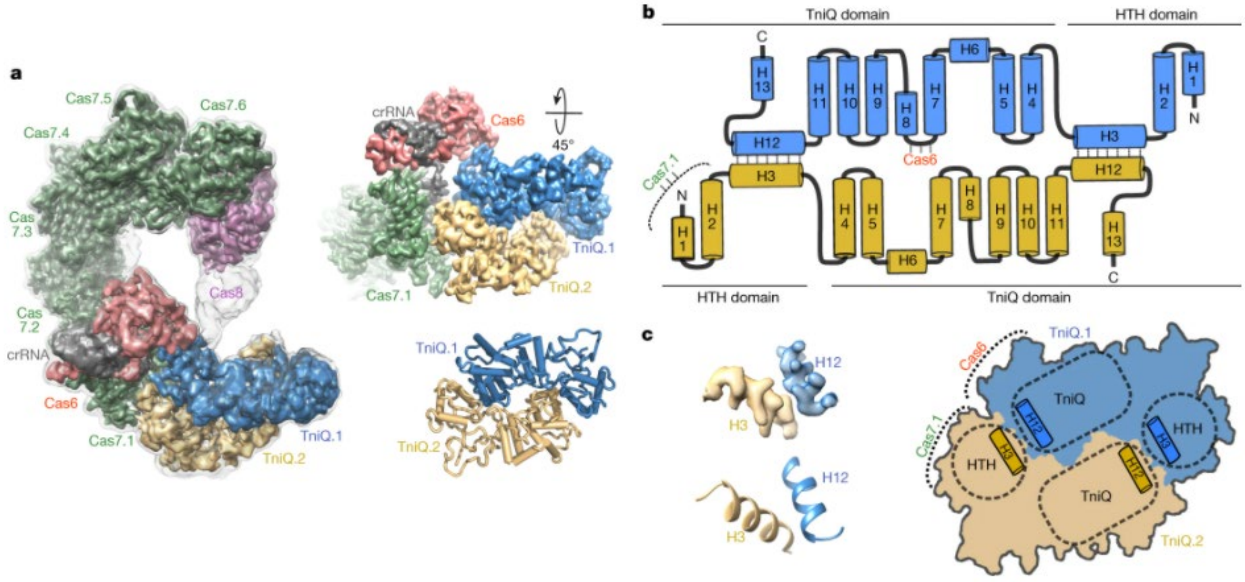


Figure 36: TniQ binds Cascade in a dimeric, head-to-tail configuration.

a, Left, overall view of the TniQ–Cascade cryo-EM unsharpened map (grey) overlaid on the post-processed map segmented and colored as in Figure 35. Right, cryo-EM map (top) and refined model (bottom) of the TniQ dimer. The two monomers interact with each other in a head-to-tail configuration and are anchored to Cascade by Cas6 and Cas7.1. **b**, Secondary structure diagram of the TniQ dimer: thirteen α-helices are organized into an N-terminal HTH domain and a C-terminal TniQ domain. Dimer interactions between H3 and H12 are indicated, as are interaction sites with Cas6 and Cas7.1. **c**, Cryo-EM density for the H3–H12 interaction shows clear side-chain features (top), allowing accurate modelling of the interaction (bottom). **d**, Schematic of the dimer interaction, showing the important dimerization interface between the HTH and TniQ domain.

4.4: TniQ Binds to Cascade as a dimer

TniQ is composed of two domains: an N-terminal domain of approximately 100 residues formed by three short α-helices and a second, larger domain of approximately 300 residues with a signature sequence for the TniQ family. A DALI[163] search using the refined TniQ model as a probe yielded marked structural similarity of the N-terminal domain to proteins containing helix–turn–helix (HTH) domains. This domain is often involved in nucleic acid recognition; however, there are examples where it has been re-purposed for protein–protein interactions[164]. The remaining C-terminal TniQ domain is formed by ten α-helices of variable length and is predicted to contain two tandem zinc finger motifs, although this region was poorly defined in the maps (Figure 36). Overall, the double domain composition of TniQ results in an elongated structure, bent at the junction of the HTH and the TniQ domain (Figure 36). The HTH domain of one monomer engages the TniQ domain of the other monomer via interactions between α-helix 3 (H3) and α-helix 12 (H12), respectively, in a tight protein–protein interaction (Figure 36c). This reciprocal interaction is complemented by multiple interactions established between the TniQ domains from both monomers (up to 45 non-covalent interactions as reported by PISA[165]).

Tethering of the TniQ dimer to Cascade is accomplished by specific interactions established with both Cas6 and Cas7.1 (Figure 37). One monomer of TniQ interacts with Cas6 via its C-terminal TniQ domain, whereas the other TniQ monomer contacts Cas7.1 through its N-terminal HTH domain (Figure 36b, 37). The loop connecting α-helices H7 and H8 of the TniQ domain of the first TniQ monomer is inserted in a hydrophobic cavity formed at the interface of two α-helices of Cas6 (Figure 37b, d). The TniQ histidine residue 265 is involved in rearranging the hydrophobic loop connecting H7 and H8 (Figure 37d), which is inserted in the hydrophobic pocket of Cas6 formed by residues L20, Y74, M78, Y83 and F84. The buried surface in the Cas6–TniQ.1 interaction interface has an area of 420 Å^2. The HTH domain of the other TniQ monomer interacts with Cas7.1 through a network of interactions established mainly by α-helix H2 and the linker connecting H2 and H3, burying a surface area of 595 Å^2 (Figure 37c, e). Thus, the HTH domain and the TniQ domain exert dual roles to drive TniQ dimerization and dock onto Cascade.

The aggregate buried surface area for the TniQ–Cascade interaction is 1,015 Å^2, significantly smaller than other Cascade–effector interactions such as with the nuclease Cas3, in which 2,433 Å^2 is buried[166]. This difference is not surprising given the flexibility observed for the TniQ dimer in its association with Cascade.

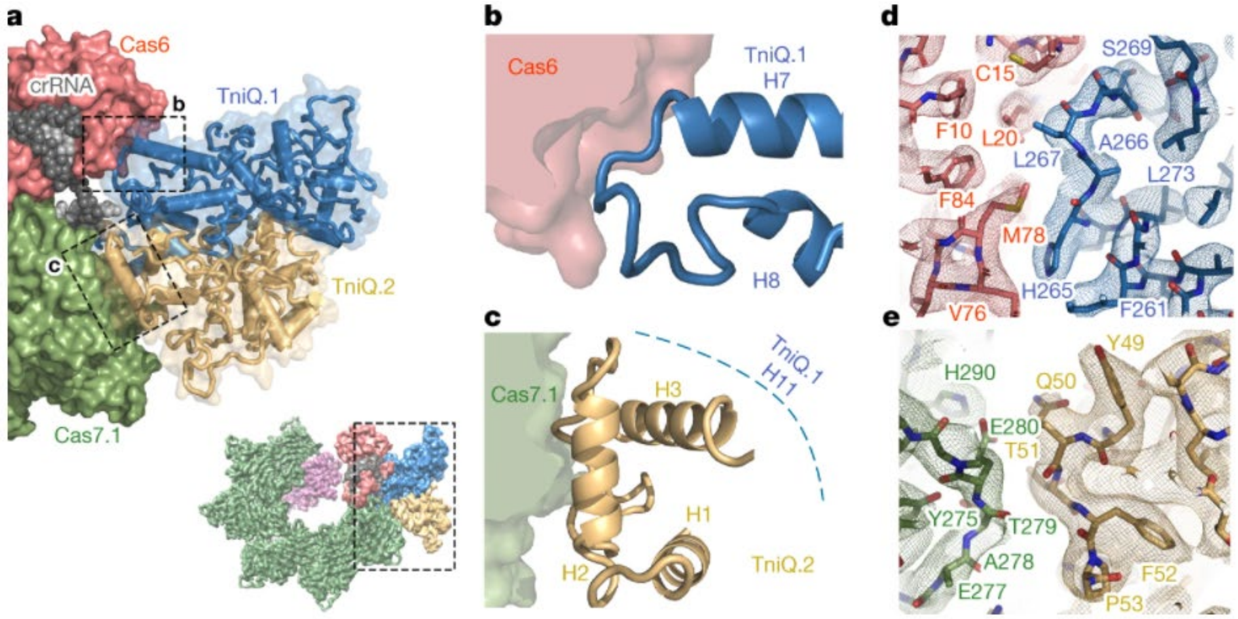


Figure 37: Cas6 and Cas7.1 form a binding platform for TniQ

a, Top, magnified area showing the interaction site of Cascade and the TniQ dimer. Cas6 and Cas7.1 are displayed as molecular Van der Waals surfaces, the crRNA is shown as grey spheres and the TniQ monomers are shown as ribbons. **b**, The loop connecting TniQ.1 α-helices H7 and H8 (blue) binds within a hydrophobic cavity of Cas6. **c**, Cas7.1 interacts with the HTH domain of the TniQ.2 monomer (yellow), mainly through H2 and the loop connecting H2 and H3. **d, e**, Experimental cryo-EM densities observed for the TniQ–Cas6 (**d**) and TniQ–Cas7.1 (**e**) interactions.

4.5: Structure of the DNA-bound TniQ-Cascade Complex

To investigate the structural determinants of DNA recognition by the TniQ–Cascade complex, we determined the structure of the complex bound to a double-stranded DNA (dsDNA) substrate containing the 32-bp target sequence, 5′-CC-3′ PAM, and 20 bp of flanking dsDNA on both ends (Figure 38, 46). Density for 28 nucleotides of the target strand and 8 nucleotides of the non-target strand could be confidently assigned in the reconstructed maps (Figure 38c). As with previous I-F Cascade structures, Cas8 recognizes the double-stranded PAM within the minor groove[127] (Figure 47), and an arginine residue (R246) establishes a stacking interaction with a guanine nucleotide on the target strand, which acts as a wedge to separate the double-stranded PAM from the neighboring unwound DNA where base-pairing with the crRNA begins (Figure 38c).

Twenty-two nucleotides of the target strand within the 32-bp target showed clear density, but surprisingly, the terminal nine nucleotides were not ordered. The target-strand base pairs with the spacer region of the crRNA in short, discontinuous helical segments, as observed previously for I-E and I-F DNA-bound Cascade complexes[88,127], with every sixth base flipped out of the heteroduplex by the insertion of a Cas7 finger (Figure 44b). The observed 22-bp heteroduplex is stabilized by the four Cas7 monomers proximal to the PAM (Cas7.6–Cas7.3), but even after local masked refinements, no density could be observed for any target strand nucleotides that would base-pair with the 3′ end of the crRNA spacer bound by Cas7.2 and Cas7.1. These two Cas7 monomers are proximal to Cas6 and in the region previously described to exhibit dynamics owing to the interaction of the Cas8 flexible domain with the inner face of the TniQ dimer. In addition, the disordered nucleotides also correspond to positions 25–28 of the target site where RNA–DNA mismatches are detrimental for RNA-guided DNA integration[157]. Thus, we propose that the partial R-loop structure that we observed could represent an intermediate conformation refractory to integration, and that further structural rearrangements may be critical for further stabilization of an open conformation, possibly driven by recruitment of the TnsC ATPase.

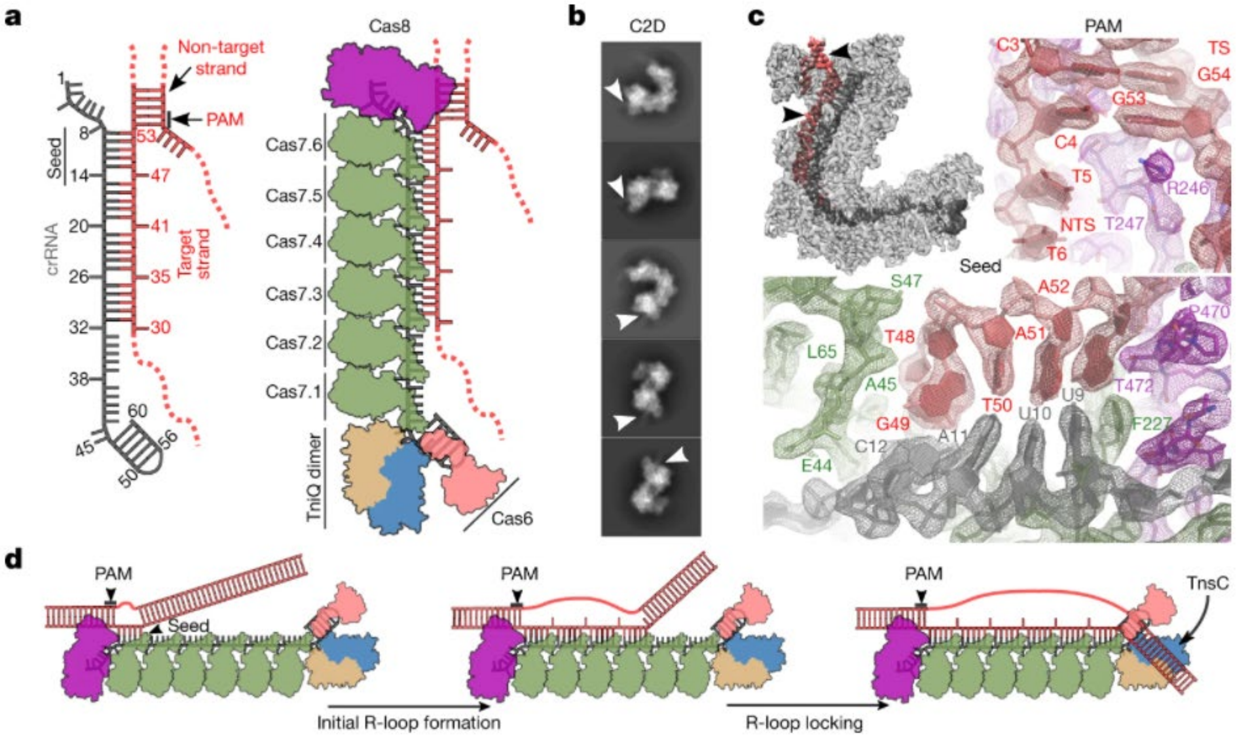


Figure 38: DNA-bound structure of the TniQ-Cascade complex.

a, Schematic of crRNA and the portion of the dsDNA substrate that was experimentally observed within the electron density map for DNA-bound TniQ–Cascade. The target strand, non-target strand, PAM and seed regions are indicated (left); protein components are shown on the right. **b**, Selected cryo-EM reference-free two-dimensional class averages for DNA-bound TniQ–Cascade; density corresponding to dsDNA could be directly observed protruding from the Cas8 component in the two-dimensional class averages (white arrows). **c**, Cryo-EM map for DNA-bound TniQ–Cascade. The crRNA is shown in dark grey and the DNA is shown in red. Right bottom, detailed views of the PAM and seed-recognition regions of the map, with refined models represented as sticks within the electron density. Cas8 is shown in purple, Cas7 is shown in green, the crRNA is in grey and DNA is shown in red. NTS, non-target strand. **d**, The *V. cholerae* transposon encodes a TniQ–Cascade co-complex that uses the sequence content of the crRNA to bind complementary DNA target sites. We propose that the incomplete R-loop observed in our structure (middle) represents an intermediate state that may precede a downstream 'locking' step involving proofreading of the RNA–DNA complementarity. TniQ is positioned at the PAM-distal end of the DNA-bound Cascade complex, where it probably interacts with TnsC during downstream steps of RNA-guided DNA insertion.

4.6: Discussion

Here we present cryo-EM structures of a CRISPR–Cas effector complex bound to the transposition protein TniQ, with and without target DNA. These structures reveal the unexpected presence of TniQ as a dimer that forms bipartite interactions with Cas6 and Cas7.1 within the Cascade complex, forming a probable recruitment platform for downstream-acting transposition proteins[167] (Figure 38d). Our structures further reveal a possible fidelity checkpoint, whereby formation of a complete R-loop requires conformational rearrangements that may depend on extensive RNA–DNA complementarity and/or downstream factor recruitment; this proofreading step could account for the highly specific RNA-guided DNA integration that we previously reported for the *V. cholerae* transposon[157]. In light of recent work demonstrating exaptation of type V-K CRISPR–Cas systems by similar *Tn7*-like transposons that also encode TniQ[73,93], it will be informative to determine whether tethering of TniQ to evolutionarily distinct crRNA effector complexes— Cascade or Cas12k—is a general theme of RNA-guided transposition.

4.7: Materials and Methods

Statistics

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

TniQ-Cascade Purification

TniQ–Cascade purification Protein components of TniQ–Cascade were expressed from a pET-derivative vector containing the native *V. cholerae tniQ-cas8-cas7- cas6* operon with an N-terminal His10-MBP-TEV site fusion on TniQ. The crRNA was expressed separately from a pACYC-derivative vector containing a minimal repeat–spacer–repeat CRISPR array encoding a spacer from the endogenous *V. cholerae* CRISPR array. The TniQ–Cascade complex was overexpressed and purified as described previously[157], and was stored in Cascade storage buffer (20 mM Tris-Cl, pH 7.5, 200 mM NaCl, 1 mM DTT, 5% glycerol).

Sample Preparation for Electron Microscopy

For negative staining, 3 µl of purified TniQ–Cascade ranging from 100 nM to 2 µM was incubated with plasma treated (H_2/O_2 gas mix, Gatan Solarus) CF400 carbon-coated grids (EMS) for 1 min. Excess solution was blotted and 3 µl of 0.75% uranyl formate was added for an additional minute. Excess stain was blotted away and grids were air-dried overnight. Grid screening for both negative staining and cryo conditions was performed on a Tecnai-F20 microscope (FEI) operated at 200 KeV and equipped with a Gatan K2-Summit direct detector. Microscope operation and data collection were carried out using the Leginon/Appion software. Initial negative staining grid screening allowed determination of a suitable concentration range for cryo conditions. Several grid geometries were tested in the 1–4 µM concentration range for cryo conditions using a Vitrobot Mark-II operated at 4 °C, 100% humidity, blot force 3, drain time 0, waiting time 15 s, and blotting times ranging from 3–5 s. The best ice distribution and particle density was obtained with 0.6/1 UltrAuFoil grids (Quantifoil).

Electron Microscopy

A preliminary dataset of 300 images in cryo was collected with the Tecnai-F20 microscope using a pixel size of 1.22 Å/pixel with illumination conditions adjusted to 8 e^{-}/pixel/second with a frame window of 200 ms. Preprocessing and image processing were integrally done in Relion3[168] with ctf estimation integrated via a wrapper to Gctf[169]. An initial model computed using the SGD algorithm[170] implemented in Relion3 was used as initial reference for a refine three-dimensional job that generated a sub-nanometric reconstruction with approximately 10,000 selected particles. Clear secondary structure features in the two-dimensional averages and the three-dimensional reconstruction could be identified. For the DNA-bound TniQ–Cascade complex containing DNA, we pre-incubated two complementary 74-nt oligonucleotides:

(NTS: 5′-
TTCATCAAGCCATTGGACCGCCTTACAGGACGCTTTGGCTTCATTGCTTTTCAGCTTC
GCCTTGACGGCCAAAA-3′,
TS: 5′-
TTTTGGCCGTCAAGGCGAAGCTGAAAAGCAATGAAGCCAAAGCGTCCTGTAAGGCG
GTCCAATGGCTTGATGAA-3′)

for 5 min at 95 °C in hybridization buffer (20 mM Tris-Cl, pH 7.5, 100 mM KCl, 5 mM $MgCl_2$) to form dsDNA,which was subsequently aliquoted and flash-frozen. Complex formation was performed by incubating a 3× molar excess of dsDNA with TniQ–Cascade at 37 °C for 5 min before vitrification, which followed the conditions optimized for the apo complex (defined as TniQ–Cascade with High-resolution data for the apo complex were collected in a Tecnai-Polara-F30 microscope operated at 300 KeV equipped with a K3 direct detector (Gatan). A 30-µm C2 aperture was used with a pixel size of 0.95 Å/pixel and illumination conditions in microprobe mode adjusted to a fluence of 16 e^-/pixel/second. Four-second images with a frame width of 100 ms (1.77 $e^-/Å^2$/frame) were collected in counting mode. For the DNA-bound complex, high-resolution data were collected in a Titan Krios microscope (FEI) equipped with an energy filter

(20 eV slit width) and a K2 direct detector (Gatan) operated at 300 KeV. A 50-µm C2 aperture was used with a pixel size of 1.06 Å/pixel and illumination conditions adjusted in nanoprobe mode to a fluence of 8 e^{-}/pixel/second. Eight-second images with a frame width of 200 ms(1.42 e^{-}/Å^{2}/frame) were collected in counting mode.

Image Processing

Motion correction was performed for every micrograph applying the algorithm described for Motioncor2[171] implemented in Relion3 with 5-by-5 patches for the K2 data and 7 by 5 patches for the K3 data. Parameters of the contrast transfer function for each motion-corrected micrograph were obtained using Gctf[169] integrated in Relion3. Initial particle picking of a subset of 200 images randomly chosen was performed with the Laplacian tool of the Auto-picking module of Relion3, using an estimated size for the complex of 200 Å. Then, 15,000 particles were extracted in a 300-pixel box size and binned 3 times for an initial two- dimensional classification job. Selected two-dimensional averages from this job were used as templates for Auto-picking of the full dataset. The full dataset of binned particles was subjected to a two-dimensional classification job to identify particles able to generate averages with clear secondary structure features. The selected subgroup of binned particles after the two-dimensional classification selection was refined against a three-dimensional volume obtained by SGD with the F20 data. This consensus volume was inspected to localize areas of heterogeneity that were clearly identified at both ends of the crescent shape characteristic of this complex. Both ends were then individually masked using soft masks of around 20 pixels that were subsequently used in classification jobs without alignments in Relion3. The T parameter used for this classification job was 6 and the total number of classes was 10. This strategy allowed us to identify two main population of particles which correspond to an open and closed state of the complex. Particles from both subgroups were separately re-extracted to obtain unbinned data sets for further refinement. New features implemented in Relion3, namely Bayesian polishing and ctf parameters refinement, allowed the extension of the resolution to 3.4, 3.5 and 2.9 Å for the two apo and the DNA-bound complexes, respectively. Post-processing was performed with a soft-mask of 5 pixels being the B-factor estimated automatically in Relion3 following standard practice. A final set of local refinements was performed with the masks used for classification. The locally aligned maps exhibit very good quality for the ends of the C-shape. These maps were used for de novo modelling and initial model refinement.

Model Building and Refinement

For the Cas7 and Cas6 monomers, the *E. coli* homologs (PDB accession code 4TVX) were initially docked with Chimera[172] and transformed to poly-alanine models. Substantial rearrangement of the finger region of Cas7 monomers, as well as other secondary structure elements of Cas6, were performed manually in COOT[173] before amino acid substitution of the poly-alanine model. Well-defined bulky side chains of aromatic residues allowed a confident assignment of the register. The crRNA was also well defined in the maps and was traced de novo with COOT. For Cas8 and TniQ in particular, no structural similarity was found in the published structures that was able to explain our densities. Locally refined maps using soft masks at both ends of the crescent-shaped complex rendered well-defined maps below 3.5 Å resolution. These maps were used for manual de novo tracing of a poly-alanine model in COOT that was subsequently mutated to the *V. cholerae* sequences. Bulky side chains for aromatic residues showed excellent density and were used as landmarks to adjust the register of the sequence. For refinement, an initial step of real-space refinement against the cryo-EM maps was performed with the phenix.real_space refinement tool of the Phenix package[174], with secondary structure restraints activated. A second step of reciprocal space refinement was performed in Refmac5[175], with secondary restraints calculated with Prosmart[176] and LibG[177]. Weight of the geometry term versus the experimental term was adjusted to avoid overfitting of the model into cryo-EM map, as previously reported[178]. Model validation was performed in Molprobity[179].

4.8: Figures

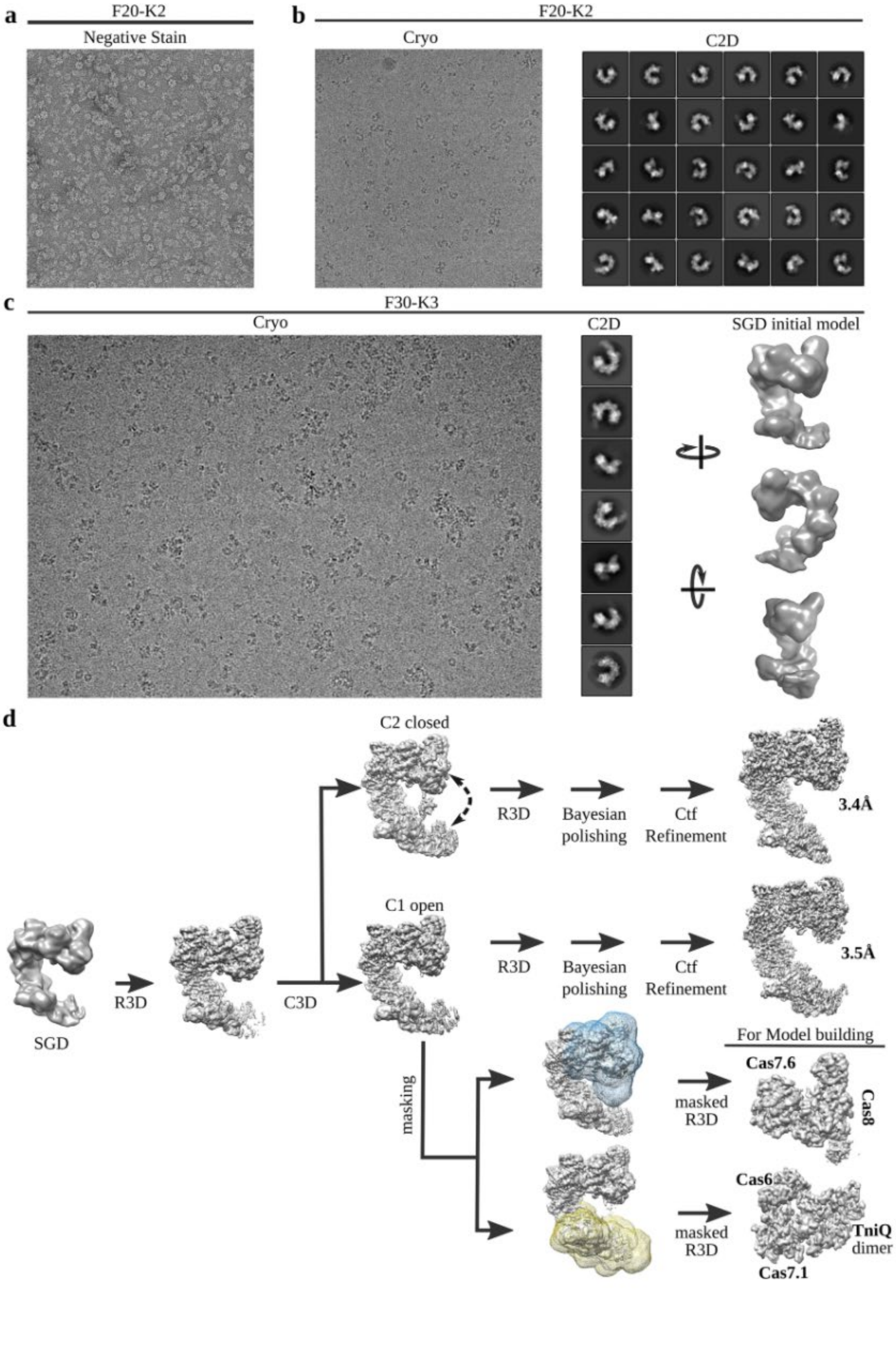

a
F20-K2
Negative Stain
b
F20-K2
Cryo
C2D
c
F30-K3
Cryo
C2D
SGD initial model
d
C2 closed
R3D
Bayesian polishing
Ctf Refinement
3.4Å
C1 open
R3D
Bayesian polishing
Ctf Refinement
3.5Å
SGD
R3D
C3D
masking
For Model building
masked R3D
Cas7.6
Cas8
masked R3D
Cas6
TniQ dimer
Cas7.1

Figure 39: Cryo-EM sample optimization and image processing workflow

a, Representative negatively stained micrograph for 500 nM TniQ–Cascade. **b**, Left, representative cryo-EM image for 2 μM TniQ–Cascade. A small dataset of 200 images was collected in a Tecnai-F20 microscope equipped with a Gatan K2 camera. Right, reference-free two-dimensional class averages for this initial cryo-EM dataset. **c**, Left, representative image from a large dataset collected in a Tecnai Polara microscope equipped with a Gatan K3 detector. Middle, detailed two-dimensional class averages were obtained that were used for initial model generation using the SGD algorithm[170] implemented in Relion3[161] (right). **d**, Image processing workflow used to identify the two main classes of the TniQ–Cascade complex in open and closed conformations. Local refinements with soft masks were used to improve the quality of the map within the terminal protuberances of the complex. These maps were instrumental for de novo modelling and initial model refinement.

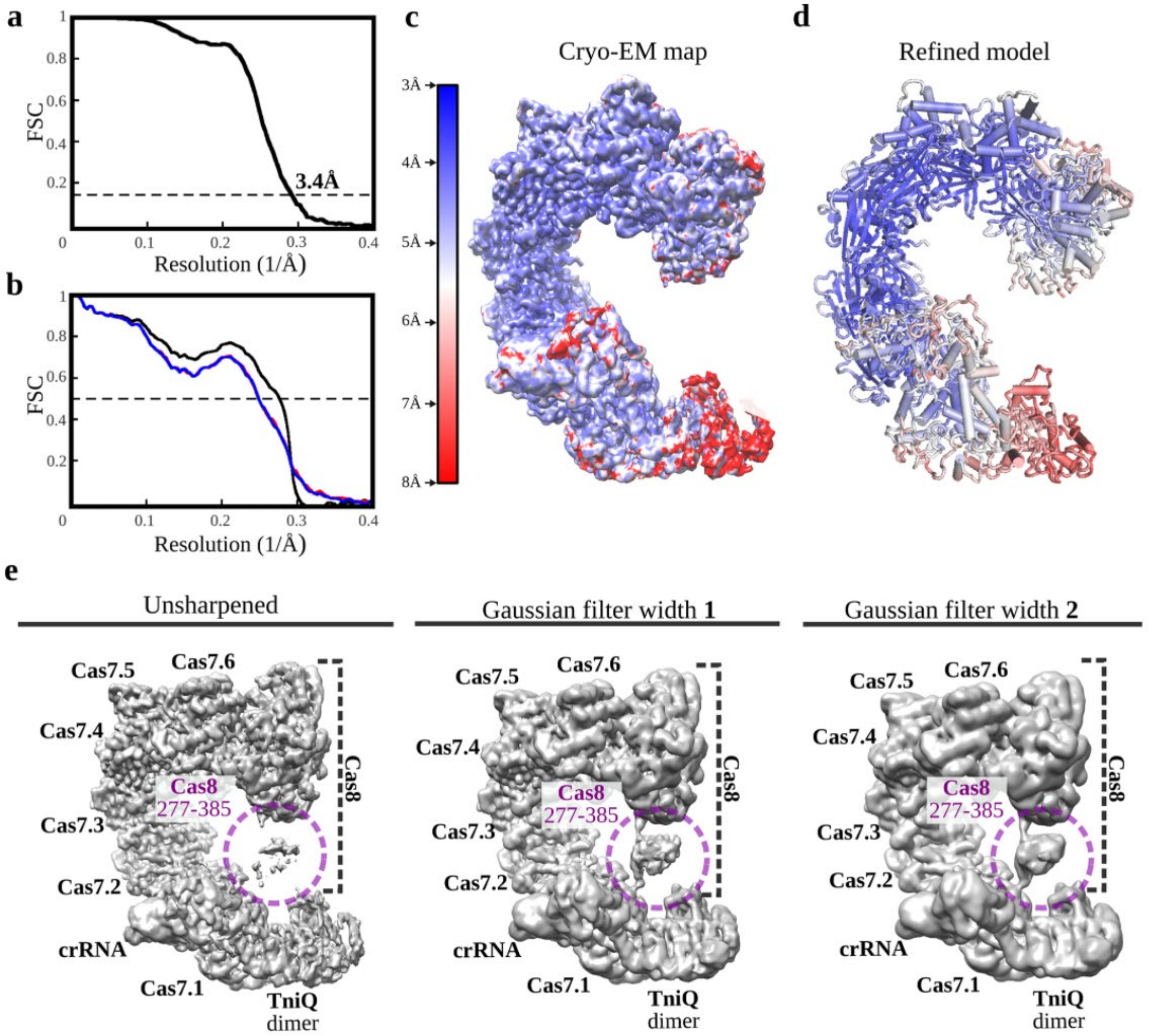


Figure 40: Fourier shell correlation curves, local resolution, and unsharpened filter maps for the TniQ-Cascade complex in closed conformation

a, Gold-standard Fourier shell correlation (FSC) curve using half maps; the global resolution estimation is 3.4 Å by the FSC 0.143 criterion. **b**, Cross-validation model-versus-map FSC. Blue curve, FSC between the shacked model refined against half map 1; red curve, FSC against half map 2, not included in the refinement; black curve, FSC between final model against the final map. The overlap observed between the blue and red curves guarantees a non-overfitted model. **c**, Unsharpened map colored according to local resolutions, as reported by RESMAP[180]. **d**, Final model colored according to B-factors calculated by REFMAC[175]. **e**, A flexible Cas8 domain encompassing residues 277–385 contacts the TniQ dimer at the other side of the crescent shape. Applying a Gaussian filter of increasing width to the unsharpened map allows for a better visualization of this flexible region.

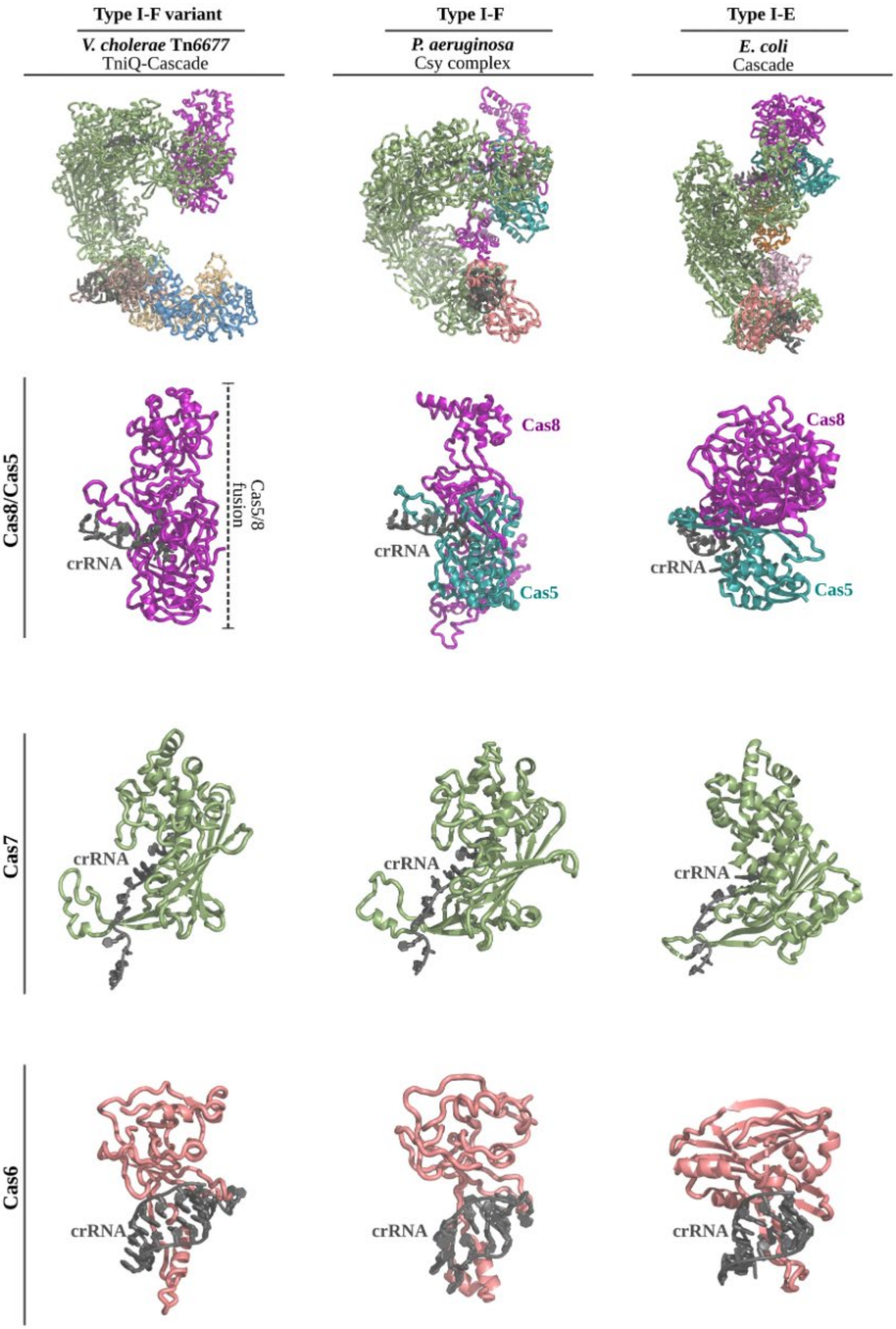

Type I-F variant
V. cholerae Tn*6677*
TniQ-Cascade
Type I-F
P. aeruginosa
Csy complex
Type I-E
E. coli
Cascade
Cas8/Cas5
Cas5/8 fusion
crRNA
Cas8
Cas5
Cas7
Cas6

Figure 41: Alignment of TniQ-Cascade with structurally similar Cascade complexes.

The *V. cholerae* I-F variant TniQ–Cascade complex (left) was superposed with *Pseudomonas aeruginosa* I-F Cascade[127] (also known as Csy complex; middle, PDB ID: 6B45) and *E. coli* I-E Cascade[158] (right, PDB ID: 4TVX). Shown are alignments of the entire complex (top), the Cas8 and Cas5 subunits with the 5′ crRNA handle (second from top), the Cas7 subunit with a fragment of crRNA (second from bottom) and the Cas6 subunit with the 3′ crRNA handle (bottom).

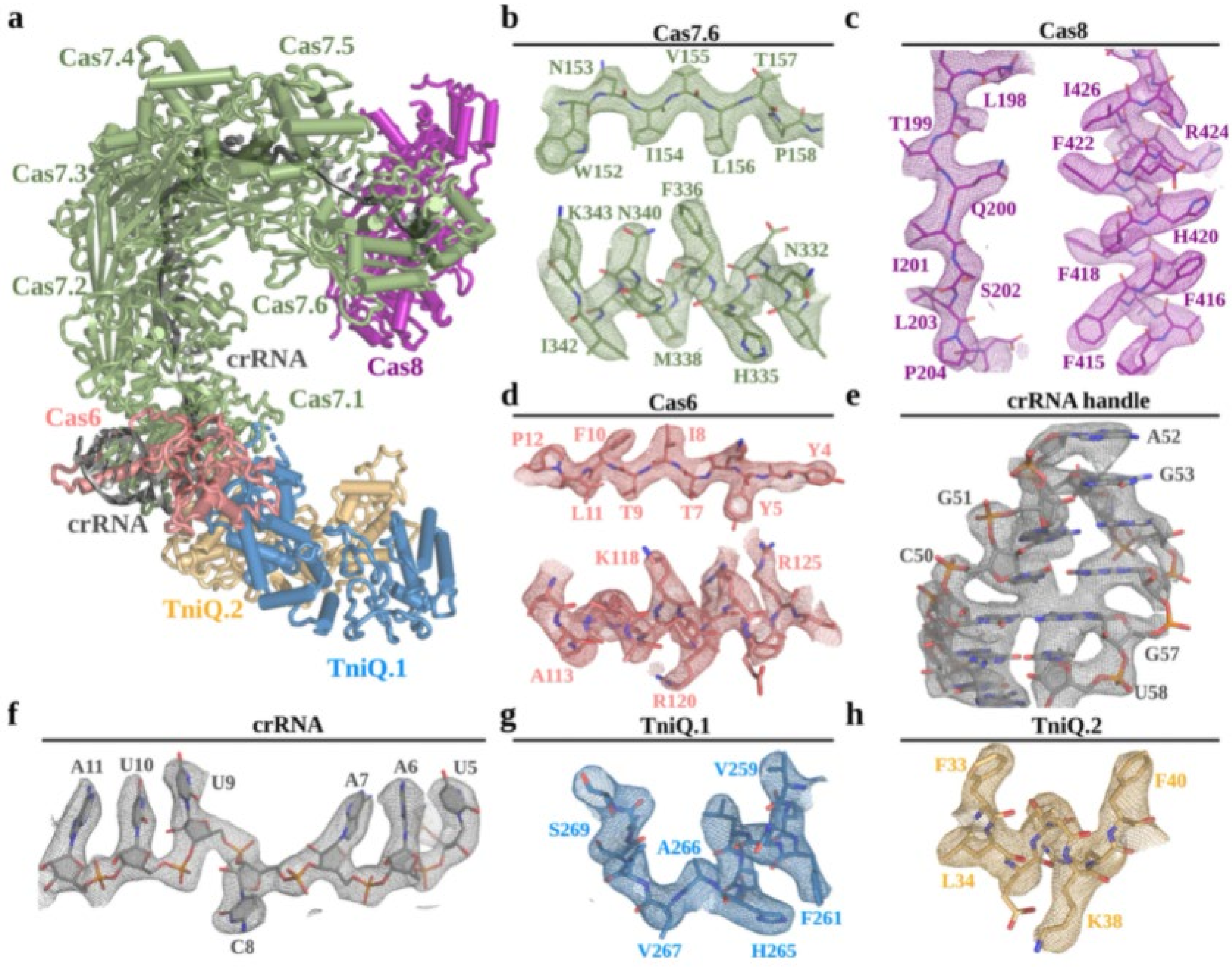


Figure 42: Representative cryo-EM densities for all the components of TniQ-Cascade complex in closed conformation

a, Final refined model of TniQ–Cascade, with Cas8 in purple, Cas7 monomers in green, Cas6 in salmon, the TniQ monomers in blue and yellow, and the crRNA in grey. **b**–**h**, Final refined model inserted in the final cryo-EM density for select regions of all the molecular components of the TniQ–Cascade complex. Residues are numbered.

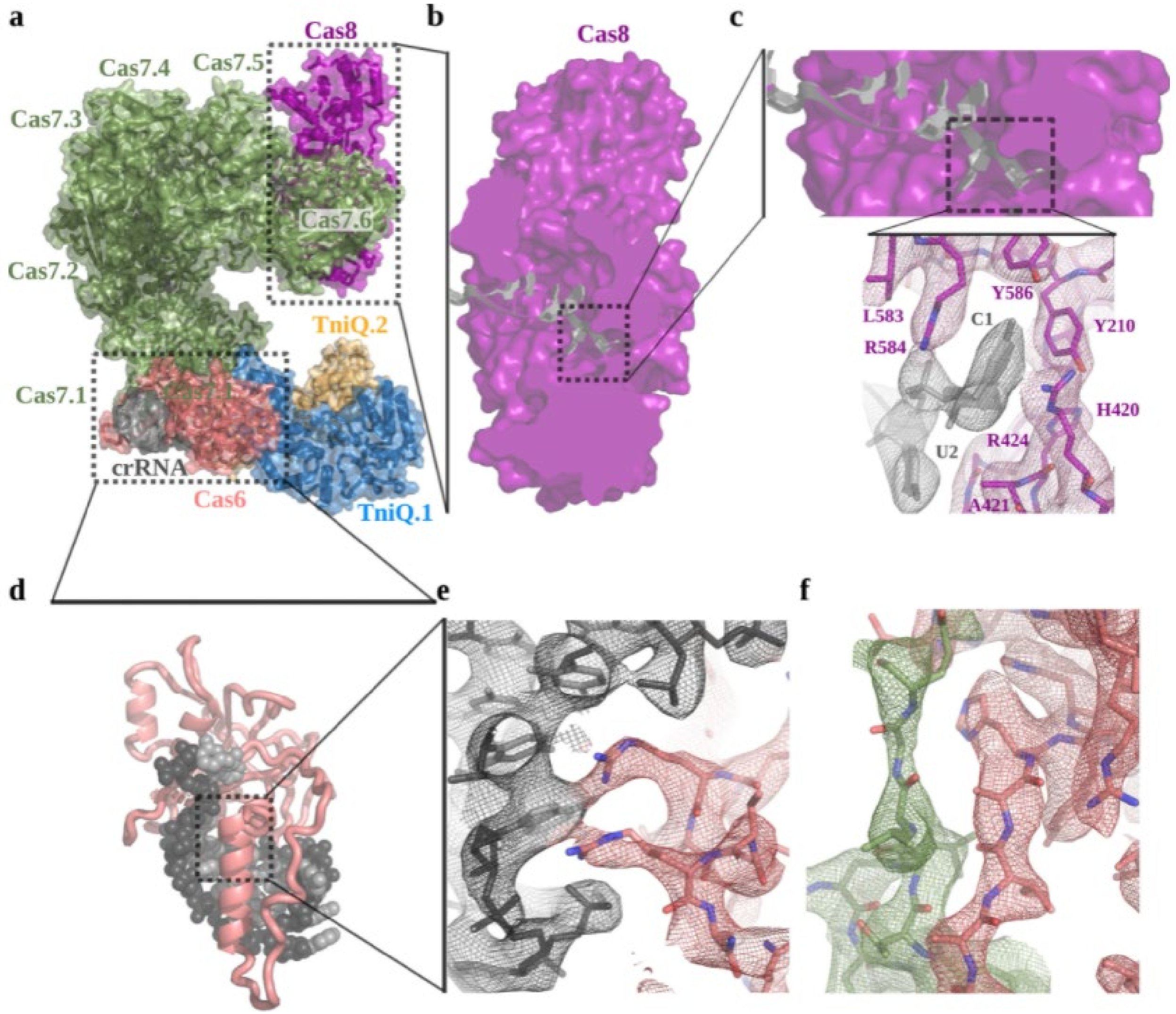


Figure 43: Cas8 and Cas6 interaction with the crRNA

a, Refined model for the TniQ–Cascade shown as ribbons inserted in the semi-transparent Van der Waals surface, coloured as in Figure 35. **b**, **c**, Magnified view of Cas8, which interacts with the 5′ end of the crRNA. The inset shows electron density for the highlighted region, where the base of nucleotide C1 is stabilized by stacking interactions with arginine residues R584 and R424. **d**, Cas6 interacts with the 3′ end of the crRNA 'handle' (nucleotides 45–60). **e**, An arginine-rich α-helix is deeply inserted within the major groove of the terminal stem–loop. This interaction is mediated by electrostatic interactions between basic residues of Cas6 and the negatively charged phosphate backbone of the crRNA. **f**, Cas6 (salmon) also interacts with Cas7.1 (green), establishing a β-sheet formed by β-strands contributed from both proteins.

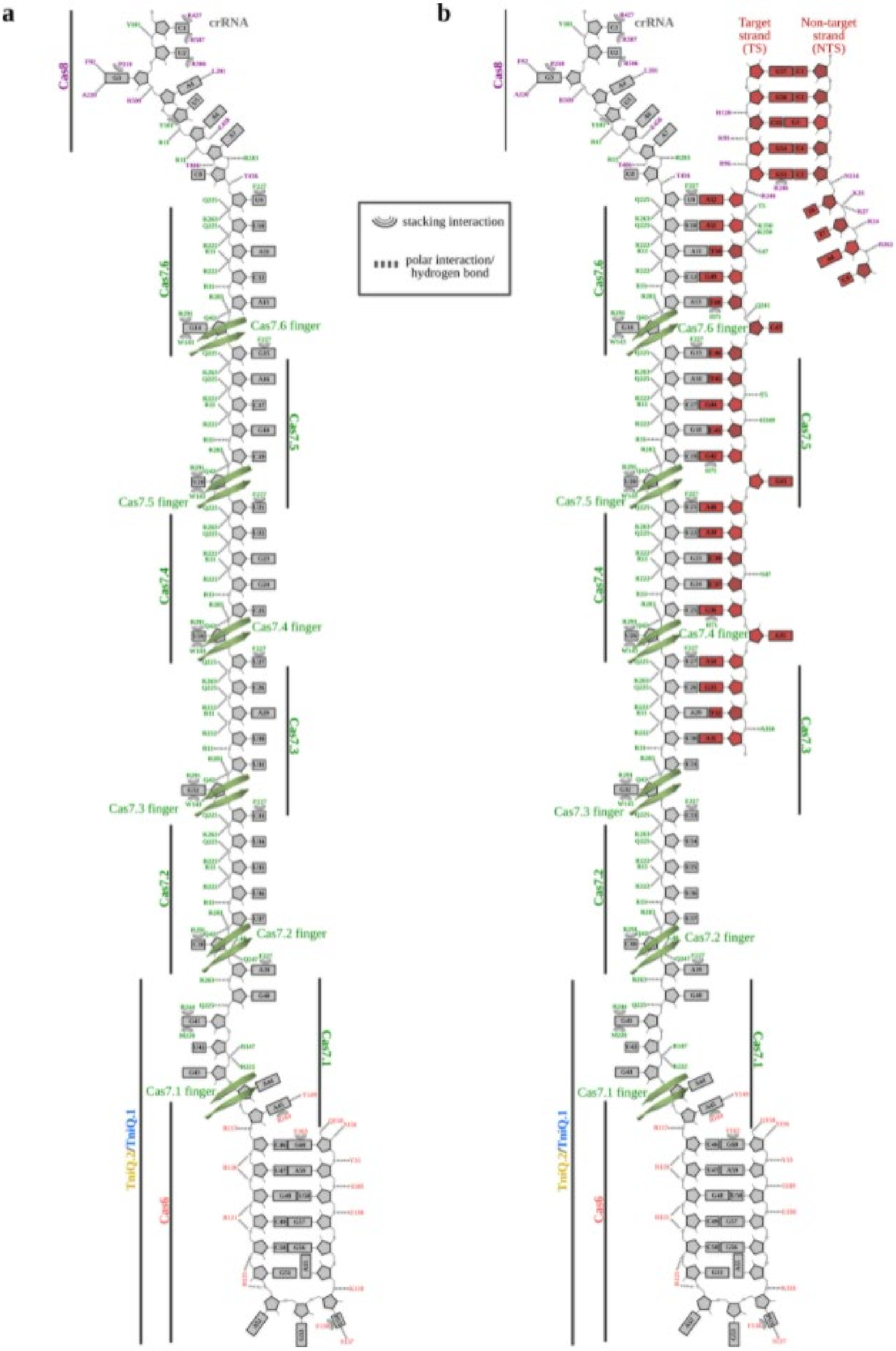

a
b
crRNA
Cas8
Target strand (TS)
Non-target strand (NTS)
stacking interaction
polar interaction/ hydrogen bond
Cas7.6
Cas7.6 finger
Cas7.5
Cas7.5 finger
Cas7.4
Cas7.4 finger
Cas7.3
Cas7.3 finger
Cas7.2
Cas7.2 finger
Cas7.1
Cas7.1 finger
TniQ.2/TniQ.1
Cas6

Figure 44: Schematic representation of crRNA and target DNA recognition by TniQ Cascade

a, TniQ–Cascade residues that interact with the crRNA are indicated. Approximate location for all protein components of the complex are also shown, as well as the position of each Cas7 finger.**b**, TniQ–Cascade residues that interact with crRNA and target DNA, shown as in **a**.

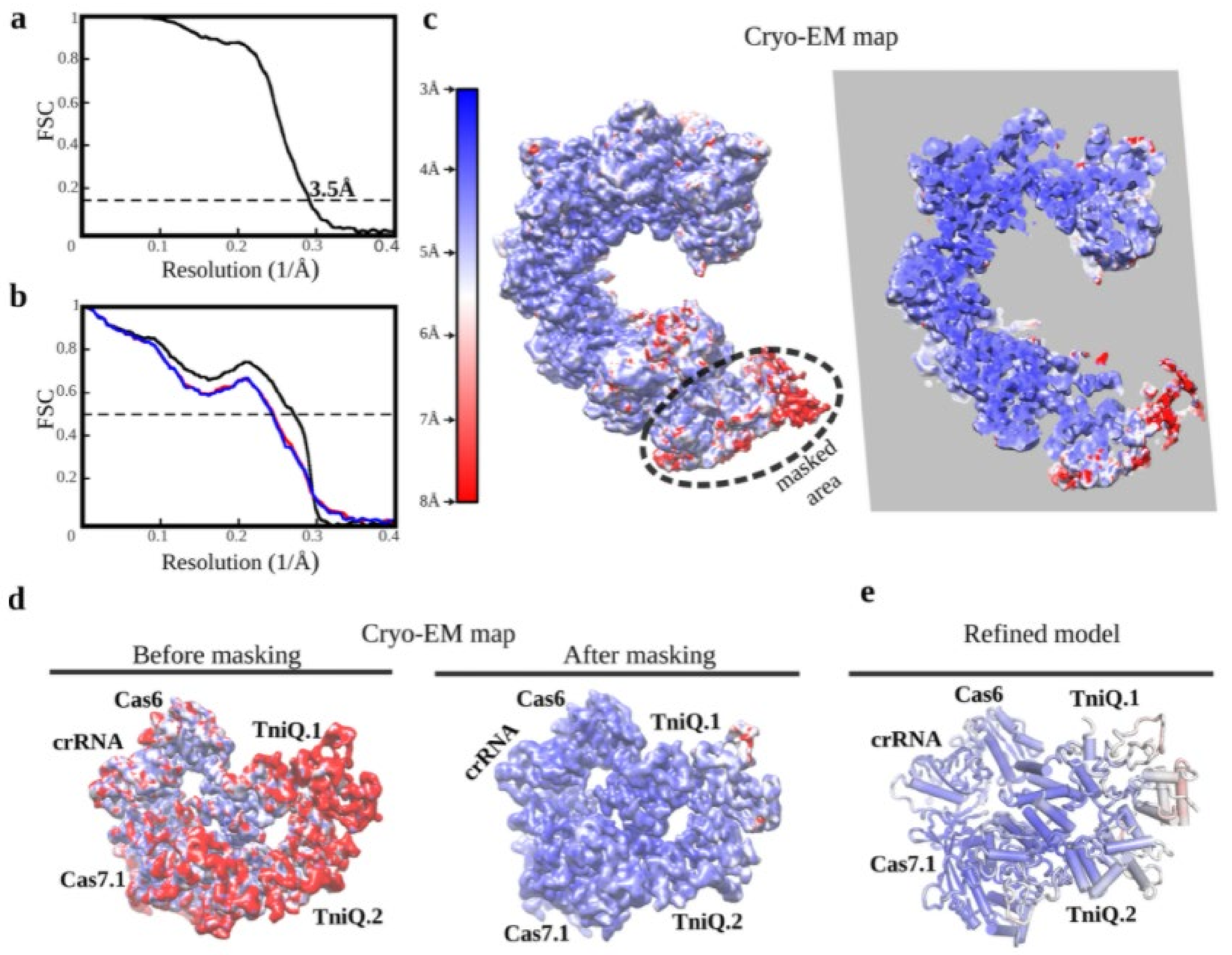


Figure 45: FSC curves, local resolution, and local refined maps for the TniQ-Cascade complex in open conformation

a, Gold-standard FSC curve using half maps; the global resolution estimation is 3.5 Å by the FSC 0.143 criterion. **b**, Cross-validation model-versus-map FSC. Blue curve, FSC between shacked model refined against half map 1; red curve, FSC against half map 2, not included in the refinement; black curve, FSC between final model against the final map. The overlapping between the blue and red curves guarantees a non-overfitted model. **c**, Unsharpened map colored according to local resolutions, as reported by RESMAP[180]. Right, slice through the map shown on the left. **d**, Local refinements with soft masks improved the maps in flexible regions. Shown is the region of the map corresponding to the TniQ dimer. Unsharpened maps colored according to the local resolution estimations are shown before (left) and after (right) masked refinements. **e**, Final model for the TniQ dimer region, colored according to the local B-factors calculated by REFMAC[175].

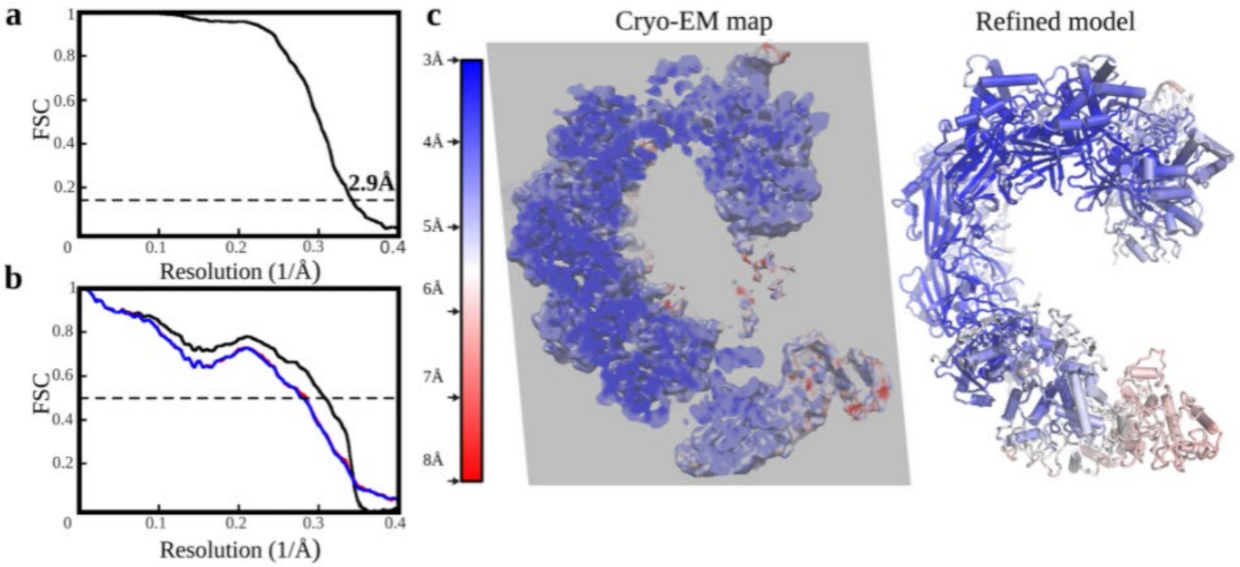


Figure 46: FSC curves, local resolution, and unsharpened filter maps for the DNA-bound TniQ-Cascade complex

a, Gold-standard FSC curve using half maps; the global resolution estimation is 2.9 Å by the FSC 0.143 criterion. **b**, Cross-validation model-versus-map FSC. Blue curve, FSC between the shacked model refined against half map 1; red curve, FSC against half map 2, not included in the refinement; black curve, FSC between final model against the final map. The overlap observed between the blue and red curves guarantees a non-overfitted model. **c**, Left, unsharpened map colored according to local resolutions, as reported by RESMAP[180]. dsDNA is visible at the top right projecting outside of the complex. Right, final model colored according to B-factors calculated by REFMAC[175].

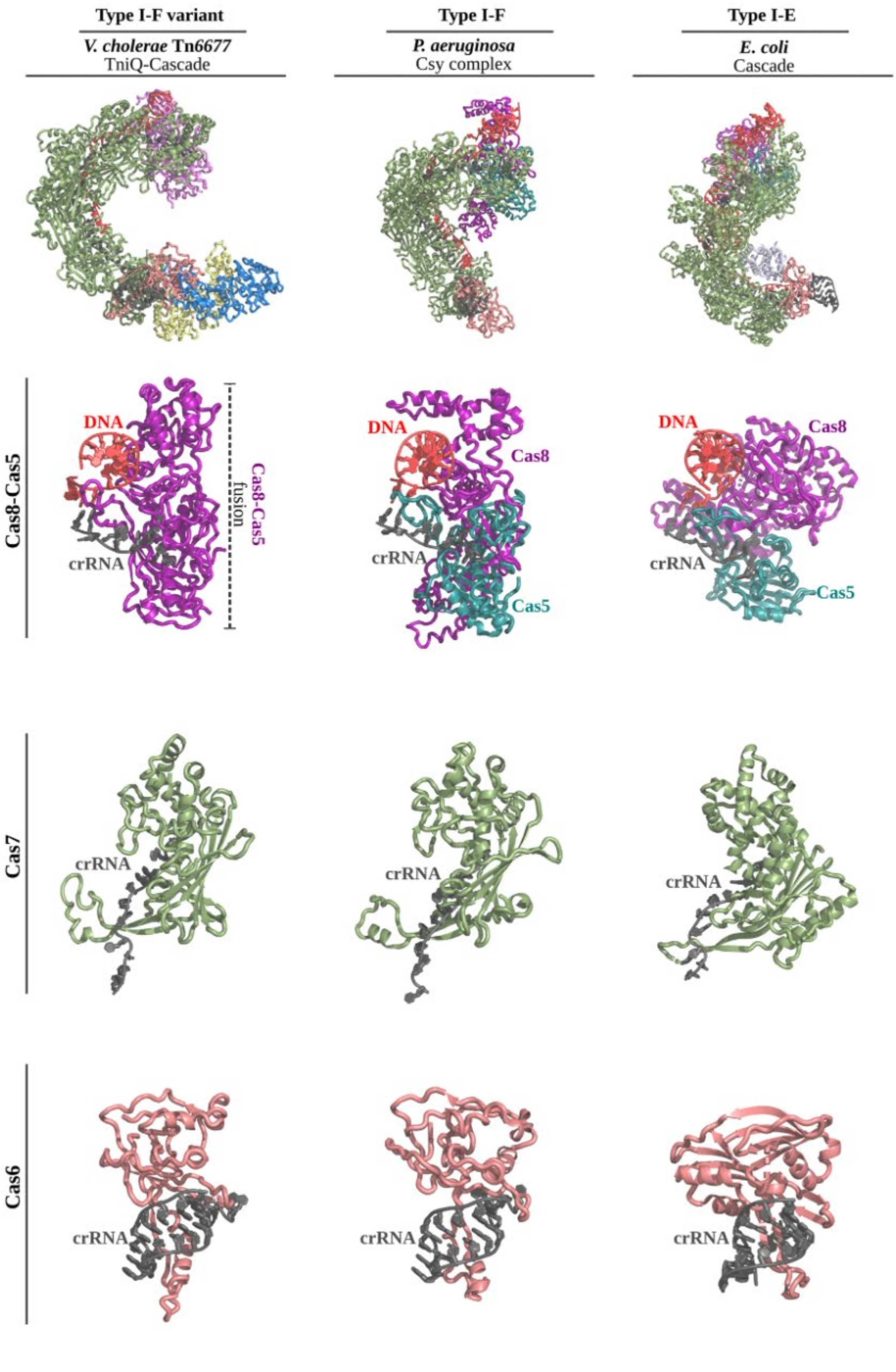

Type I-F variant
V. cholerae Tn6677
TniQ-Cascade
Type I-F
P. aeruginosa
Csy complex
Type I-E
E. coli
Cascade
Cas8-Cas5
DNA
crRNA
fusion
Cas8-Cas5
DNA
Cas8
crRNA
Cas5
DNA
Cas8
crRNA
Cas5
Cas7
crRNA
crRNA
crRNA
Cas6
crRNA
crRNA
crRNA

Figure 47: Alignment of DNA-bound TniQ-Cascade with structurally similar Cascade complexes

The DNA-bound structure of *V. cholerae* I-F variant TniQ–Cascade complex (left) was superposed with DNA-bound structures of *P. aeruginosa* I-F Cascade[127] (also known as Csy complex; middle, PDB ID: 6B44) and *E. coli* I-E Cascade[158]sc (right, PDB ID: 5H9F). Shown are alignments of the entire complex (top), the Cas8 and Cas5 subunits with the 5′ crRNA handle and double-stranded PAM DNA (middle top), the Cas7 subunit with a fragment of crRNA (middle bottom), and the Cas6 subunit with the 3′ crRNA handle (bottom).

4.9: Supplementary Table

Table 1: Supplementary Table 1

Cryo-EM data collection, refinement and validation statistics

	TniQ-Cascade (open) (EMDB-20349) (PDB 6PIF)	TniQ-Cascade (closed) (EMDB-20350) (PDB 6PIG)	TniQ-Cascade-DNA (EMDB-20351) (PDB 6PIJ)
Data collection and processing			
Magnification	96,000	96,000	130,000
Voltage (kV)	300	300	300
Electron exposure (e–/Å^{2})	70.91	70.91	56.95
Defocus range (μm)	-1/-3	-1/-3	-0.5/-2.5
Pixel size (Å)	0.95	0.95	1.06
Symmetry imposed	C1	C1	C1
Initial particle images (no.)	356,222	356,222	188,675
Final particle images (no.)	52,316	52,987	88,055
Map resolution (Å)	3.5	3.5	2.9
FSC threshold	0.143	0.143	0.143
Map resolution range (Å)	3-8	3-8	2.8-8
Refinement			
Initial model used (PDB code)	manual built	manual built	manual built
Model resolution (Å)	3.8	3.8	3.2
FSC threshold	0.5	0.5	0.5
Model resolution range (Å)	3.5-8	3.5-8	2.8-8
Map sharpening *B* factor (Å^{2})	-71.91	-77.01	-34.23
Model composition			
Non-hydrogen atoms	28,877	28,877	29,666
Protein residues	28,877	28,877	28,877
Ligands	-	-	789
B factors (Å^{2})			
Protein	92.47	96.5	93.9
Ligand	-	-	103.9
R.m.s. deviations			
Bond lengths (Å)	0.014	0.014	0.016
Bond angles (°)	1.77	1.78	1.78
Validation			
MolProbity score	2.34	1.81	1.82
Clashscore	6.89	1.82	3.32
Poor rotamers (%)	18.62	18.6	17.69
Ramachandran plot			
Favored (%)	81.38	81.41	82.31
Allowed (%)	96.09	96.55	96.53
Disallowed (%)	3.91	3.45	3.47

Chapter 5: Epilogue & Future Directions

The structural and functional analysis of Tn*6677* along with the Type V-K system have further expanded the CRISPR toolkit, paving the way for facile programmable insertions of specified genetic cargos[41,106,181–185]. The work presented in this thesis is already being used for discovery of related systems, and rational engineering of well-studied CRISPR-guided transposons to assess not only their molecular origins, but also how they might be adapted as tools for the scientific community. These systems could again prove the benefit of basic research as they appear tailor-made to address the difficulties of targeted integration; the same difficulties that have had scientists working for more than a decade to engineer a system that nature, and the microbial pangenome, has had eons to evolve. A facet of that evolution that others have started to untangle is how the transposons encoding CRISPR-Cas systems segregate their CRISPR-Cas machinery from host CRISPR-Cas machinery. CRISPR-associated transposons are mobile genetic elements, and since MGEs move between hosts, in doing so they may encounter traditional CRISPR-Cas immune systems. It is not unlikely that there is "crosstalk" between the CRISPR-Cas immune system and the newly- acquired transposon-encoded CRISPR-Cas system.

While we have demonstrated the stringent specificity of both the protein-protein interactions of the Tn*6677*-derived transposon (see Figure 21) and integration itself, others have shown that Type I-F1 CRISPR-Cas immune systems are fully capable of processing and utilizing the mature crRNA produced from a transposon-encoded Type I-F3 CRISPR array for targeted cleavage, creating the potential for Type I-F3 spacers to result in cleavage of the host genome[186]. To avoid this, and the resulting demise of their host, the Type I-F3 transposon-encoded CRISPR-Cas systems have adopted at least two strategies to avoid targeting host genomes. First, they have incorporated the use of autoinhibitory regulatory elements, such as Xre-family transcriptional regulators. As a result, there is a burst of expression from the transposon-encoded CRISPR-Cas locus upon initial host infiltration, and subsequent silencing of the transposon-encoded CRISPR locus when the Xre-family transcriptional regulators achieve a high enough concentration, presumably after the transposon has integrated into the host genome[186]. Second, crRNA's from the transposon-encoded CRISPR-Cas system are segregated by the targets of the spacers within said CRISPR array. The spacers within the transposon-encoded CRISPR array do not correspond to genomic targets within potential hosts, as they are instead complementary to sequences found in mobile plasmids, and phage genomes[32]. Therefore, if concentration-dependent regulation fails, the host's I-F1 CRISPR system will not target the host's genome, and will instead cleave a different mobile genetic element. Interestingly, this mechanism of regulation could potentially be used by the transposon to out-compete other MGEs within the same host. However, the CRISPR-containing transposons must still integrate into host genomes. To accomplish this, Type I-F3 and Type V-K systems appear to have a repeat-spacer-repeat sequence with atypical repeats, and imperfectly matching spacers, encoded shortly downstream of the transposon-encoded CRISPR array. Allowing for imperfect crRNA:DNA pairing, these sequences are complementary to host genomic sequences that are the expected distance from the end of the transposon[186,188]. The atypical crRNA not only has altered repeats that are poorly recognized by some Type I-F CRISPR immune systems, but also contains mismatches in the spacer region (other than in the tolerated N+6 positions) that further ablate target cleavage should this atypical crRNA be used by a I-F1 immune system[186]. This strategy is expanded upon in V-K systems where the atypical and delocalized crRNAs are also shortened[188]. Regardless of the system in question, it appears that nature has evolved a mechanism similar to what has been engineered with Cas9 to reduce off-target cleavage[189]. By utilizing a crRNA that is suboptimal for target-binding, the energetic threshold is increased enough that binding to the desired genomic target is now less efficient, but the mismatches between the spacer and target will near-completely abolish off-target binding, and thus cleavage, or in this case, transposition. While spacer-target mismatches are one mechanism of control, altered RNA structure with atypical repeats, like those found in the I-F3 systems, mitigate off-target insertion likely by again, increasing the energetic threshold required to fully assemble the RNP complexes and bind the correct target.

Of note, the mismatches between the atypical spacers and their targets do not occur in a fashion that would allow for flexibility of target sequences (should they have mutated or drifted), but instead are positioned early in the crRNA spacer region, sometimes even in the seed-sequence[186]. Tolerance of mismatches in the N+3 wobble-position of a codon is thought to be utilized by Tn*7*'s TnsD DNA-interacting residues so that Tn*7* may still transpose downstream of homologous, but not identical, *glmS* sequences. While it appears that both Type I-F3 systems and Type V-K systems utilize distinct types of crRNAs for different target types, the Type I-B systems, (see Figure 12) appear to utilize either the TnsA/B/C + Cascade + TniQ pathway where integration occurs downstream of the Cascade-bound target, or proceed through a TnsA/B/C + TnsD pathway, where, like in Tn*7*, the target-site is specified by TnsD through protein-DNA interactions[188]. The Type I-B systems appear to have encountered a similar problem as the I-F3 and V-K systems, but instead of providing sub-optimal crRNAs as a solution, the I-B systems retained, or incorporated, both TniQ and TnsD pathways for targeting MGEs and the host genome respectively. Interestingly, the TniQ and TnsD pathways appear to be in competition. When provided with targets for both TnsD and the crRNA around which Cascade is scaffolded, removal of either TniQ or TnsD increases transposition efficiency of the pathway that remains[188]. Of particular interest is that there appear to be two different mechanisms of host genome securitization that have evolved. One being whether integration proceeds through the TniQ or TnsD pathway, and the other being the utilization of crRNAs with sub-optimal target-complementarity and altered structural elements that may partially inhibit Cascade assembly.

As a testament to the speed at which the CRISPR field moves, two pre-prints were posted to BioRxiv describing solved Cryo-EM structures of TnsC homologs, shortly after initial drafting of this thesis. The first article describes the structure of the TnsC heptameric complex from *E. coli* Tn*7*, and the second describes the TnsC heptameric complex from the ShCAST Type V-K system[187,190]. Interestingly, the authors propose vastly different models for TnsC recruitment to their cognate DNA-targeting protein (TnsD for Tn*7* and TniQ & Cas12k for ShCAST). Shen *et al* purify Tn*7*'s TnsC assembled on distorted dsDNA that contains a 7bp mismatch flanked by 20bp on both sides of complementary DNA, forming what they call the 20-7-20 substrate. The 7-bp of mismatched bases provides a bubble that mimics the DNA distortion imposed by TnsD under native conditions[22]. Using a gain-of-function mutation and truncated TnsC for improved protein stability, they solved the structure of a shortened TnsC mutant, TnsC$^{A225\Delta504\text{-}555}$, bound to the 20-7-20 substrate. They found that TnsC forms a ring of seven TnsC protomers, and that this ring is stabilized by bound, non-hydrolyzed, ATP[187]. ATP hydrolysis prevents stabilization of the ring and, as TnsB induces hydrolysis of ATP in TnsC, is likely the mechanism through which target immunity is established[40]. Their proposed model has each TnsC protomer recruit a single copy of TnsA at the end of an extended "arm" of each TnsC protein. This TnsC/TnsA complex then builds a ring around the distorted DNA produced by TnsD, and the TnsC-bound TnsA proteins are then responsible for the recruitment of TnsB-decorated Tn*7* Ends. The entire complex is then able to integrate the transposon at the proscribed site, immediately downstream of *glmS*[187].

Park *et al* describe ShCAST TnsC, both as a filamentous species bound to DNA, and in complex with TniQ[190]. Unfortunately, the differences between ShCAST TniQ and *V. cholerae* TniQ (from Tn*6677*) prevent readers from drawing any conclusions, beyond the most obvious, about recruitment of their cognate TnsC proteins. Despite this, there is nothing in Halpin-Healy *et al*, nor their paper, that precludes their speculation of where a TnsC ring may contact the TniQ dimer at the PAM-distal end of a target-bound I-F3 Cascade[185,190]. The authors propose that unlike Tn*7*, TnsC searches for TniQ bound to DNA by forming filaments along dsDNA until it encounters a TniQ-Cas12k complex, and that this filament is then disassembled by TnsB (already bound to the ends of the transposon) until only the heptameric ring contacting TniQ directly remain. This TnsC "spring-washer" then accounts for the stretch of DNA between the target site and the insertion site. Furthermore, the disassembly of this filament may not be perfectly regular, and may account for the wide range of distances between the target site and insertion site that one can see in ShCAST[190].

Between the TnsC papers and the tendency of new CRISPR niches to explode in popularity, it is likely not long until there is an assembled structure of all components of CRISPR-associated transposons, and their mechanism is unraveled as the different structures and complexes are captured in various states of transposition. This is likely to happen concomitantly with the development of these systems as tools in heterologous systems. Of particular interest is the use of these tools in mammalian cells as they address near-all current shortcomings of modern DNA-integration methods, such as cargo size limitations, unpredictable genomic scarring, and targetability[191].

It is likely that the coming years will see a further refinement of the classification of transposons, CRISPR-Cas systems, and their components, that will allow a greater understanding of how these two systems intersected and established the observed mutualism. As more details are uncovered, those working in the field will incorporate the new data to better optimize these systems as tools in organisms that, to our knowledge, do not contain CRISPR-associated transposons.

Bibliography

1. Hickman, A. B. & Dyda, F. DNA Transposition at Work. *Chem. Rev.* **116**, 12758–12784 (2016).

2. McClintock, B. The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci. U. S. A.* **36**, 337–344 (1950).

3. Aziz, R. K., Breitbart, M. & Edwards, R. A. Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Res.* **38**, 4207–4217 (2010).

4. Feschotte, C. & Pritham, E. J. DNA Transposons and the Evolution of Eukaryotic Genomes. *Annu. Rev. Genet.* **41**, 331–368 (2007).

5. Hof, A. E. V. t. *et al.* The industrial melanism mutation in British peppered moths is a transposable element. *Nature* **534**, 102–105 (2016).

6. Zhang, Y. *et al.* Transposon molecular domestication and the evolution of the RAG recombinase. *Nature* **569**, 79–84 (2019).

7. Soucy, S. M., Huang, J. & Gogarten, J. P. Horizontal gene transfer: Building the web of life. *Nat. Rev. Genet.* **16**, 472–482 (2015).

8. Craig, N. L. *et al. Mobile DNA III*. (ASMScience, 2015). doi:10.1055/s-0032-1329178.

9. Koonin, E. V. & Makarova, K. S. Mobile genetic elements and evolution of crispr-cas systems: All theway there and back. *Genome Biol. Evol.* **9**, 2812–2825 (2017).

10. Deniz, Ö., Frost, J. M. & Branco, M. R. Regulation of transposable elements by DNA modifications. *Nat. Rev. Genet.* **20**, 417–431 (2019).

11. Gray, Y. H. M. It takes two transposons to tango: Transposable-element-mediated chromosomal rearrangements. *Trends Genet.* **16**, 461–468 (2000).

12. Munoz-Lopez, M. & Garcia-Perez, J. DNA Transposons: Nature and Applications in Genomics. *Curr. Genomics* **11**, 115–128 (2010).

13. Peters, J. E. & Craig, N. L. Tn7 : SMARTER THAN WE THOUGHT. **2**, 806–814 (2001).

14. tenOever, B. R. The Evolution of Antiviral Defense Systems. *Cell Host Microbe* **19**, 142–

149 (2016).

15. Felden, B. & Paillard, L. When eukaryotes and prokaryotes look alike: the case of regulatory RNAs. *FEMS Microbiol. Rev.* **41**, 624–639 (2017).

16. Kumar, A. Jump around: Transposons in and out of the laboratory. *F1000Research* **9**, 1–12 (2020).

17. Thomason, M. K. & Storz, G. Bacterial antisense RNAs: How many are there, and what are they doing? *Annu. Rev. Genet.* **44**, 167–188 (2010).

18. Swarts, D. C. *et al.* DNA-guided DNA interference by a prokaryotic Argonaute. *Nature* **507**, 258–261 (2014).

19. Steiniger-White, M., Rayment, I. & Reznikoff, W. S. Structure/function insights into Tn5 transposition. *Curr. Opin. Struct. Biol.* **14**, 50–57 (2004).

20. Davies, D. R., Goryshin, I. Y., Reznikoff, W. S. & Rayment, I. Three-dimensional structure of the tn5 synaptic complex transposition intermediate. *Science (80-.).* **289**, 77–85 (2000).

21. Finnegan, D. J. Retrotransposons. *Curr. Biol.* **22**, 432–437 (2012).

22. Mitra, R., McKenzie, G. J., Yi, L., Lee, C. A. & Craig, N. L. Characterization of the TnsD-attTn7 complex that promotes site-specific insertion of Tn7. *Mob. DNA* **1**, 1–14 (2010).

23. Querques, I. *et al.* A highly soluble Sleeping Beauty transposase improves control of gene insertion. *Nat. Biotechnol.* **37**, 1502–1512 (2019).

24. Parks, A. R. *et al.* Transposition into Replicating DNA Occurs through Interaction with the Processivity Factor. *Cell* **138**, 685–695 (2009).

25. Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982 (2007).

26. Lovell, S., Goryshin, I. Y., Reznikoff, W. R. & Rayment, I. Two-metal active site binding of a Tn5 transposase synaptic complex. *Nat. Struct. Biol.* **9**, 278–281 (2002).

27. Bhasin, A., Goryshin, I. Y. & Reznikoff, W. S. Hairpin formation in Tn5 transposition. *J. Biol. Chem.* **274**, 37021–37029 (1999).

28. Nicolas, E. *et al.* The Tn 3 -family of Replicative Transposons . *Mob. DNA III* 693–726 (2015) doi:10.1128/9781555819217.ch32.

29. May, E. W. & Craig, N. L. Switching from cut-and-paste to replicative Tn7 transposition. *Science (80-.).* **272**, 401–404 (1996).

30. Barth, P. T., Datta, N., Hedges, R. W. & Grinter, N. J. Transposition of a deoxyribonucleic acid sequence endcoding trimethoprim and streptomycin resistances from R483 to other replicons. *J. Bacteriol.* **125**, 800–810 (1976).

31. Barth, P. T. & Datta, N. Two naturally occurring transposons indistinguishable from Tn7. *J. Gen. Microbiol.* **102**, 129–134 (1977).

32. Peters, J. E., Makarova, K. S., Shmakov, S. & Koonin, E. V. Recruitment of CRISPR-Cas systems by Tn7-like transposons. *Proc. Natl. Acad. Sci.* **114**, E7358–E7366 (2017).

33. Arciszewska, L. K. & Craig, N. L. Interaction of the Tn7-encoded transposition protein TnsB with the ends of the transposon. *Nucleic Acids Res.* **19**, 5021–5029 (1991).

34. Sarnovsky, R. J., May, E. W. & Craig, N. L. The Tn7 transposase is a heteromeric complex in which DNA breakage and joining activities are distributed between different gene products. *EMBO J.* **15**, 6348–6361 (1996).

35. Skelding, Z., Sarnovsky, R. & Craig, N. L. Formation of a nucleoprotein complex containing Tn7 and its target DNA regulates transposition initiation. *EMBO J.* **21**, 3494–3504 (2002).

36. Choi, K. Y., Spencer, J. M. & Craig, N. L. The Tn7 transposition regulator TnsC interacts with the transposase subunit TnsB and target selector TnsD. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 2858–2865 (2014).

37. Bainton, R. J., Kubo, K. M., Feng, J. nong & Craig, N. L. Tn7 transposition: Target DNA recognition is mediated by multiple Tn7-encoded proteins in a purified in vitro system. *Cell* **72**, 931–943 (1993).

38. Peters, J. E. Targeted transposition with Tn7 elements: safe sites, mobile plasmids, CRISPR/Cas and beyond. *Mol. Microbiol.* **112**, 1635–1644 (2019).

39. Kuduvalli, P. N., Rao, J. E. & Craig, N. L. Target DNA structure plays a critical role in Tn7 transposition. *EMBO J.* **20**, 924–932 (2001).

40. Stellwagen, A. E. & Craig, N. L. Analysis of gain-of-function mutants of an ATP-dependent regulator of Tn7 transposition. *J. Mol. Biol.* **305**, 633–642 (2001).

41. Vo, P. L. H. *et al.* CRISPR RNA-guided integrases for high-efficiency, multiplexed bacterial genome engineering. *Nat. Biotechnol.* **39**, 480–489 (2021).

42. Stellwagen, A. E. & Craig, N. L. Avoiding self: Two Tn7-encoded proteins mediate target immunity in Tn7 transposition. *EMBO J.* **16**, 6823–6834 (1997).

43. Holder, J. W. & Craig, N. L. Architecture of the Tn7 posttransposition complex: An elaborate nucleoprotein structure. *J. Mol. Biol.* **401**, 167–181 (2010).

44. Ronning, D. R. *et al.* The carboxy-terminal portion of TnsC activates the Tn7 transposase through a specific interaction with TnsA. *EMBO J.* **23**, 2972–2981 (2004).

45. Hickman, A. B. *et al.* Unexpected structural diversity in DNA recombination: The restriction endonuclease connection. *Mol. Cell* **5**, 1025–1034 (2000).

46. Choi, K. Y., Li, Y., Sarnovsky, R. & Craig, N. L. Direct interaction between the TnsA and TnsB subunits controls the heteromeric Tn7 transposase. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E2038–E2045 (2013).

47. Lu, F. & Craig, N. L. Isolation and characterization of Tn7 transposase gain-of-function mutants: A model for transposase activation. *EMBO J.* **19**, 3446–3457 (2000).

48. Bainton, R., Gamas, P. & Craig, N. L. Tn7 transposition in vitro proceeds through an excised transposon intermediate generated by staggered breaks in DNA. *Cell* **65**, 805–816 (1991).

49. Stellwagen, A. E. & Craig, N. L. Gain-of-function mutations in TnsC, an ATP-dependent transposition protein that activates the bacterial transposon Tn7. *Genetics* **145**, 573–585 (1997).

50. Stellwagen, A. E. & Craig, N. L. Mobile DNA elements: Controlling transposition with ATP-dependent molecular switches. *Trends Biochem. Sci.* **23**, 486–490 (1998).

51. Sharpe, P. L. & Craig, N. L. Host proteins can stimulate Tn7 transposition: A novel role for the ribosomal protein L29 and the acyl carrier protein. *EMBO J.* **17**, 5822–5831 (1998).

52. Peters, J. E. & Craig, N. L. Tn7 recognizes transposition target structures associated with DNA replication using the DNA-binding protein TnsE. *Genes Dev.* **15**, 737–747 (2001).

53. Skelding, Z., Queen-Baker, J. & Craig, N. L. Alternative interactions between the Tn7 transposase and the Tn7 target DNA binding protein regulate target immunity and transposition. *EMBO J.* **22**, 5904–5917 (2003).

54. Yue, Y. *et al.* Extensive Mammalian Germline Genome Engineering. *bioRxiv* 1–15 (2019) doi:10.1101/2019.12.17.876862.

55. Bhatt, S. & Chalmers, R. Targeted DNA transposition in vitro using a dCas9-transposase fusion protein. *Nucleic Acids Res.* **47**, 8126–8135 (2019).

56. Chen, S. P. & Wang, H. H. An Engineered Cas-Transposon System for Programmable and Site-Directed DNA Transpositions. *Cris. J.* **X**, 1–20 (2019).

57. Kovač, A. *et al.* RNA-guided retargeting of Sleeping Beauty transposition in human cells. *Elife* **9**, 1–19 (2020).

58. Kebriaei, P. *et al.* Phase I trials using Sleeping Beauty to generate Find the latest version : Phase I trials using Sleeping Beauty to generate CD19-specific CAR T cells. *J. Clin. Invest.* **126**, 3363–3376 (2016).

59. Tipanee, J., Chai, Y. C., Driessche, T. Vanden & Chuah, M. K. Preclinical and clinical advances in transposon-based gene therapy. *Biosci. Rep.* **37**, 1–20 (2017).

60. Suttle, C. Viruses in the sea. *Nature* **437**, 356–361 (2005).

61. Darwin, C. *Notebook A*. (1837).

62. Doron, S. *et al.* Systematic discovery of antiphage defense systems in the microbial pangenome. *Science (80-.).* **359**, 0–12 (2018).

63. Pingoud, A., Wilson, G. G. & Wende, W. Type II restriction endonucleases - A historical perspective and more. *Nucleic Acids Res.* **42**, 7489–7527 (2014).

64. Hille, F. *et al.* The Biology of CRISPR-Cas: Backward and Forward. *Cell* **172**, 1239–1259 (2018).

65. Makarova, K. S. *et al.* Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nat. Rev. Microbiol.* **18**, 67–83 (2020).

66. Klompe, S. E. & Sternberg, S. H. Harnessing “A Billion Years of Experimentation”: The Ongoing Exploration and Exploitation of CRISPR–Cas Immune Systems. *Cris. J.* **1**, 141–158 (2018).

67. Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507**, 62–67 (2014).

68. Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Soria, E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol.* **60**, 174–182 (2005).

69. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science (80-.).* **337**, 816–821 (2012).

70. Gorski, S. A., Vogel, J. & Doudna, J. A. RNA-based recognition and targeting: Sowing the seeds of specificity. *Nat. Rev. Mol. Cell Biol.* **18**, 215–228 (2017).

71. Yamano, T. *et al.* Crystal Structure of Cpf1 in Complex with Guide RNA and Target DNA. *Cell* **165**, 949–962 (2016).

72. Rollins, M. C. F. *et al.* Structure Reveals a Mechanism of CRISPR-RNA-Guided Nuclease Recruitment and Anti-CRISPR Viral Mimicry. *Mol. Cell* **74**, 132-142.e5 (2019).

73. Faure, G., Scott, D. A. & Peters, J. E. CRISPR–Cas in mobile genetic elements: counter-defence and beyond. *Nat. Rev. Microbiol.* doi:10.1038/s41579-019-0204-7.

74. Makarova, K. S. *et al.* An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol.* **13**, 722–36 (2015).

75. Wu, K. J., Zimmer, C. & Peltier, E. Nobel Prize in Chemistry Awarded to 2 Scientists for Work on Genome Editing. *The New York Times* (2020).

76. Wiedenheft, B. *et al.* Erratum: RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions (Proceedings of the National Academy of Sciences of the United States of America (2011) 108, 25 (10092-10097) DOI: 10.1073/pnas.11027. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 15010 (2011).

77. Lin, J. *et al.* DNA targeting by subtype I-D CRISPR-Cas shows type i and type III features. *Nucleic Acids Res.* **48**, 10470–10478 (2020).

78. Sternberg, S. H., Haurwitz, R. E. & Doudna, J. A. Mechanism of substrate selection by a highly specific CRISPR endoribonuclease. *Rna* **18**, 661–672 (2012).

79. Sashital, D. G., Jinek, M. & Doudna, J. A. An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease Cse3. *Nat. Struct. Mol. Biol.* **18**, 680–687 (2011).

80. Wiedenheft, B. *et al.* RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *Proc. Natl. Acad. Sci.* **108**, 10092–10097 (2011).

81. Haurwitz, R. E., Jinek, M., Wiedenheft, B., Zhou, K. & Doudna, J. A. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science (80-.).* **329**, 1355–1358 (2010).

82. Wiedenheft, B. *et al.* Structures of the RNA-guided surveillance complex from a bacterial immune system. *Nature* **477**, 486–489 (2011).

83. Hayes, R. P. *et al.* Structural basis for promiscuous PAM recognition in type I-E Cascade from E. coli. *Nature* **530**, 499–503 (2016).

84. Wiedenheft, B. *et al.* Structures of the RNA-guided surveillance complex from a bacterial immune system. *Nature* **477**, 486–489 (2011).

85. Hochstrasser, M. L. *et al.* CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference. *Proc. Natl. Acad. Sci.* **111**, 6618–6623 (2014).

86. Chowdhury, S. *et al.* Structure Reveals Mechanisms of Viral Suppressors that Intercept a CRISPR RNA-Guided Surveillance Complex. *Cell* **169**, 47-57.e11 (2017).

87. Jore, M. M. *et al.* Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nat. Struct. Mol. Biol.* **18**, 529–536 (2011).

88. Mulepati, S., Héroux, A. & Bailey, S. Structural biology. Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target. *Science* **345**, 1479–1484 (2014).

89. Chowdhury, S. *et al.* Structure Reveals Mechanisms of Viral Suppressors that Intercept a CRISPR RNA-Guided Surveillance Structure Reveals Mechanisms of Viral Suppressors that Intercept a CRISPR RNA-Guided Surveillance Complex. *Cell* **169**, 47-51.e11 (2017).

90. Zetsche, B. *et al.* Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. *Cell* **163**, 759–771 (2015).

91. Kleinstiver, B. P. *et al.* Engineered CRISPR–Cas12a variants with increased activities and improved targeting ranges for gene, epigenetic and base editing. *Nat. Biotechnol.* **37**, 276–282 (2019).

92. Swarts, D. C. & Jinek, M. Mechanistic Insights into the cis- and trans-Acting DNase Activities of Cas12a. *Mol. Cell* **73**, 589-600.e4 (2019).

93. Strecker, J. *et al.* RNA-guided DNA insertion with CRISPR-associated transposases. *Science (80-.).* **364**, 48–53 (2019).

94. Rollins, M. C. F. *et al.* Cas1 and the Csy complex are opposing regulators of Cas2/3 nuclease activity. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E5113–E5121 (2017).

95. Makarova, K. S. *et al.* Evolution and classification of the CRISPR-Cas systems. *Nat. Rev. Microbiol.* **9**, 467–477 (2011).

96. Kuznedelov, K. *et al.* Altered stoichiometry Escherichia coli Cascade complexes with shortened CRISPR RNA spacers are capable of interference and primed adaptation. *Nucleic Acids Res.* **44**, 10849–10861 (2016).

97. Gleditzsch, D. *et al.* Modulating the Cascade architecture of a minimal Type I-F CRISPR-Cas system. *Nucleic Acids Res.* **44**, 5872–5882 (2016).

98. Dolan, A. E. *et al.* Introducing a Spectrum of Long-Range Genomic Deletions in Human Embryonic Stem Cells Using Type I CRISPR-Cas. *Mol. Cell* **74**, 936-950.e5 (2019).

99. Cameron, P. *et al.* Harnessing type I CRISPR–Cas systems for genome engineering in human cells. *Nat. Biotechnol.* 1–7 (2019) doi:10.1038/s41587-019-0310-0.

100. Pickar-Oliver, A. *et al.* Targeted transcriptional modulation with type I CRISPR–Cas systems in human cells. *Nat. Biotechnol.* (2019) doi:10.1038/s41587-019-0235-7.

101. Mali, P. *et al.* CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat. Biotechnol.* **31**, 833–838 (2013).

102. Liu, Y., Wan, X. & Wang, B. Engineered CRISPRa enables programmable eukaryote-like gene activation in bacteria. *Nat. Commun.* **10**, (2019).

103. Strecker, J. *et al.* Engineering of CRISPR-Cas12b for human genome editing. *Nat. Commun.* **10**, (2019).

104. Koonin, E. V., Makarova, K. S., Wolf, Y. I. & Krupovic, M. Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. *Nat. Rev. Genet.* **21**, 119–131 (2020).

105. Faure, G. *et al.* CRISPR–Cas in mobile genetic elements: counter-defence and beyond. *Nat. Rev. Microbiol.* **17**, 513–525 (2019).

106. Rice, P. A., Craig, N. L. & Dyda, F. Comment on RNA-guided DNA insertion with CRISPR-associated transposases. *Science.* **364**, 48–53 (2019).

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

108. Thomas, C. M. & Nielsen, K. M. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol.* **3**, 711–721 (2005).

109. Koonin, E. V. The Turbulent Network Dynamics of Microbial Evolution and the Statistical Tree of Life. *J. Mol. Evol.* **80**, 244–250 (2015).

110. Toussaint, A. & Chandler, M. *Prokaryote genome fluidity: Toward a system approach of the mobilome*. *Methods in Molecular Biology* vol. 804 (2012).

111. Dy, R. L., Richter, C., Salmond, G. P. C. & Fineran, P. C. Remarkable mechanisms in microbes to resist phage infections. *Annu. Rev. Virol.* **1**, 307–331 (2014).

112. Doron, S. *et al.* Systematic discovery of antiphage defense systems in the microbial pangenome. *Science (80-.).* **359**, 0–12 (2018).

113. Koonin, E. V., Makarova, K. S. & Wolf, Y. I. Evolutionary Genomics of Defense Systems in Archaea and Bacteria. *Annu. Rev. Microbiol.* **71**, 233–261 (2017).

114. Koonin, E. V. & Makarova, K. S. Mobile genetic elements and evolution of crispr-cas systems: All theway there and back. *Genome Biol. Evol.* **9**, 2812–2825 (2017).

115. Broecker, F. & Moelling, K. Evolution of Immune Systems From Viruses and Transposable Elements. *Front. Microbiol.* **10**, 1–15 (2019).

116. Kapitonov, V. V., Makarova, K. S. & Koonin, E. V. ISC, a novel group of bacterial and archaeal DNA transposons that encode Cas9 homologs. *J. Bacteriol.* **198**, 797–807 (2016).

117. Shmakov, S. *et al.* Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. *Mol. Cell* **60**, 385–397 (2015).

118. Krupovic, M., Béguin, P. & Koonin, E. V. Casposons: mobile genetic elements that gave rise to the CRISPR-Cas adaptation machinery. *Curr. Opin. Microbiol.* **38**, 36–43 (2017).

119. Peters, J. E. Tn7 30. *Microbiol. Spectr.* **2**, (2014).

120. Waddell, C. S. & Craig, N. L. Tn7 transposition: two transposition pathways directed by five Tn7-encoded genes. *Genes Dev.* **2**, 137–149 (1988).

121. Lichtenstein, C. & Brenner, S. Unique insertion site of Tn*7* in the *E. coli* chromosome. *Nature* **297**, 601–603 (1982).

122. McKown, R. L., Orle, K. A., Chen, T. & Craig, N. L. Sequence requirements of

Escherichia coli attTn7, a specific site of transposon Tn7 insertion. *J. Bacteriol.* **170**, 352–358 (1988).

123. McDonald, N. D., Regmi, A., Morreale, D. P., Borowski, J. D. & Fidelma Boyd, E. CRISPR-Cas systems are present predominantly on mobile genetic elements in Vibrio species. *BMC Genomics* **20**, 1–23 (2019).

124. Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Classification and Nomenclature of CRISPR-Cas Systems: Where from Here? *Cris. J.* **1**, 325–336 (2018).

125. Rollins, M. F., Schuman, J. T., Paulus, K., Bukhari, H. S. T. & Wiedenheft, B. Mechanism of foreign DNA recognition by a CRISPR RNA-guided surveillance complex from Pseudomonas aeruginosa. *Nucleic Acids Res.* **43**, 2216–2222 (2015).

126. Sarnovsky, R. J., May, E. W. & Craig, N. L. The Tn7 transposase is a heteromeric complex in which DNA breakage and joining activities are distributed between different gene products. *EMBO J.* **15**, 6348–61 (1996).

127. Guo, T. W. *et al.* Cryo-EM Structures Reveal Mechanism and Inhibition of DNA Targeting by a CRISPR-Cas Surveillance Complex. *Cell* **171**, 414-426.e12 (2017).

128. Xue, C. & Sashital, D. G. Mechanisms of Type I-E and I-F CRISPR-Cas Systems in Enterobacteriaceae. *EcoSal Plus* **8**, (2019).

129. Blosser, T. R. *et al.* Two distinct DNA binding modes guide dual roles of a CRISPR-cas protein complex. *Mol. Cell* **58**, 60–70 (2015).

130. Cooper, L. A., Stringer, A. M. & Wade, J. T. Determining the specificity of cascade binding, interference, and primed adaptation In Vivo in the Escherichia coli type I-E CRISPR-cas system. *MBio* **9**, 1–18 (2018).

131. Rutkauskas, M. *et al.* Directional R-loop formation by the CRISPR-cas surveillance complex cascade provides efficient off-target site rejection. *Cell Rep.* **10**, 1534–1543 (2015).

132. Luo, M. L. *et al.* The CRISPR RNA-guided surveillance complex in Escherichia coli accommodates extended RNA spacers. *Nucleic Acids Res.* **44**, 7385–7394 (2016).

133. Goodman, A. L. *et al.* Identifying Genetic Determinants Needed to Establish a Human Gut Symbiont in Its Habitat. *Cell Host Microbe* **6**, 279–289 (2009).

134. van Opijnen, T., Bodi, K. L. & Camilli, A. Tn-seq: High-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat. Methods* **6**, 767–772

(2009).

135. Wiles, T. J. *et al.* Combining Quantitative Genetic Footprinting and Trait Enrichment Analysis to Identify Fitness Determinants of a Bacterial Pathogen. *PLoS Genet.* **9**, (2013).

136. Sobecky, P. A. & Hazen, T. H. Horizontal gene transfer and mobile genetic elements in marine systems. *Methods Mol. Biol.* **532**, 435–453 (2009).

137. Makarova, K. S. Beyond the adaptive immunity: sub- and neofunctionalization of CRISPR–Cas systems and their components. in (2018).

138. Cheng, D. R., Yan, W. X. & Scott, D. A. Discovery of Type VI-D CRISPR-Cas Systems. in (2018).

139. Dunbar, C. E. *et al.* Gene therapy comes of age. *Science.* **359**, 175 (2018).

140. Gelvin, S. B. Integration of Agrobacterium T-DNA into the Plant Genome . *Annu. Rev. Genet.* **51**, 195–217 (2017).

141. Wurm, F. M. Production of recombinant protein therapeutics in cultivated mammalian cells. *Nat. Biotechnol.* **22**, 1393–1398 (2004).

142. Kvaratskhelia, M., Sharma, A., Larue, R. C., Serrao, E. & Engelman, A. Molecular mechanisms of retroviral integration site selection. *Nucleic Acids Res.* **42**, 10209–10225 (2014).

143. Di Matteo, M., Belay, E., Chuah, M. K. & VandenDriessche, T. Recent developments in transposon-mediated gene therapy. *Expert Opin. Biol. Ther.* **12**, 841–858 (2012).

144. Zelensky, A. N., Schimmel, J., Kool, H., Kanaar, R. & Tijsterman, M. Inactivation of Pol θ and C-NHEJ eliminates off-target integration of exogenous DNA. *Nat. Commun.* **8**, 1–7 (2017).

145. Cox, D. B. T., Platt, R. J. & Zhang, F. Therapeutic genome editing: Prospects and challenges. *Nat. Med.* **21**, 121–131 (2015).

146. Pawelczak, K. S., Gavande, N. S., VanderVere-Carozza, P. S. & Turchi, J. J. Modulating DNA Repair Pathways to Improve Precision Genome Engineering. *ACS Chem. Biol.* **13**, 389–396 (2018).

147. Schmidt, F., Cherepkova, M. Y. & Platt, R. J. Transcriptional recording by CRISPR spacer acquisition from RNA. *Nature* **562**, 380–385 (2018).

148. Myhrvold, C. *et al.* Field-deployable viral diagnostics using CRISPR-Cas13. *Science.* **360**, 444–448 (2018).

149. Harrington, L. B. *et al.* Programmed DNA destruction by miniature CRISPR-Cas14 enzymes. *Science (80-.).* **362**, 839–842 (2018).

150. Robert, X. & Gouet, P. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* **42**, 320–324 (2014).

151. Heidrich, N., Dugar, G., Vogel, J. & Sharma, C. M. Investigating CRISPR RNA Biogenesis and Function Using RNA-seq. in *CRISPR: Methods and Protocols* (eds. Lundgren, M., Charpentier, E. & Fineran, P. C.) 1–21 (Springer New York, 2015). doi:10.1007/978-1-4939-2687-9_1.

152. Reiter, W.-D., Palm, P. & Yeats, S. Transfer RNA genes frequently serve as integration sites for prokaryotic genetic elements. *Nucleic Acids Res.* **17**, 1907–1914 (1989).

153. Boyd, E. F., Almagro-Moreno, S. & Parent, M. A. Genomic islands are dynamic, ancient integrative elements in bacterial evolution. *Trends Microbiol.* **17**, 47–53 (2009).

154. Biswas, A., Gagnon, J. N., Brouns, S. J. J., Fineran, P. C. & Brown, C. M. CRISPRTarget: Bioinformatic prediction and analysis of crRNA targets. *RNA Biol.* **10**, 817–827 (2013).

155. Sinkunas, T. *et al.* In vitro reconstitution of Cascade-mediated CRISPR immunity in Streptococcus thermophilus. *EMBO J.* **32**, 385–394 (2013).

156. Redding, S. *et al.* Surveillance and Processing of Foreign DNA by the Escherichia coli CRISPR-Cas System. *Cell* **163**, 854–865 (2015).

157. Klompe, S. E., Vo, P. L. H., Halpin-Healy, T. S. & Sternberg, S. H. Transposon-encoded CRISPR–Cas systems direct RNA-guided DNA integration. *Nature* **571**, 219–225 (2019).

158. Jackson, R. N. *et al.* Structural biology. Crystal structure of the CRISPR RNA-guided surveillance complex from Escherichia coli. *Science* **345**, 1473–1479 (2014).

159. Guo, T. W. Cryo-EM structures reveal mechanism and inhibition of DNA targeting by a CRISPR–Cas surveillance complex. *Cell* **171**, (2017).

160. Mulepati, S., Héroux, A. & Bailey, S. Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target. *Science.* **345**, 1479–1484 (2014).

161. Zivanov, J. *et al.* RELION-3: New tools for automated high-resolution cryo-EM structure determination. *bioRxiv* 1–22 (2018) doi:10.1101/421123.

162. Holder, J. W. & Craig, N. L. Architecture of the Tn7 posttransposition complex: An elaborate nucleoprotein structure. *J. Mol. Biol.* **401**, 167–181 (2010).

163. Holm, L. & Laakso, L. M. Dali server update. *Nucleic Acids Res.* **44**, W351–W355 (2016).

164. Aravind, L., Anantharaman, V., Balaji, S., Babu, M. M. & Iyer, L. M. The many faces of the helix-turn-helix domain: Transcription regulation and beyond. *FEMS Microbiol. Rev.* **29**, 231–262 (2005).

165. Krissinel, E. Stock-based detection of protein oligomeric states in jsPISA. *Nucleic Acids Res.* **43**, W314–W319 (2015).

166. Xiao, Y., Luo, M., Dolan, A. E., Liao, M. & Ke, A. Structure basis for RNA-guided DNA degradation by Cascade and Cas3. *Science.* **361**, 1–7 (2018).

167. Choi, K. Y., Spencer, J. M. & Craig, N. L. The Tn7 transposition regulator TnsC interacts with the transposase subunit TnsB and target selector TnsD. *Proc. Natl. Acad. Sci.* **111**, E2858–E2865 (2014).

168. Scheres, S. H. W. Semi-automated selection of cryo-EM particles in RELION-1.3. *J. Struct. Biol.* **189**, 114–122 (2015).

169. Zhang, K. Gctf: Real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).

170. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. CryoSPARC: Algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).

171. Zheng, S. Q. *et al.* MotionCor2: Anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).

172. Pettersen, E. F. *et al.* UCSF Chimera - A visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).

173. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 486–501 (2010).

174. Afonine, P. V. *et al.* Real-space refinement in PHENIX for cryo-EM and crystallography. *Acta Crystallogr. Sect. D Struct. Biol.* **74**, 531–544 (2018).

175. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. Sect. D Biol.*

Crystallogr. **53**, 240–255 (1997).

176. Nicholls, R. A., Fischer, M., Mcnicholas, S. & Murshudov, G. N. Conformation-independent structural comparison of macromolecules with ProSMART. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **70**, 2487–2499 (2014).

177. Brown, A. *et al.* Tools for macromolecular model building and refinement into electron cryo-microscopy reconstructions. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **71**, 136–153 (2015).

178. Fernández, I. S., Bai, X. C., Murshudov, G., Scheres, S. H. W. & Ramakrishnan, V. Initiation of translation by cricket paralysis virus IRES requires its translocation in the ribosome. *Cell* **157**, 823–831 (2014).

179. Williams, C. J. *et al.* MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci.* **27**, 293–315 (2018).

180. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* **11**, 63–65 (2014).

181. Klompe, S. E., Vo, P. L. H., Halpin-Healy, T. S. & Sternberg, S. H. Transposon-encoded CRISPR–Cas systems direct RNA-guided DNA integration. *Nature* **571**, 219–225 (2019).

182. Li, Z., Zhang, H., Xiao, R. & Chang, L. Cryo-EM structure of a type I-F CRISPR RNA guided surveillance complex bound to transposition protein TniQ. *Cell Research* vol. 30 179–181 (2020).

183. Wang, B., Xu, W. & Yang, H. Structural basis of a Tn7-like transposase recruitment and DNA loading to CRISPR-Cas surveillance complex. *Cell Res.* **30**, 185–187 (2020).

184. Jia, N., Xie, W., de la Cruz, M. J., Eng, E. T. & Patel, D. J. Structure–function insights into the initial step of DNA integration by a CRISPR–Cas–Transposon complex. *Cell Research* vol. 30 182–184 (2020).

185. Halpin-Healy, T. S., Klompe, S. E., Sternberg, S. H. & Fernández, I. S. Structural basis of DNA targeting by a transposon-encoded CRISPR–Cas system. *Nature* **577**, 271–274 (2020).

186. Petassi, M., Hsieh, S.-C. & Peters, J. Guide RNA categorization enables target site choice in Tn7-CRISPR-Cas transposons. *Cell* **183**, 1–15 (2020).

187. Shen, Y., Gomez-blanco, J., Petassi, M. T., Peters, J. E. & Ortega, J. Structural basis for DNA targeting by the Tn7 transposon. *bioRxiv* 1–36 (2021)

doi:10.1101/2021.05.24.445525.

188. Saito, M. *et al.* Dual modes of CRISPR-associated transposon homing. *Cell* **184**, 1–13 (2021).

189. Fu, Y., Sander, J. D., Reyon, D., Cascio, V. M. & Joung, J. K. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat. Biotechnol.* **32**, 279–284 (2014).

190. Kellogg, E. Structural basis for target-site selection in RNA-guided DNA transposition systems. *bioRxiv* 1–50 (2021) doi:10.1101/2021.05.25.445634.

191. Pickar-Oliver, A. & Gersbach, C. A. The next generation of CRISPR–Cas technologies and applications. *Nat. Rev. Mol. Cell Biol.* **20**, 490–507 (2019).

Appendix A

Klompe *et al*, 2019

https://doi.org/10.1038/s41586-019-1323-z

Transposon-encoded CRISPR–Cas systems direct RNA-guided DNA integration

Sanne E. Klompe[1], Phuc L. H. Vo[2,3], Tyler S. Halpin-Healy[1,3] & Samuel H. Sternberg[1]*

Conventional CRISPR–Cas systems maintain genomic integrity by leveraging guide RNAs for the nuclease-dependent degradation of mobile genetic elements, including plasmids and viruses. Here we describe a notable inversion of this paradigm, in which bacterial Tn7-like transposons have co-opted nuclease-deficient CRISPR–Cas systems to catalyse RNA-guided integration of mobile genetic elements into the genome. Programmable transposition of *Vibrio cholerae* Tn*6677* in *Escherichia coli* requires CRISPR- and transposon-associated molecular machineries, including a co-complex between the DNA-targeting complex Cascade and the transposition protein TniQ. Integration of donor DNA occurs in one of two possible orientations at a fixed distance downstream of target DNA sequences, and can accommodate variable length genetic payloads. Deep-sequencing experiments reveal highly specific, genome-wide DNA insertion across dozens of unique target sites. This discovery of a fully programmable, RNA-guided integrase lays the foundation for genomic manipulations that obviate the requirements for double-strand breaks and homology-directed repair.

Horizontal gene transfer, a process that allows genetic information to be transmitted between phylogenetically unrelated species, is a major driver of genome evolution across the three domains of life[1–3]. Mobile genetic elements that facilitate horizontal gene transfer are especially pervasive in bacteria and archaea, in which viruses, plasmids and transposons constitute the vast prokaryotic mobilome[4]. In response to the ceaseless assault of genetic parasites, bacteria have evolved numerous innate and adaptive defence strategies for protection, including RNA-guided immune systems encoded by clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR-associated (Cas) genes[5–7]. Remarkably, the evolution of CRISPR–Cas is intimately linked to the large reservoir of genes provided by mobile genetic elements, with core enzymatic machineries involved in both new spacer acquisition (Cas1) and RNA-guided DNA targeting (Cas9 and Cas12) derived from transposable elements[8–13]. These examples support a 'guns-for-hire' model, in which the rampant shuffling of genes between offensive and defensive roles results from the perennial arms race between bacteria and mobile genetic elements.

We set out to uncover examples of functional associations between defence systems and mobile genetic elements. In this regard, we were inspired by a recent report that described a class of bacterial Tn*7*-like transposons encoding evolutionarily linked CRISPR–Cas systems and proposed a functional relationship between RNA-guided DNA targeting and transposition[14]. The well-studied *E. coli* Tn*7* transposon is unique in that it mobilizes via two mutually exclusive pathways—one that involves non-sequence-specific integration into the lagging-strand template during replication, and a second that involves site-specific integration downstream of a conserved genomic sequence[15]. Notably, those Tn*7*-like transposons that specifically associate with CRISPR–Cas systems lack a key gene involved in DNA targeting, and the CRISPR–Cas systems that they encode lack a key gene involved in DNA degradation. We therefore hypothesized that transposon-encoded CRISPR–Cas systems have been repurposed for a role other than adaptive immunity, in which RNA-guided DNA targeting is leveraged for a novel mode of transposon mobilization.

Here we demonstrate that a CRISPR–Cas effector complex from *V. cholerae* directs an accompanying transposase to integrate DNA downstream of a genomic target site complementary to a guide RNA, representing the discovery of a programmable integrase. Beyond revealing an elegant mechanism by which mobile genetic elements have hijacked RNA-guided DNA targeting for their evolutionary success, our work highlights an opportunity for facile, site-specific DNA insertion without requiring homologous recombination.

Cascade directs site-specific DNA integration

We set out to develop assays for monitoring transposition from a plasmid-encoded donor into the genome, first using *E. coli* Tn*7*, a well-studied cut-and-paste DNA transposon[16] (Extended Data Fig. 1a). The Tn*7* transposon contains characteristic left- and right-end sequences and encodes five *tns* genes, *tnsA*–*tnsE*, which collectively encode a heteromeric transposase: TnsA and TnsB are catalytic enzymes that excise the transposon donor via coordinated double-strand breaks; TnsB, a member of the retroviral integrase superfamily, catalyses DNA integration; TnsD and TnsE constitute mutually exclusive targeting factors that specify DNA insertion sites; and TnsC is an ATPase that communicates between TnsAB and TnsD or TnsE[15]. Previous studies have shown that *E. coli* TnsD (*Eco*TnsD) mediates site-specific Tn*7* transposition into a conserved Tn*7* attachment site (*att*Tn*7*) downstream of the *glmS* gene in *E. coli*[17,18], whereas *Eco*TnsE mediates random transposition into the lagging-strand template during replication[19]. We recapitulated TnsD-mediated transposition by transforming *E. coli* BL21(DE3) cells with pEcoTnsABCD and pEcoDonor, and detecting genomic transposon insertion events by PCR and Sanger sequencing (Supplementary Table 1 and Extended Data Fig. 1).

To test the hypothesis that CRISPR-associated targeting complexes direct transposons to genomic sites complementary to a guide RNA (Fig. 1a), we selected a representative transposon from *V. cholerae* strain HE-45, Tn*6677*, which encodes a variant type I-F CRISPR–Cas system[20,21] (Extended Data Fig. 1f, Supplementary Note, Supplementary Table 2 and Supplementary Figs. 2–8). This transposon is bounded by

[1]Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY, USA. [2]Department of Pharmacology, Columbia University, New York, NY, USA. [3]These authors contributed equally: Phuc L. H. Vo, Tyler S. Halpin-Healy. *e-mail: shsternberg@gmail.com

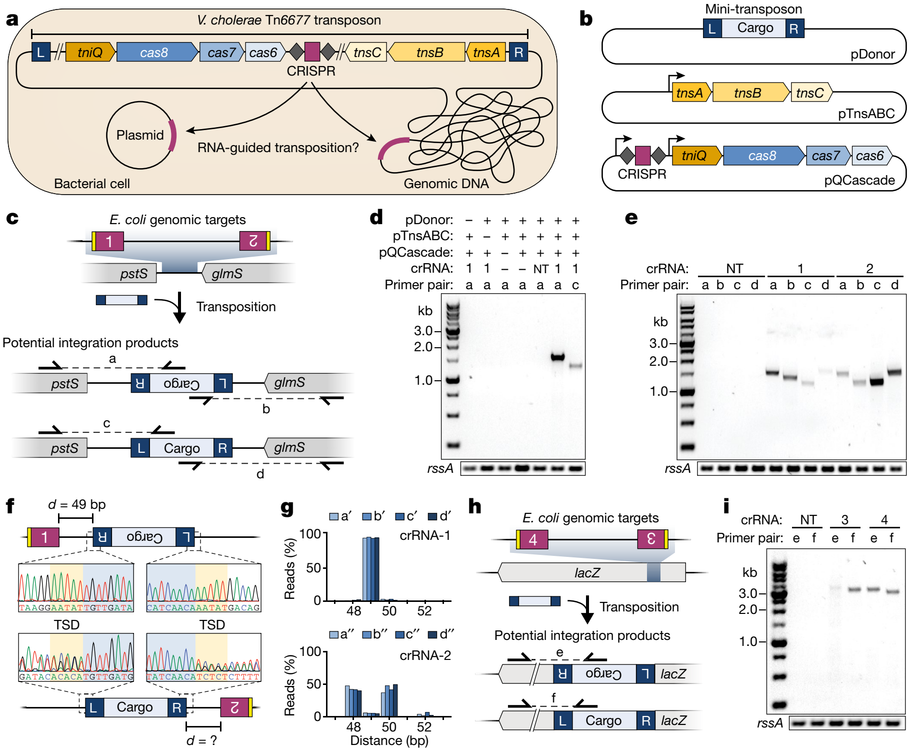


Fig. 1 | RNA-guided DNA integration with a *V. cholerae* transposon. **a**, Hypothetical scenario for Tn*6677* transposition into plasmid or genomic target sites complementary to a crRNA. **b**, Plasmid schematics for transposition experiments in which a mini-transposon on pDonor is mobilized in *trans*. The CRISPR array comprises two repeats (grey diamonds) and a single spacer (maroon rectangle). **c**, Genomic locus targeted by crRNA-1 and crRNA-2, two potential transposition products, and the PCR primer pairs to selectively amplify them. The PAMs and target sites are in yellow and maroon, respectively. **d**, PCR analysis of transposition with a non-targeting crRNA (crRNA-NT) and crRNA-1, resolved by agarose gel electrophoresis. **e**, PCR analysis of transposition with crRNA-NT, crRNA-1 and crRNA-2 using four distinct primer pairs, resolved by agarose gel electrophoresis. **f**, Sanger sequencing chromatograms for upstream and downstream junctions of genomically integrated transposons from experiments with crRNA-1 and crRNA-2. Overlapping peaks for crRNA-2 suggest the presence of multiple integration sites. The distance between the 3′ end of the target site and the first base of the transposon sequence is designated ‘*d*’. TSD, target-site duplication. **g**, NGS analysis of the distance between the Cascade target site and transposon integration site, determined for crRNA-1 and crRNA-2 with four primer pairs. **h**, Genomic locus targeted by crRNA-3 and crRNA-4. **i**, PCR analysis of transposition with crRNA-NT, crRNA-3 and crRNA-4, resolved by agarose gel electrophoresis. For **d**, **e** and **i**, amplification of *rssA* serves as a loading control; gel source data may be found in Supplementary Fig. 1.

left- and right-end sequences, distinguishable by their TnsB-binding sites, and includes a terminal operon that comprises the *tnsA*, *tnsB* and *tnsC* genes. Notably, the *tniQ* gene, a homologue of *E. coli tnsD*, is encoded within the *cas* rather than the *tns* operon, whereas *tnsE* is absent entirely. Like other such transposon-encoded CRISPR–Cas systems[14], the *cas1* and *cas2* genes responsible for spacer acquisition are conspicuously absent, as is the *cas3* gene responsible for target DNA degradation. The putative DNA-targeting complex Cascade (also known as Csy complex[6]) is encoded by three genes: *cas6*, *cas7* and a natural *cas8*–*cas5* fusion[21] (hereafter referred to simply as *cas8*). The native CRISPR array, comprising four repeat and three spacer sequences, encodes mature CRISPR RNAs (crRNAs) that we also refer to as guide RNAs.

We transformed *E. coli* with plasmids that encode components of the *V. cholerae* transposon, including a mini-transposon donor (pDonor), the *tnsA*-*tnsB*-*tnsC* operon (pTnsABC), and the *tniQ*-*cas8*-*cas7*-*cas6* operon alongside a synthetic CRISPR array (pQCascade) (Fig. 1b). The CRISPR array was designed to produce a non-targeting crRNA or crRNA-1, which targets a genomic site downstream of *glmS* flanked by a 5′-CC-3′ protospacer adjacent motif (PAM)[22] (Supplementary Table 3). Notably, we observed PCR products from cellular lysate between a genome-specific primer and either of two transposon-specific primers in experiments containing pTnsABC, pDonor and pQCascade expressing crRNA-1, but not with a non-targeting crRNA or any empty vector controls (Fig. 1c, d).

Because parallel reactions with oppositely oriented transposon primers revealed integration events within the same biological sample, we hypothesized that, unlike *E. coli* Tn*7*, RNA-guided transposition might occur in either orientation. We tested this by performing additional PCRs, by adding a downstream genomic primer, and by targeting an additional site with crRNA-2 found in the same genomic locus but on the opposite strand. For both crRNA-1 and crRNA-2, transposition products in both orientations were present, although with distinct orientation preferences based on relative band intensities (Fig. 1e). Given the presence of discrete bands, it appeared that integration was occurring at a set distance from the target site, and Sanger and next-generation sequencing (NGS) analyses revealed that more than 95% of integration events for crRNA-1 occurred 49 base pairs (bp) from the 3′ edge of the target site. The observed pattern with crRNA-2 was more complex, with integration clearly favouring distances of 48 and 50 bp over 49 bp. Both sequencing approaches also revealed the expected 5-bp target-site duplication that is a hallmark feature of Tn*7* transposition products[15] (Fig. 1f, g).

The *V. cholerae* Tn*6677* transposon is not naturally present downstream of *glmS*, and we saw no evidence of site-specific transposition within this locus when we omitted the crRNA (Fig. 1d). Nevertheless, we wanted to ensure that integration specificity was solely guided by the crRNA sequence, and not by any intrinsic preference for the *glmS* locus. We therefore cloned and tested crRNA-3 and crRNA-4, which target opposite strands within the *lacZ* coding sequence. We again observed bidirectional integration 48–50 bp downstream of both target sites, and were able to isolate clonally integrated, *lacZ*-knockout strains after performing blue–white colony screening on X-gal-containing LB-agar plates (Fig. 1h, i and Extended Data Fig. 2). Collectively, these experiments demonstrate transposon integration downstream of genomic target sites complementary to guide RNAs.

Protein requirements of RNA-guided DNA integration

To confirm the involvement of transposon- and CRISPR-associated proteins in catalysing RNA-guided DNA integration, we cloned and tested a series of plasmids in which each individual *tns* and *cas* gene was deleted, or in which the active site of each individual enzyme was mutated. Removal of any protein component abrogated transposition activity, as did mutations in the active site of the TnsB transposase, which catalyses DNA integration[23], the TnsC ATPase, which regulates target site selection[24], and the Cas6 RNase, which catalyses pre-crRNA processing[25] (Fig. 2a). A TnsA mutant that is catalytically impaired still facilitated RNA-guided DNA integration. On the basis of previous studies of *E. coli* Tn*7*, this variant system is expected to mobilize via replicative transposition as opposed to cut-and-paste transposition[26].

In *E. coli*, site-specific transposition requires *attTn7* binding by *Eco*TnsD, followed by interactions with the *Eco*TnsC regulator protein to directly recruit the *Eco*TnsA-TnsB-donor DNA[27]. Given the essential nature of *tniQ* (a *tnsD* homologue) in RNA-guided transposition, and its location within the *cas8-cas7-cas6* operon, we envisioned that the Cascade complex might directly bind TniQ and thereby deliver it to genomic target sites. We tested this hypothesis by recombinantly expressing CRISPR RNA and the *V. cholerae tniQ-cas8-cas7-cas6* operon containing an N-terminal His$_{10}$ tag on the TniQ subunit (Extended Data Fig. 3a). TniQ co-purified with Cas8, Cas7 and Cas6, as shown by SDS–PAGE and mass spectrometry analysis, and the relative band intensities for each Cas protein were similar to TniQ-free Cascade and consistent with the 1:6:1 Cas8:Cas7:Cas6 stoichiometry expected for a I-F variant Cascade complex[28] (Fig. 2b and Extended Data Fig. 3b). The complex migrated through a gel filtration column with an apparent molecular mass of roughly 440 kDa, in good agreement with its approximate expected mass, and both Cascade and TniQ–Cascade co-purified with a 60-nucleotide RNA species, which we confirmed was a mature crRNA by deep sequencing (Fig. 2c, d and Extended Data Fig. 3c, d). To validate the interaction between Cascade and TniQ further, we incubated separately purified samples in vitro and demonstrated complex formation by size-exclusion chromatography (Extended Data Fig. 3e). Together, these results reveal the existence of a novel TniQ–Cascade co-complex, highlighting a direct functional link between a CRISPR RNA-guided effector complex and a transposition protein.

To determine whether specific TniQ–Cascade interactions are required, or whether TniQ could direct transposition adjacent to generic R-loop structures or via artificial recruitment to DNA, we used *Streptococcus pyogenes* Cas9 (*Spy*Cas9)[29] and *Pseudomonas aeruginosa* Cascade (*Pae*Cascade)[28] as orthogonal RNA-guided DNA-targeting systems. After generating protein–RNA expression plasmids and programming both effector complexes with crRNAs that target the same *lacZ* sites as our earlier transposition experiments, we first validated DNA targeting by demonstrating efficient cell killing in the presence of an active Cas9 nuclease or the *Pae*Cascade-dependent Cas2-3 nuclease (Extended Data Fig. 4a, b). When we transformed strains containing pTnsABCQ and pDonor with a plasmid encoding either catalytically deactivated Cas9-sgRNA (dCas9-sgRNA) or *Pae*Cascade and performed PCR analysis of the resulting cell lysate, we found no evidence of site-specific transposition (Fig. 2e), indicating that a genomic R-loop is insufficient for site-specific integration. We also failed to detect transposition when TniQ was directly fused to either terminus of dCas9, or to the Cas8 or Cas6 subunit of *Pae*Cascade (Fig. 2e), at least for the linker sequences tested. Notably, however, a similar fusion of TniQ to the Cas6 subunit of *V. cholerae* Cascade, but not to the Cas8 subunit, restored RNA-guided transposition activity (Fig. 2e and Extended Data Fig. 4c).

Together with our biochemical results, we conclude that TniQ forms essential interactions with Cascade, possibly via the Cas6 subunit, which could account for our finding that RNA-guided DNA insertion occurs downstream of the PAM-distal end of the target site where Cas6 is bound[30,31] (Fig. 2f). Because TniQ is required for transposition, we propose that it serves as an important connection between the CRISPR- and transposon-associated machineries during DNA targeting and integration, although further biochemical and structural studies will be required to define these mechanistic steps in greater detail.

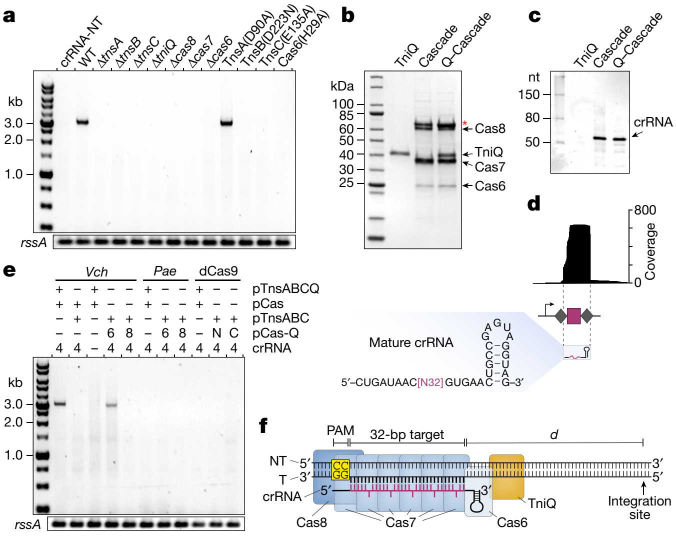


Fig. 2 | TniQ forms a complex with Cascade and is necessary for RNA-guided DNA integration. a, PCR analysis of transposition with crRNA-4 and a panel of gene deletions or point mutations, resolved by agarose gel electrophoresis. **b**, SDS–PAGE analysis of purified TniQ, Cascade and a TniQ–Cascade (Q–Cascade) co-complex. Asterisk denotes an HtpG contaminant. **c**, Denaturing urea–PAGE analysis of co-purifying nucleic acids. nt, nucleotides. **d**, Top, RNA sequencing analysis of RNA co-purifying with Cascade. Bottom, reads mapping to the CRISPR array reveal the mature crRNA sequence. **e**, PCR analysis of transposition experiments testing whether generic R-loop formation or artificial TniQ tethering can direct targeted integration. The *V. cholerae* transposon and TnsA-TnsB-TnsC were combined with DNA-targeting components that comprise *V. cholerae* (*Vch*) Cascade, *P. aeruginosa* (*Pae*) Cascade, or *S. pyogenes* dCas9-RNA (dCas9). TniQ was expressed either on its own from pTnsABCQ or as a fusion to the targeting complex (pCas-Q) at the Cas6 C terminus (6), Cas8 N terminus (8), or dCas9 N or C terminus. **f**, Schematic of the R-loop formed upon target DNA binding by Cascade, with the approximate position of each protein subunit denoted. The putative TniQ-binding site and the distance to the primary integration site are indicated. NT, non-target strand; T, target strand. For **a** and **e**, amplification of *rssA* serves as a loading control; gel source data are in Supplementary Fig. 1.

Donor requirements of RNA-guided DNA integration

To determine the minimal donor requirements for RNA-guided DNA integration, as well as the effects of truncating the transposon ends and altering the cargo size, we first developed a quantitative PCR (qPCR) method for scoring transposition efficiency that could accurately and sensitively measure genomic integration events in both orientations (Extended Data Fig. 5). Analysis of cell lysates from transposition experiments using *lacZ*-targeting crRNA-3 and crRNA-4 yielded overall integration efficiencies of 62% and 42% without selection, respectively. The preference for integrating the 'right' versus the 'left' transposon end proximal to the genomic site targeted by Cascade was 39-to-1 for crRNA-3 and 1-to-1 for crRNA-4, suggesting the existence of additional sequence determinants that regulate integration orientation (Fig. 3a, b).

With a quantitative assay in place, we were curious to investigate the effect of transposon size on RNA-guided integration efficiency and determine possible size constraints. When we progressively shortened or lengthened the DNA cargo in between the donor ends, beginning with our original mini-transposon donor plasmid (977 bp), we found that integration efficiency with our three-plasmid expression system was maximal with a 775-bp transposon and decayed with both the shorter and longer cargos tested (Fig. 3c). Interestingly, naturally occurring Tn*7*-like transposons that encode CRISPR–Cas systems range from

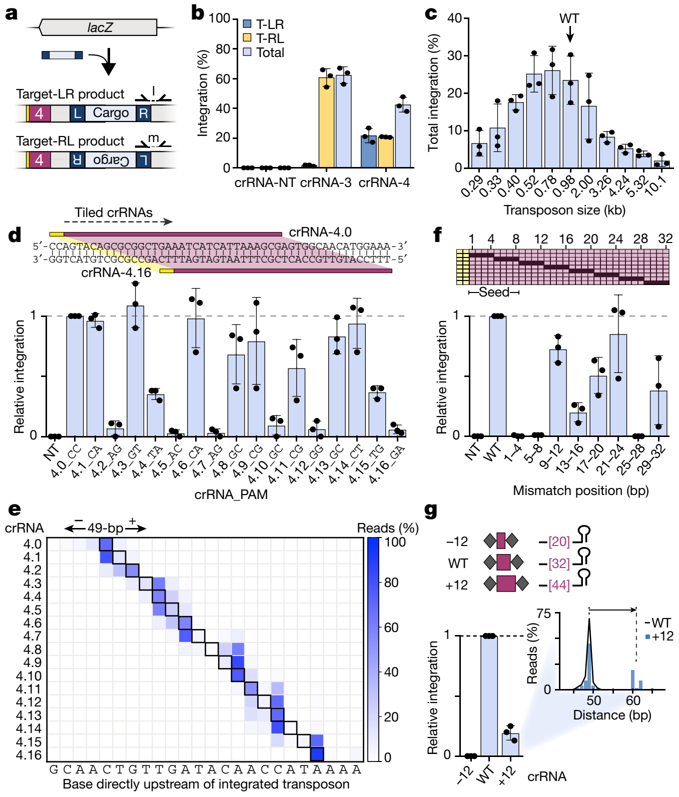


Fig. 3 | Influence of cargo size, PAM sequence, and crRNA mismatches on RNA-guided DNA integration. a, Schematic of alternative integration orientations and the primer pairs to selectively detect them by qPCR. **b**, qPCR-based quantification of transposition efficiency in both orientations with crRNA-NT, crRNA-3 and crRNA-4. T-LR and T-RL denote transposition products in which the transposon left end and right end are proximal to the target site, respectively. **c**, Total integration efficiency with crRNA-4 as a function of transposon size. The arrow denotes the wild-type (WT) pDonor used in most assays throughout this study. **d**, crRNAs were tiled along the *lacZ* gene in 1-bp increments relative to crRNA-4 (4.0) (top), and the resulting integration efficiencies were determined by qPCR (bottom). Data are normalized to crRNA-4.0, and the 2-nucleotide PAM for each crRNA is shown. **e**, Heat map showing the integration site distribution (*x* axis) for each of the tiled crRNAs (*y* axis) in **d**, determined by NGS. The 49-bp distance for each crRNA is denoted by a black box. **f**, crRNAs were mutated in 4-nucleotide blocks to introduce crRNA-target DNA mismatches (black, top), and the resulting integration efficiencies were determined by qPCR (bottom). Data are normalized to crRNA-4. **g**, The crRNA-4 spacer length was shortened or lengthened by 12 nucleotides (top), and the resulting integration efficiencies were determined by qPCR (bottom). Data are normalized to crRNA-4 (WT). The inset shows a comparison of integration site distributions for crRNA-4 and crRNA-4.+12, determined by NGS. Data in **b–d**, **f** and **g** are shown as mean ± s.d. for $n = 3$ biologically independent samples.

20 to more than 100 kb in size[14], although their capacity for active mobility is unknown.

We next separately truncated both ends of the transposon. We found that around 105 bp of the left end and 47 bp of the right end were absolutely crucial for efficient RNA-guided DNA integration, corresponding to three and two intact putative TnsB-binding sites, respectively (Extended Data Fig. 6). Shorter transposons containing right-end truncations were integrated more efficiently, accompanied by a notable change in the orientation bias.

These experiments reveal crucial parameters for the development of programmable DNA integration technology. Future efforts will be required to explore how transposition is affected by vector design, to what extent transposon end mutations are tolerated, and whether rational engineering allows for integration of larger cargos and/or greater control over integration orientation.

Guide RNA and target DNA requirements

The Tn*6677*-encoded CRISPR–Cas system is most closely related to the I-F subtype, in which DNA target recognition by Cascade requires a consensus 5′-CC-3′ PAM[22], a high degree of sequence complementarity within a PAM-proximal seed sequence[28], and additional base-pairing across the entire 32-bp protospacer[32]. To determine sequence determinants of RNA-guided DNA integration, we first tested 12 dinucleotide PAMs by sliding the guide sequence in 1-bp increments along the *lacZ* gene relative to crRNA-4 (Fig. 3d). In total, 8 distinct dinucleotide PAMs supported transposition at levels that were more than 25% of the 5′-CC-3′ PAM, and transposition occurred at over 1% total efficiency across the entire set of PAMs tested (Fig. 3d). Additional deep sequencing revealed that the distance between the Cascade target site and primary transposon insertion site remained fixed at approximately 47–51 bp across the panel of crRNAs tested, although interesting patterns emerged, suggesting an additional layer of insertion site preference that requires further investigation (Fig. 3e and Extended Data Fig. 7a). Nevertheless, these experiments highlight how PAM recognition plasticity can be harnessed to direct a high degree of insertion flexibility and specificity at base-pair resolution.

To probe the sensitivity of transposition to RNA–DNA mismatches, we tested consecutive blocks of 4-nucleotide mismatches along the guide portion of crRNA-4 (Fig. 3f). As expected from previous studies with Cascade homologues[33–35], mismatches within the 8-nucleotide seed sequence severely reduced transposition, probably owing to the inability to form a stable R-loop. Unexpectedly, however, our results highlighted a second region of mismatches at positions 25–28 that abrogated DNA integration, despite previous studies demonstrating that the stability of DNA binding is largely insensitive to mismatches in this region[33–35]. For the terminal mismatch block, which retained 17% integration activity, the distribution of observed insertion sites was markedly skewed to shorter distances from the target site relative to crRNA-4 (Extended Data Fig. 7b), which we hypothesize is the result of R-loop conformational heterogeneity.

Our emerging model for RNA-guided DNA integration involves Cascade-mediated recruitment of TniQ to target DNA. In the absence of any structural data, we realized that we could investigate whether TniQ may be positioned near the PAM-distal end of the R-loop by testing engineered crRNAs that contain spacers of variable lengths. Previous work with *E. coli* Cascade has demonstrated that crRNAs with extended spacers form complexes that contain additional Cas7 subunits[36], which would increase the distance between the PAM-bound Cas8 and the Cas6 at the other end of the R-loop. We therefore cloned and tested modified crRNAs containing spacers that were either shortened or lengthened in 6-nucleotide increments from the 3′ end. crRNAs with truncated spacers showed little or no activity, whereas extended spacers facilitated targeted integration, albeit at reduced levels with increasing length (Extended Data Fig. 7c, d). The +12-nucleotide crRNA directed transposition to two distinct regions: one approximately 49 bp from the 3′ end of the wild-type 32-nucleotide spacer, and an additional region shifted 11–13 bp away, in agreement with the expected increase in the length of the R-loop measured from the PAM (Fig. 3g). Although more experiments are required to deduce the underlying mechanisms that explain this bimodal distribution, as well as the insertion site distribution observed for other extended crRNAs, these data, together with the mismatch panel, provide further evidence that TniQ is tethered to the PAM-distal end of the R-loop structure.

Programmability and genome-wide specificity

We lastly sought to examine both the programmability and the genome-wide specificity of our RNA-guided DNA integration system. First, we cloned and tested a series of crRNAs targeting additional genomic sites flanked by 5′-CC-3′ PAMs within the *lac* operon. Using the same primer pair for each resulting cellular lysate, we showed by PCR analysis that transposition was predictably repositioned with each distinct crRNA (Fig. 4a).

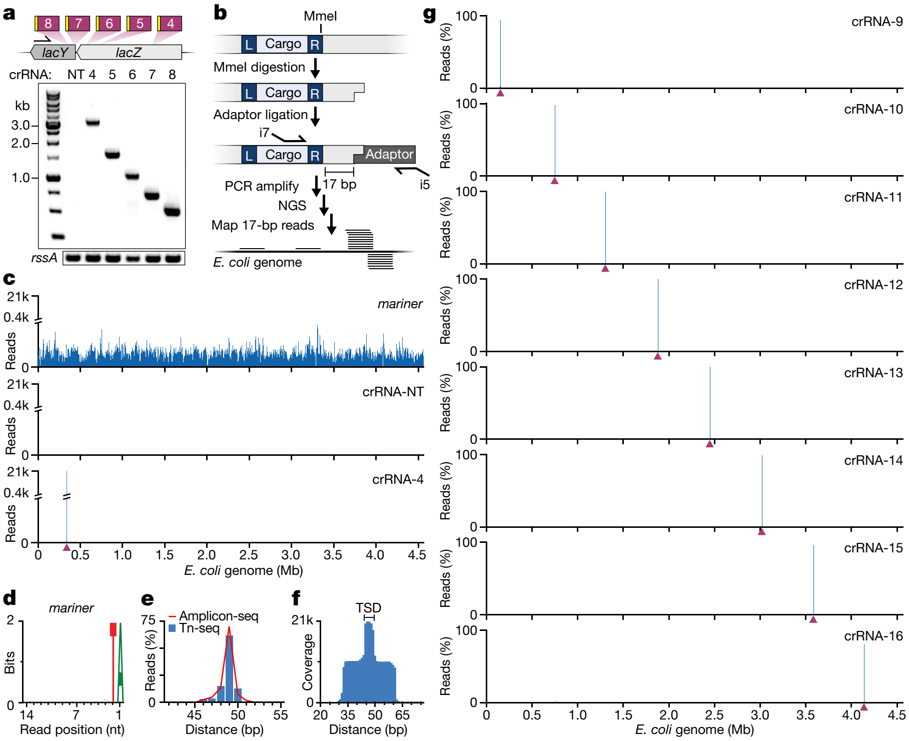


Fig. 4 | Genome-wide analysis of programmable RNA-guided DNA integration. a, Genomic locus targeted by crRNAs 4–8 (top), and PCR analysis of transposition resolved by agarose gel electrophoresis (bottom). Amplification of *rssA* serves as a loading control; gel source data may be found in Supplementary Fig. 1. **b**, Tn-seq workflow for deep sequencing of genome-wide transposition events. **c**, Mapped Tn-seq reads from transposition experiments with the *mariner* transposon, and with the *V. cholerae* transposon programmed with either crRNA-NT or crRNA-4. The crRNA-4 target site is denoted by a maroon triangle. **d**, Sequence logo of all *mariner* Tn-seq reads, highlighting the TA dinucleotide target-site preference. **e**, Comparison of integration site distributions for crRNA-4 determined by PCR amplicon sequencing and Tn-seq, for the T-RL product; the distance between the Cascade target site and transposon integration site is plotted. **f**, Zoomed-in view of Tn-seq read coverage at the primary integration site for experiments with crRNA-4, highlighting the 5-bp target-site duplication (TSD); the distance from the Cascade target site is plotted. **g**, Genome-wide distribution of genome-mapping Tn-seq reads from transposition experiments with crRNAs 9–16 for the *V. cholerae* transposon. The location of each target site is denoted by a maroon triangle.

Our experiments thus far specifically interrogated genomic loci containing the anticipated integration products, and it therefore remained possible that non-specific integration was simultaneously occurring elsewhere, either at off-target genomic sites bound by Cascade, or independently of Cascade targeting. We thus adopted a transposon insertion sequencing (Tn-seq) pipeline previously developed for *mariner* transposons[37,38], in which all integration sites genome-wide are revealed by NGS (Fig. 4b, Extended Data Fig. 8a, b and Methods). We first applied Tn-seq to a plasmid-encoded *mariner* transposon and found that our pipeline successfully recapitulated the genome-wide integration landscape previously observed with the Himar1c9 transposase[37,39] (Fig. 4c, d and Extended Data Fig. 8c, d).

When we performed the same analysis for the RNA-guided *V. cholerae* transposon programmed with crRNA-4, we observed exquisite selectivity for *lacZ*-specific DNA integration (Fig. 4c). The observed integration site, which accounted for 99.0% of all Tn-seq reads that passed our filtering criteria (Methods and Supplementary Table 4), precisely matched the site observed by previous PCR amplicon NGS analysis (Fig. 4e), and we did not observe reproducible off-target integration events elsewhere in the genome across three biological replicates (Extended Data Fig. 8e, f). Our Tn-seq data furthermore yielded diagnostic read pile-ups that highlighted the 5-bp target-site duplication and corroborated our previous measurements of transposon insertion orientation bias (Fig. 4f). Tn-seq libraries from *E. coli* strains expressing pQCascade programmed with the non-targeting crRNA, or from strains lacking Cascade altogether (but still containing pDonor and pTnsABCQ), yielded far fewer genome-mapping reads, and no integration sites were consistently observed across several biological replicates (Fig. 4c, Extended Data Fig. 8g, h and Supplementary Table 4).

In addition to performing Tn-seq with the crRNAs targeting *glmS* and *lacZ* genomic loci (Extended Data Fig. 9a), we cloned and tested an additional 16 crRNAs targeting the *E. coli* genome at 8 arbitrary locations spaced equidistantly around the circular chromosome. Beyond requiring that target sites were unique, were flanked by a 5′-CC-3′ PAM, and would direct DNA insertion to intergenic regions, we applied no further design rules or empirical selection criteria. Remarkably, when we analysed the resulting Tn-seq data, we found that 16 out of 16 crRNAs directed highly precise RNA-guided DNA integration 46–55 bp downstream of the Cascade target, with around 95% of all filtered Tn-seq reads mapping to the on-target insertion site (Fig. 4g and Extended Data Fig. 9b, c). These experiments highlight the high degree of intrinsic programmability and genome-wide integration specificity directed by transposon-encoded CRISPR–Cas systems.

Discussion

Transposases and integrases are generally thought to mobilize their specific genetic payloads either by integrating randomly, with a low degree of sequence specificity, or by targeting specialized genomic loci through inflexible, sequence-specific homing mechanisms[40]. We have discovered a fully programmable integrase, in which the DNA insertion activity of a heteromeric transposase from *V. cholerae* is directed by an RNA-guided complex known as Cascade, the DNA-targeting specificity of which can be easily tuned. Beyond defining fundamental parameters that govern this activity, our work also reveals a complex between Cascade and TniQ that mechanistically connects the transposon- and CRISPR-associated machineries. On the basis of our results, and of previous studies of Tn*7* transposition[15], we propose a model for the RNA-guided mobilization of Tn*7*-like transposons encoding CRISPR–Cas systems (Fig. 5). Because integration does not disrupt the

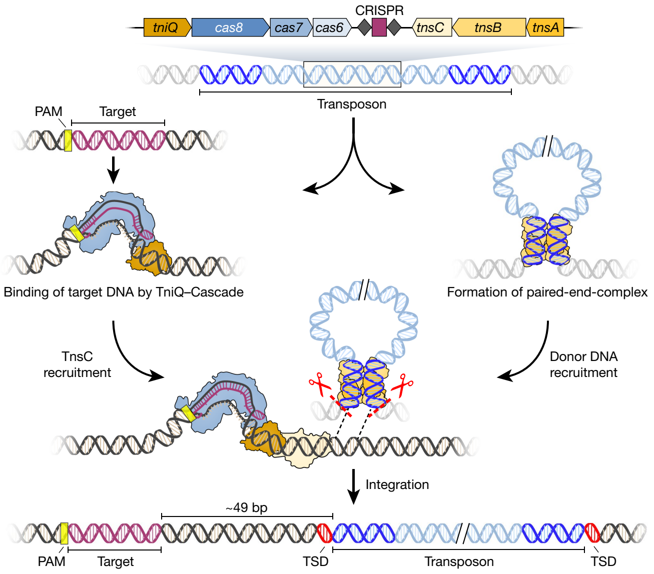


Fig. 5 | Proposed model for RNA-guided DNA integration by Tn*7*-like transposons encoding CRISPR–Cas systems. The *V. cholerae* Tn*6677* transposon encodes a programmable RNA-guided DNA-binding complex called Cascade, which we have shown forms a co-complex with TniQ. We propose that TniQ–Cascade complexes survey the cell for matching DNA target sites, which may be found on the host chromosome or mobile genetic elements. After target binding and R-loop formation, TniQ presumably recruits the non-sequence-specific DNA-binding protein TnsC, based on previous studies of *E. coli* Tn*7* (reviewed in ref. [15]). The transposon itself is bound at the left and right ends by TnsA and TnsB, forming a so-called paired-end complex that is recruited to the target DNA by TnsC. Excision of the transposon from its donor site allows for targeted integration at a fixed distance downstream of DNA-bound TniQ–Cascade, resulting in a 5-bp target-site duplication.

Cascade-binding site, an important question for future investigation is whether the *V. cholerae* transposon exhibits a similar mode of target immunity as *E. coli* Tn*7*[41], in which repeated transposition into the same genomic locus is prevented.

Almost all type I-F CRISPR–Cas systems within the *Vibrionaceae* family are associated with mobile genetic elements, and those found within Tn*7*-like transposons frequently co-occur with restriction-modification and type three secretion systems[14,20]. It is therefore tempting to speculate that RNA-guided DNA integration may facilitate sharing of innate immune systems and virulence mechanisms via horizontal gene transfer, particularly within marine environments[42]. Interestingly, we and others[43,44] recently observed a unique clade of type V CRISPR–Cas systems that also reside within bacterial transposons, which bear many of the same features as *V. cholerae* Tn*6677*: the presence of the *tniQ* gene, the lack of predicted DNA cleavage activity by the RNA-guided effector complex[45], and cargo genes that frequently include other innate immune systems (Extended Data Fig. 10). Although future experiments will be necessary to determine whether these systems also possess RNA-guided DNA integration activity, the bioinformatic evidence points to a more pervasive functional coupling between CRISPR–Cas systems and transposable elements than previously appreciated.

Many biotechnology products require genomic integration of large genetic payloads, including gene therapies[46], engineered crops[47] and biopharmaceuticals[48], and the advent of CRISPR-based genome editing has increased the need for effective knock-in methods. Yet current genome engineering solutions are limited by a lack of specificity, as with viral transduction[49], randomly integrating transposases[50] and non-homologous end joining[51] approaches, or by a lack of efficiency and cell-type versatility, as with homology-directed repair[52,53]. The ability to INsert Transposable Elements by Guide RNA-Assisted TargEting (INTEGRATE) offers an opportunity for site-specific DNA integration that would obviate the need for double-strand breaks in the target DNA, homology arms in the donor DNA, and host DNA repair factors. By virtue of its facile programmability, this technology could furthermore be leveraged for multiplexing and large-scale screening using guide RNA libraries. Together with other recent studies[54–57], our work highlights the far-reaching possibilities for genetic manipulation that continue to emerge from the diverse functions of CRISPR–Cas systems.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1323-z.

Received: 15 March 2019; Accepted: 4 June 2019;
Published online: 12 June 2019

1. Thomas, C. M. & Nielsen, K. M. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol.* **3**, 711–721 (2005).
2. Soucy, S. M., Huang, J. & Gogarten, J. P. Horizontal gene transfer: building the web of life. *Nat. Rev. Genet.* **16**, 472–482 (2015).
3. Koonin, E. V. The turbulent network dynamics of microbial evolution and the statistical tree of life. *J. Mol. Evol.* **80**, 244–250 (2015).
4. Toussaint, A. & Chandler, M. Prokaryote genome fluidity: toward a system approach of the mobilome. *Methods Mol. Biol.* **804**, 57–80 (2012).
5. Dy, R. L., Richter, C., Salmond, G. P. C. & Fineran, P. C. Remarkable mechanisms in microbes to resist phage infections. *Annu. Rev. Virol.* **1**, 307–331 (2014).
6. Hille, F. et al. The biology of CRISPR-Cas: backward and forward. *Cell* **172**, 1239–1259 (2018).
7. Doron, S. et al. Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* **359**, eaar4120 (2018).
8. Koonin, E. V., Makarova, K. S. & Wolf, Y. I. Evolutionary genomics of defense systems in archaea and bacteria. *Annu. Rev. Microbiol.* **71**, 233–261 (2017).
9. Koonin, E. V. & Makarova, K. S. Mobile genetic elements and evolution of CRISPR-Cas systems: all the way there and back. *Genome Biol. Evol.* **9**, 2812–2825 (2017).
10. Broecker, F. & Moelling, K. Evolution of immune systems from viruses and transposable elements. *Front. Microbiol.* **10**, 51 (2019).
11. Kapitonov, V. V., Makarova, K. S. & Koonin, E. V. ISC, a novel group of bacterial and archaeal DNA transposons that encode Cas9 homologs. *J. Bacteriol.* **198**, 797–807 (2016).
12. Shmakov, S. et al. Discovery and functional characterization of diverse class 2 CRISPR-Cas systems. *Mol. Cell* **60**, 385–397 (2015).
13. Krupovic, M., Béguin, P. & Koonin, E. V. Casposons: mobile genetic elements that gave rise to the CRISPR-Cas adaptation machinery. *Curr. Opin. Microbiol.* **38**, 36–43 (2017).
14. Peters, J. E., Makarova, K. S., Shmakov, S. & Koonin, E. V. Recruitment of CRISPR-Cas systems by Tn7-like transposons. *Proc. Natl Acad. Sci. USA* **114**, E7358–E7366 (2017).
15. Peters, J. E. Tn*7*. *Microbiol. Spectr.* **2**, MDNA3-0010-2014 (2014).
16. Waddell, C. S. & Craig, N. L. Tn7 transposition: two transposition pathways directed by five Tn7-encoded genes. *Genes Dev.* **2**, 137–149 (1988).
17. Lichtenstein, C. & Brenner, S. Unique insertion site of Tn7 in the *E. coli* chromosome. *Nature* **297**, 601–603 (1982).
18. McKown, R. L., Orle, K. A., Chen, T. & Craig, N. L. Sequence requirements of *Escherichia coli* attTn7, a specific site of transposon Tn7 insertion. *J. Bacteriol.* **170**, 352–358 (1988).
19. Parks, A. R. et al. Transposition into replicating DNA occurs through interaction with the processivity factor. *Cell* **138**, 685–695 (2009).
20. McDonald, N. D., Regmi, A., Morreale, D. P., Borowski, J. D. & Boyd, E. F. CRISPR-Cas systems are present predominantly on mobile genetic elements in *Vibrio* species. *BMC Genomics* **20**, 105 (2019).
21. Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Classification and nomenclature of CRISPR-Cas systems: where from here? *CRISPR J.* **1**, 325–336 (2018).
22. Rollins, M. F., Schuman, J. T., Paulus, K., Bukhari, H. S. T. & Wiedenheft, B. Mechanism of foreign DNA recognition by a CRISPR RNA-guided surveillance complex from *Pseudomonas aeruginosa*. *Nucleic Acids Res.* **43**, 2216–2222 (2015).
23. Sarnovsky, R. J., May, E. W. & Craig, N. L. The Tn7 transposase is a heteromeric complex in which DNA breakage and joining activities are distributed between different gene products. *EMBO J.* **15**, 6348–6361 (1996).
24. Stellwagen, A. E. & Craig, N. L. Gain-of-function mutations in TnsC, an ATP-dependent transposition protein that activates the bacterial transposon Tn7. *Genetics* **145**, 573–585 (1997).
25. Haurwitz, R. E., Jinek, M., Wiedenheft, B., Zhou, K. & Doudna, J. A. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science* **329**, 1355–1358 (2010).
26. May, E. W. & Craig, N. L. Switching from cut-and-paste to replicative Tn7 transposition. *Science* **272**, 401–404 (1996).

27. Choi, K. Y., Spencer, J. M. & Craig, N. L. The Tn7 transposition regulator TnsC interacts with the transposase subunit TnsB and target selector TnsD. *Proc. Natl Acad. Sci. USA* **111**, E2858–E2865 (2014).
28. Wiedenheft, B. et al. RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *Proc. Natl Acad. Sci. USA* **108**, 10092–10097 (2011).
29. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
30. Wiedenheft, B. et al. Structures of the RNA-guided surveillance complex from a bacterial immune system. *Nature* **477**, 486–489 (2011).
31. Guo, T. W. et al. Cryo-EM structures reveal mechanism and inhibition of DNA targeting by a CRISPR-Cas surveillance complex. *Cell* **171**, 414–426.e12 (2017).
32. Xue, C. & Sashital, D. G. Mechanisms of type I-E and I-F CRISPR-Cas systems in *Enterobacteriaceae*. *EcoSal Plus* **8**, ESP-0008-2018 (2019).
33. Blosser, T. R. et al. Two distinct DNA binding modes guide dual roles of a CRISPR-Cas protein complex. *Mol. Cell* **58**, 60–70 (2015).
34. Cooper, L. A., Stringer, A. M. & Wade, J. T. Determining the specificity of cascade binding, interference, and primed adaptation *in vivo* in the *Escherichia coli* type I-E CRISPR-Cas system. *MBio* **9**, e02100-17 (2018).
35. Rutkauskas, M. et al. Directional R-loop formation by the CRISPR-Cas surveillance complex cascade provides efficient off-target site rejection. *Cell Reports* **10**, 1534–1543 (2015).
36. Luo, M. L. et al. The CRISPR RNA-guided surveillance complex in *Escherichia coli* accommodates extended RNA spacers. *Nucleic Acids Res.* **44**, 7385–7394 (2016).
37. Goodman, A. L. et al. Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host Microbe* **6**, 279–289 (2009).
38. van Opijnen, T., Bodi, K. L. & Camilli, A. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat. Methods* **6**, 767–772 (2009).
39. Wiles, T. J. et al. Combining quantitative genetic footprinting and trait enrichment analysis to identify fitness determinants of a bacterial pathogen. *PLoS Genet.* **9**, e1003716 (2013).
40. Craig, N. L., Craigie, R., Gellert, M. & Lambowitz, A. M. *Mobile DNA III* (2014).
41. Stellwagen, A. E. & Craig, N. L. Avoiding self: two Tn7-encoded proteins mediate target immunity in Tn7 transposition. *EMBO J.* **16**, 6823–6834 (1997).
42. Sobecky, P. A. & Hazen, T. H. Horizontal gene transfer and mobile genetic elements in marine systems. *Methods Mol. Biol.* **532**, 435–453 (2009).
43. Makarova, K. S. Beyond the adaptive immunity: sub- and neofunctionalization of CRISPR–Cas systems and their components. Paper presented at: CRISPR 2018 Meeting; Jun 20; Vilnius, Lithuania. (2018).
44. Cheng, D. R., Yan, W. X. & Scott, D. A. Discovery of Type VI-D CRISPR-Cas Systems. Paper presented at: CRISPR 2018 Meeting; Jun 21; Vilnius, Lithuania. (2018).
45. Shmakov, S. et al. Diversity and evolution of class 2 CRISPR–Cas systems. *Nat. Rev. Microbiol.* **15**, 169–182 (2017).
46. Dunbar, C. E. et al. Gene therapy comes of age. *Science* **359**, eaan4672 (2018).
47. Gelvin, S. B. Integration of agrobacterium T-DNA into the plant genome. *Annu. Rev. Genet.* **51**, 195–217 (2017).
48. Wurm, F. M. Production of recombinant protein therapeutics in cultivated mammalian cells. *Nat. Biotechnol.* **22**, 1393–1398 (2004).
49. Kvaratskhelia, M., Sharma, A., Larue, R. C., Serrao, E. & Engelman, A. Molecular mechanisms of retroviral integration site selection. *Nucleic Acids Res.* **42**, 10209–10225 (2014).
50. Di Matteo, M., Belay, E., Chuah, M. K. & Vandendriessche, T. Recent developments in transposon-mediated gene therapy. *Expert Opin. Biol. Ther.* **12**, 841–858 (2012).
51. Zelensky, A. N., Schimmel, J., Kool, H., Kanaar, R. & Tijsterman, M. Inactivation of Pol θ and C-NHEJ eliminates off-target integration of exogenous DNA. *Nat. Commun.* **8**, 66 (2017).
52. Cox, D. B. T., Platt, R. J. & Zhang, F. Therapeutic genome editing: prospects and challenges. *Nat. Med.* **21**, 121–131 (2015).
53. Pawelczak, K. S., Gavande, N. S., VanderVere-Carozza, P. S. & Turchi, J. J. Modulating DNA repair pathways to improve precision genome engineering. *ACS Chem. Biol.* **13**, 389–396 (2018).
54. Schmidt, F., Cherepkova, M. Y. & Platt, R. J. Transcriptional recording by CRISPR spacer acquisition from RNA. *Nature* **562**, 380–385 (2018).
55. Myhrvold, C. et al. Field-deployable viral diagnostics using CRISPR-Cas13. *Science* **360**, 444–448 (2018).
56. Yan, W. X. et al. Functionally diverse type V CRISPR-Cas systems. *Science* **363**, 88–91 (2019).
57. Harrington, L. B. et al. Programmed DNA destruction by miniature CRISPR-Cas14 enzymes. *Science* **362**, 839–842 (2018).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Plasmid construction. All plasmids used in this study are described in Supplementary Table 1, and a subset is available from Addgene. In brief, genes encoding *V. cholerae* strain HE-45 TnsA-TnsB-TnsC and TniQ-Cas8-Cas7-Cas6 (Supplementary Table 2 and Supplementary Figs. 2–8) were synthesized by GenScript and cloned into pCOLADuet-1 and pCDFDuet-1, respectively, yielding pTnsABC and pQCascadeΔCRISPR. A pQCascade entry vector (pQCascade_entry) was generated by inserting tandem BsaI restriction sites flanked by two CRISPR repeats downstream of the first T7 promoter, and specific spacers (Supplementary Table 3) were subsequently cloned by oligoduplex ligation, yielding pQCascade. To generate pDonor, a gene fragment (GenScript) encoding both transposon ends was cloned into pUC19, and a chloramphenicol-resistance gene was subsequently inserted within the mini-transposon. Further derivatives of these plasmids were cloned using a combination of methods, including Gibson assembly, restriction digestion-ligation, ligation of hybridized oligonucleotides, and around-the-horn PCR. Plasmids were cloned and propagated in NEB Turbo cells (NEB), purified using Miniprep Kits (Qiagen), and verified by Sanger sequencing (GENEWIZ).

For transposition experiments involving the *E. coli* Tn*7* transposon, pEcoDonor was generated similarly to pDonor, and pEcoTnsABCD was subcloned from pCW4 (a gift from N. Craig, Addgene plasmid 8484). For transposition and cell killing experiments involving the I-F system from *P. aeruginosa*, genes encoding Cas8-Cas5-Cas7-Cas6 (also known as Csy1-Csy2-Csy3-Csy4) were subcloned from pBW64 (a gift from B. Wiedenheft), and the gene encoding the natural Cas2-3 fusion protein was subcloned from pCas1_Cas2/3 (a gift from B. Wiedenheft, Addgene plasmid 89240). For transposition and cell killing experiments involving the II-A system from *S. pyogenes*, the gene encoding Cas9 was subcloned from a vector in-house. For control Tn-seq experiments using the *mariner* transposon and Himar1C9 transposase, the relevant portions were subcloned from pSAM_Ec (a gift from M. Mulvey, Addgene plasmid 102939).

Expression plasmids for protein purification were subcloned from pQCascade into p2CT-10 (a gift from the QB3 MacroLab, Addgene plasmid 55209), and the crRNA expression construct was cloned into pACYCDuet-1.

Multiple sequence alignments (Supplementary Figs. 2–8) were performed using Clustal Omega with default parameters and visualized with ESPript 3.0[58]. Analysis of spacers from C2c5 CRISPR arrays (Extended Data Fig. 10) was performed using CRISPRTarget[59].

Transposition experiments. All transposition experiments were performed in *E. coli* BL21(DE3) cells (NEB). For experiments including pDonor, pTnsABC and pQCascade (or variants thereof), chemically competent cells were first co-transformed with pDonor and pTnsABC, pDonor and pQCascade, or pTnsABC and pQCascade, and transformants were isolated by selective plating on double antibiotic LB-agar plates. Liquid cultures were then inoculated from single colonies, and the resulting strains were made chemically competent using standard methods, aliquoted and snap frozen. The third plasmid was introduced in a new transformation reaction by heat shock, and after recovering cells in fresh LB medium at 37 °C for 1 h, cells were plated on triple antibiotic LB-agar plates containing 100 μg ml^{-1} carbenicillin, 50 μg ml^{-1} kanamycin, and 50 μg ml^{-1} spectinomycin. After overnight growth at 37 °C for 16 h, hundreds of colonies were scraped from the plates, and a portion was resuspended in fresh LB medium before being re-plated on triple antibiotic LB-agar plates as before, this time supplemented with 0.1 mM IPTG to induce protein expression. Solid media culturing was chosen over liquid culturing in order to avoid growth competition and population bottlenecks. Cells were incubated an additional 24 h at 37 °C and typically grew as densely spaced colonies, before being scraped, resuspended in LB medium, and prepared for subsequent analysis. Control experiments lacking one or more molecular components were performed using empty vectors and the exact same protocol as above. Experiments investigating the effect of induction level on transposition efficiency contained variable IPTG concentrations in the media (Extended Data Fig. 5d). To isolate clonal, *lacZ*-integrated strains via blue-white colony screening, cells were re-plated on triple antibiotic LB-agar plates supplemented with 1 mM IPTG and 100 μg ml^{-1} X-gal (GoldBio), and grown overnight at 37 °C before colony PCR analysis.

PCR and Sanger sequencing analysis of transposition products. Optical density measurements at 600 nm were taken of scraped colonies that had been resuspended in LB medium, and approximately 3.2×10^8 cells (the equivalent of 200 μl of $\mathrm{OD}_{600} = 2.0$) were transferred to a 96-well plate. Cells were pelleted by centrifugation at 4,000*g* for 5 min and resuspended in 80 μl of H_2O, before being lysed by incubating at 95 °C for 10 min in a thermal cycler. The cell debris was pelleted by centrifugation at 4,000*g* for 5 min, and 10 μl of lysate supernatant was removed and serially diluted with 90 μl of H_2O to generate 10- and 100-fold lysate dilutions for qPCR and PCR analysis, respectively.

PCR products were generated with Q5 Hot Start High-Fidelity DNA Polymerase (NEB) using 5 μl of 100-fold diluted lysate per 12.5 μl reaction volume serving as template. Reactions contained 200 μM dNTPs and 0.5 μM primers, and were generally subjected to 30 thermal cycles with an annealing temperature of 66 °C. Primer pairs contained one genome-specific primer and one transposon-specific primer, and were varied such that all possible integration orientations could be detected both upstream and downstream of the target site (see Supplementary Table 5 for selected oligonucleotides used in this study). Colony PCRs (Extended Data Fig. 2b) were performed by inoculating overnight cultures with individual colonies and performing PCR analysis as described above. PCR amplicons were resolved by 1–2% agarose gel electrophoresis and visualized by staining with SYBR Safe (Thermo Scientific). Negative control samples were always analysed in parallel with experimental samples to identify mispriming products, some of which presumably result from the analysis being performed on crude cell lysates that still contain the high-copy pDonor. PCRs were initially performed with different DNA polymerases, variable cycling conditions, and different sample preparation methods. We note that higher concentrations of the crude lysate appeared to inhibit successful amplification of the integrated transposition product.

To map integration sites by Sanger sequencing, bands were excised after separation by gel electrophoresis, DNA was isolated by Gel Extraction Kit (Qiagen), and samples were submitted to and analysed by GENEWIZ.

Integration site distribution analysis by NGS of PCR amplicons. PCR-1 products were generated as described above, except that primers contained universal Illumina adaptors as 5′ overhangs (Supplementary Table 5) and the cycle number was reduced to 20. These products were then diluted 20-fold into a fresh polymerase chain reaction (PCR-2) containing indexed p5/p7 primers and subjected to 10 additional thermal cycles using an annealing temperature of 65 °C. After verifying amplification by analytical gel electrophoresis, barcoded reactions were pooled and resolved by 2% agarose gel electrophoresis, DNA was isolated by Gel Extraction Kit (Qiagen), and NGS libraries were quantified by qPCR using the NEBNext Library Quant Kit (NEB). Illumina sequencing was performed using a NextSeq mid output kit with 150-cycle single-end reads and automated demultiplexing and adaptor trimming (Illumina). Individual bases with Phred quality scores under 20 (corresponding to a base miscalling rate of >1%) were changed to 'N,' and only reads with at least half the called bases above Q20 were retained for subsequent analysis.

To determine the integration site distribution for a given sample, the following steps were performed using custom Python scripts. First, reads were filtered based on the requirement that they contain 20 bp of perfectly matching transposon end sequence. Fifteen base pairs of sequence immediately flanking the transposon were then extracted and aligned to a 1-kb window of the *E. coli* BL21(DE3) genome (GenBank accession CP001509) surrounding the crRNA-matching genomic target site. The distance between the nearest transposon–genome junction and the PAM-distal edge of the 32-bp target site was determined. Histograms were plotted after compiling these distances across all the reads within a given library (see Supplementary Table 4 for NGS statistics).

Cell killing experiments. For experiments with Cas9, 40 μl chemically competent BL21(DE3) cells were transformed with 100 ng Cas9-sgRNA expression plasmid encoding either sgRNA-3 or sgRNA-4, which target equivalent *lacZ* sites as *V. cholerae* crRNA-3 or crRNA-4 but on opposite strands, or a truncated/non-functional sgRNA derived from the BsaI-containing entry vector (Supplementary Table 3). After a one-hour recovery at 37 °C, variable dilutions of cells were plated on LB-agar plates containing 100 μg ml^{-1} carbenicillin and 0.1 mM IPTG and grown an additional 16 h at 37 °C. The number of resulting colonies was quantified across three biological replicates, and the data were plotted as colony-forming units per microgram of plasmid DNA. Additional control experiments used an expression plasmid encoding Cas9 nuclease-inactivating D10A and H840A mutations (dCas9).

For experiments with Cascade and Cas2-3 from *P. aeruginosa*, BL21(DE3) cells were first transformed with a Cas2-3 expression vector, and the resulting strains were made chemically competent. Forty microlitres of these cells were then transformed with 100 ng *Pae*Cascade expression plasmid encoding either crRNA-Pae3 or crRNA-Pae4, which target equivalent *lacZ* sites as *V. cholerae* crRNA-3 or crRNA-4, or a truncated/non-functional crRNA derived from the BsaI-containing entry vector (Supplementary Table 3). After a one-hour recovery at 37 °C, variable dilutions of cells were plated on LB-agar plates containing 100 μg ml^{-1} carbenicillin and 50 μg ml^{-1} kanamycin and grown an additional 16 h at 37 °C. The number of resulting colonies was quantified across three biological replicates, and the data were plotted as colony-forming units per microgram of plasmid DNA. We found that even low concentrations of IPTG led to crRNA-independent toxicity in these experiments, whereas crRNA-dependent cell killing was readily observed in the absence of induction, presumably from leaky expression by T7 RNAP. We therefore omitted IPTG from experiments using *Pae*Cascade and Cas2-3.

qPCR analysis of transposition efficiency. For both crRNA-3 and crRNA-4, pairs of transposon- and genome-specific primers were designed to amplify an approximately 140–240-bp fragment resulting from RNA-guided DNA integration at the expected *lacZ* locus in either orientation. A separate pair of genome-specific primers was designed to amplify an *E. coli* reference gene (*rssA*) for normalization purposes (Supplementary Table 5). qPCR reactions (10 μl) contained 5 μl of SsoAdvanced Universal SYBR Green Supermix (BioRad), 1 μl H_2O, 2 μl of 2.5 μM primers, and 2 μl of tenfold diluted lysate prepared from scraped colonies, as described for the PCR analysis above. Reactions were prepared in 384-well clear/white PCR plates (BioRad), and measurements were performed on a CFX384 Real-Time PCR Detection System (BioRad) using the following thermal cycling parameters: polymerase activation and DNA denaturation (98 °C for 2.5 min), 40 cycles of amplification (98 °C for 10 s, 62 °C for 20 s), and terminal melt-curve analysis (65–95 °C in 0.5 °C per 5 s increments).

We first prepared lysates from a control BL21(DE3) strain containing pDonor and both empty expression vectors (pCOLADuet-1 and pCDFDuet-1), and from strains that underwent clonal integration into the *lacZ* locus downstream of both crRNA-3 and crRNA-4 target sites in both orientations. By testing our primer pairs with each of these samples diluted across five orders of magnitude, and then determining the resulting C_q values and PCR efficiencies, we verified that our experimental and reference amplicons were amplified with similar efficiencies, and that our primer pairs selectively amplified the intended transposition product (Extended Data Fig. 5a, b). We next simulated variable transposition efficiencies across five orders of magnitude (ranging from 0.002 to 100%) by mixing control lysates and clonally-integrated lysates in various ratios, and showed that we could accurately and reproducibly detect transposition products at both target sites, in either orientation, at levels >0.01% (Extended Data Fig. 5b). Finally, we simulated variable integration orientation biases by mixing clonally-integrated lysates together in varying ratios together with control lysates, and showed that these could also be accurately measured (Extended Data Fig. 5c).

In our final qPCR analysis protocol, each biological sample is analysed in three parallel reactions: one reaction contains a primer pair for the *E. coli* reference gene, a second reaction contains a primer pair for one of the two possible integration orientations, and a third reaction contains a primer pair for the other possible integration orientation. Transposition efficiency for each orientation is then calculated as $2^{\Delta C_q}$, in which ΔC_q is the C_q difference between the experimental reaction and the control reaction. Total transposition efficiency for a given experiment is calculated as the sum of transposition efficiencies for both orientations. All measurements presented in the text and figures were determined from three independent biological replicates.

We note that experiments with pDonor variants were performed by delivering pDonor in the final transformation step, whereas most other experiments were performed by delivering pQCascade in the final transformation step. Integration efficiencies between samples from these two experiments appeared to differ slightly as a result (compare Fig. 3b with Fig. 3c). Additionally, because we did not want to bias our qPCR analysis of the donor end truncation samples by successively shortening the PCR amplicon, different primer pairs were used for these samples. Within the left and right end truncation panel (Extended Data Fig. 6b, c), the transposon end that was not being perturbed was selectively amplified during qPCR analysis.

Recombinant protein expression and purification. The protein components for Cascade, TniQ and TniQ–Cascade were expressed from a pET-derivative vector containing an N-terminal His_{10}-MBP-TEVsite fusion on Cas8, TniQ and TniQ, respectively (see Extended Data Fig. 3a). The crRNAs for Cascade and TniQ–Cascade were expressed separately from a pACYC-derivative vector (Supplementary Table 1). *E. coli* BL21(DE3) cells containing one or both plasmids were grown in 2xYT medium with the appropriate antibiotic(s) at 37 °C to $OD_{600} = 0.5–0.7$, at which point IPTG was added to a final concentration of 0.5 mM and growth was allowed to continue at 16 °C for an additional 12–16 h. Cells were harvested by centrifugation at 4,000*g* for 20 min at 4 °C.

Cascade and TniQ–Cascade were purified as follows. Cell pellets were resuspended in Cascade lysis buffer (50 mM Tris-Cl, pH 7.5, 100 mM NaCl, 0.5 mM PMSF, EDTA-free Protease Inhibitor Cocktail tablets (Roche), 1 mM dithiothreitol (DTT), 5% glycerol) and lysed by sonication with a sonic dismembrator (Fisher) set to 40% amplitude and 12 min total process time (cycles of 10 s on and 20 s off, for a total of 4 min on and 8 min off). Lysates were clarified by centrifugation at 15,000*g* for 30 min at 4 °C. Initial purification was performed by immobilized metal-ion affinity chromatography with NiNTA Agarose (Qiagen) using NiNTA wash buffer (50 mM Tris-Cl, pH 7.5, 100 mM NaCl, 10 mM imidazole, 1 mM DTT, 5% glycerol) and NiNTA elution buffer (50 mM Tris-Cl pH 7.5, 100 mM NaCl, 300 mM imidazole, 1 mM DTT, 5% glycerol). The His_{10}-MBP fusion was removed by incubation with TEV protease overnight at 4 °C in NiNTA elution buffer, and complexes were further purified by anion exchange chromatography on an AKTApure system (GE Healthcare) using a 5 ml HiTrap Q HP Column (GE Healthcare) with a linear gradient from 100% buffer A (20 mM Tris-Cl, pH 7.5, 100 mM NaCl, 1 mM DTT, 5% glycerol) to 100% buffer B (20 mM Tris-Cl, pH 7.5, 1 M NaCl, 1 mM DTT, 5% glycerol) over 20 column volumes. Pooled fractions were identified by SDS–PAGE analysis and concentrated, and the sample was further refined by size exclusion chromatography over one or two tandem Superose 6 Increase 10/300 columns (GE Healthcare) equilibrated with Cascade storage buffer (20 mM Tris-Cl, pH 7.5, 200 mM NaCl, 1 mM DTT, 5% glycerol). Fractions were pooled, concentrated, snap frozen in liquid nitrogen, and stored at −80 °C.

TniQ was purified similarly, except the lysis, NiNTA wash, and NiNTA elution buffers contained 500 mM NaCl instead of 100 mM NaCl. Separation by ion exchange chromatography was performed on a 5 ml HiTrap SP HP Column (GE Healthcare) using the same buffer A and buffer B as above, and the final size-exclusion chromatography step was performed on a HiLoad Superdex 75 16/600 column (GE Healthcare) in Cascade storage buffer. The TniQ protein used in TniQ–Cascade binding experiments (Extended Data Fig. 3e) contained an N-terminal StrepII tag (Supplementary Table 1).

Mass spectrometry analysis. Total protein (0.5–5 μg) was separated on 4–20% gradient SDS–PAGE and stained with Imperial Protein Stain (Thermo Scientific). In-gel digestion was performed essentially as described[60], with minor modifications. Protein gel slices were excised, washed with 1:1 acetonitrile:100 mM ammonium bicarbonate (v/v) for 30 min, dehydrated with 100% acetonitrile for 10 min, and dried in a speed-vac for 10 min without heat. Gel slices were reduced with 5 mM DTT for 30 min at 56 °C and then alkylated with 11 mM iodoacetamide for 30 min at room temperature in the dark. Gel slices were washed with 100 mM ammonium bicarbonate and 100% acetonitrile for 10 min each, and excess acetonitrile was removed by drying in a speed-vac for 10 min without heat. Gel slices were then rehydrated in a solution of 25 ng μl^{-1} trypsin in 50 mM ammonium bicarbonate for 30 min on ice, and trypsin digestions were performed overnight at 37 °C. Digested peptides were collected and further extracted from gel slices in mass spectrometry (MS) extraction buffer (1:2 5% formic acid:acetonitrile (v/v)) with high-speed shaking. Supernatants were dried down in a speed-vac, and peptides were dissolved in a solution containing 3% acetonitrile and 0.1% formic acid.

Desalted peptides were injected onto an EASY-Spray PepMap RSLC C18 50 cm × 75 μm column (Thermo Scientific), which was coupled to the Orbitrap Fusion Tribrid mass spectrometer (Thermo Scientific). Peptides were eluted with a non-linear 100-min gradient of 5–30% mass spectrometry buffer B (MS buffer A: 0.1% (v/v) formic acid in water; MS buffer B: 0.1% (v/v) formic acid in acetonitrile) at a flow rate of 250 nl min^{-1}. Survey scans of peptide precursors were performed from 400 to 1,575 *m*/*z* at 120K full width at half-maximum resolution (at 200 *m*/*z*) with a 2×10^5 ion count target and a maximum injection time of 50 ms. The instrument was set to run in top speed mode with 3-s cycles for the survey and the tandem mass spectrometry (MS/MS) scans. After a survey scan, tandem mass spectrometry was performed on the most abundant precursors exhibiting a charge state from 2 to 6 of greater than 5×10^3 intensity by isolating them in the quadrupole at 1.6 Th. CID fragmentation was applied with 35% collision energy, and resulting fragments were detected using the rapid scan rate in the ion trap. The AGC target for MS/MS was set to 1×10^4 and the maximum injection time limited to 35 ms. The dynamic exclusion was set to 45 s with a 10 ppm mass tolerance around the precursor and its isotopes. Monoisotopic precursor selection was enabled.

Raw mass spectrometric data were processed and searched using the Sequest HT search engine within the Proteome Discoverer 2.2 software (Thermo Scientific) with custom sequences and the reference *E. coli* BL21(DE3) strain database downloaded from Uniprot. The default search settings used for protein identification were as follows: two mis-cleavages for full trypsin, with fixed carbamidomethyl modification of cysteine and oxidation of methionine; deamidation of asparagine and glutamine and acetylation on protein N termini were used as variable modifications. Identified peptides were filtered for a maximum 1% false discovery rate using the Percolator algorithm, and the PD2.2 output combined folder was uploaded in Scaffold (Proteome Software) for data visualization. Spectral counting was used for analysis to compare the samples.

crRNA analysis and RNA sequencing. To analyse the nucleic acid component co-purifying with Cascade and TniQ–Cascade, nucleic acids were isolated by phenol-chloroform extraction, resolved by 10% denaturing urea–PAGE, and visualized by staining with SYBR Gold (Thermo Scientific). Analytical RNase and DNase digestions were performed in 10 μl reactions with approximately 4 pmol nucleic acid and either 10 μg RNase A (Thermo Scientific) or 2 U DNase I (NEB), and were analysed by 10% denaturing urea–PAGE and SYBR Gold staining.

RNA sequencing was performed generally as previously described[61]. In brief, RNA was isolated from Cascade and TniQ–Cascade complexes by phenol-chloroform extraction, ethanol precipitated, and 5′-phosphorylated/3′-dephosphorylated using T4 polynucleotide kinase (NEB), followed by clean-up using the ssDNA/RNA Clean & Concentrator Kit (Zymo Research). A ssDNA universal Illumina adaptor containing 5′-adenylation and 3′-dideoxycytidine modifications (Supplementary Table 5) was ligated to the 3′ end with T4 RNA Ligase 1 (NEB), followed by hybridization of a ssDNA reverse transcriptase primer and ligation of ssRNA universal Illumina adaptor to the 5′ end with T4 RNA Ligase 1 (NEB). cDNA was synthesized using Maxima H Minus Reverse Transcriptase

(Thermo Scientific), followed by PCR amplification using indexed p5/p7 primers. Illumina sequencing was performed using a NextSeq mid output kit with 150-cycle single-end reads and automated demultiplexing and adaptor trimming (Illumina). Individual bases with Phred quality scores under 20 (corresponding to a base miscalling rate of >1%) were changed to 'N,' and only reads with at least half the called bases above Q20 were retained for subsequent analysis. Reads were aligned to the crRNA expression plasmid used for recombinant Cascade and TniQ–Cascade expression and purification.

TniQ–Cascade binding experiments. Binding reactions (120 μl) contained 1 μM Cascade and 5 μM StrepII-tagged TniQ, and were prepared in Cascade storage buffer and incubated at room temperature for 30 min, before being loaded into a 100 μl sample loop on an AKTApure system (GE Healthcare). Reactions were resolved by size exclusion chromatography over a Superose 6 Increase 10/300 column (GE Healthcare) in Cascade storage buffer, and proteins in each peak fraction were acetone precipitated and analysed by SDS–PAGE. Control reactions lacked either Cascade or TniQ.

Tn-seq experiments. Transposition experiments were performed as described above, except pDonor contained two point mutations in the transposon right end that introduced an MmeI restriction site (Supplementary Table 1 and Extended Data Fig. 8a, b). Colonies from triple antibiotic LB-agar plates containing IPTG (typically numbering in the range of 10^2–10^3) were resuspended in 4 ml fresh LB medium, and 0.5 ml (corresponding to around 2×10^9 cells) was used for genomic DNA (gDNA) extraction with the Wizard Genomic DNA Purification Kit (Promega). This procedure typically yielded 50 μl of 0.5–1.5 μg μl^{-1} gDNA, which is a mixture of the *E. coli* circular chromosome (4.6 Mb, copy number of 1), pDonor (3.6 kb, copy number 100+), pTnsABC (6.9 kb, copy number ~20–40), and pQCascade (8.4 kb, copy number ~20–40).

NGS libraries were prepared in parallel on 96-well plates, as follows. First, 1 μg of gDNA was digested with 4 U of MmeI (NEB) for 12 h at 37 °C in a 50 μl reaction containing 50 μM *S*-adenosyl methionine and 1× CutSmart Buffer, before heat inactivation at 65 °C for 20 min. MmeI cleaves the transposon 17–19 nucleotides outside of the terminal repeat, leaving 2-nucleotide 3′-overhangs. Reactions were cleaned up using 1.8× Mag-Bind TotalPure NGS magnetic beads (Omega) according to the manufacturer's instructions, and elutions were performed using 30 μl of 10 mM Tris-Cl, pH 7.0. MmeI-digested gDNA was ligated to a double-stranded i5 universal adaptor containing a 3′-terminal NN overhang (Supplementary Table 5) in a 20 μl ligation reaction containing 16.86 μl of MmeI-digested gDNA, 280 nM adaptor, 400 U T4 DNA ligase (NEB), and 1× T4 DNA ligase buffer. Reactions were incubated at room temperature for 30 min before being cleaned up with magnetic beads as before. To reduce the degree of pDonor contamination within our NGS libraries, since pDonor also contains the full-length transposon with an MmeI site, we took advantage of the presence of a unique HindIII restriction site just outside the transposon right end within pDonor. The entirety of the adaptor-ligated gDNA sample was thus digested with 20 Units of HindIII (NEB) in a 34.4 μl reaction for 1 h at 37 °C, before a heat inactivation step at 65 °C for 20 min. Magnetic bead-based DNA clean-up was performed as before.

Adaptor-ligated transposons were enriched in a PCR-1 step using a universal i5 adaptor primer and a transposon-specific primer containing a universal i7 adaptor as 5′ overhang. Reactions were 25 μl in volume and contained 16.75 μl of HindIII-digested gDNA, 200 μM dNTPs, 0.5 μM primers, 1× Q5 reaction buffer, and 0.5 U Q5 Hot Start High-Fidelity DNA Polymerase (NEB). Amplification was allowed to proceed for 25 cycles, with an annealing temperature of 66 °C. Reaction products were then diluted 20-fold into a second 20 μl polymerase chain reaction (PCR-2) containing indexed p5/p7 primers, and this was subjected to 10 additional thermal cycles using an annealing temperature of 65 °C. After verifying amplification for select libraries by analytical gel electrophoresis, barcoded reactions were pooled and resolved by 2% agarose gel electrophoresis, DNA was isolated by Gel Extraction Kit (Qiagen), and NGS libraries were quantified by qPCR using the NEBNext Library Quant Kit (NEB). Illumina sequencing was performed using a NextSeq mid output kit with 150-cycle single-end reads and automated demultiplexing and adaptor trimming (Illumina). Individual bases with Phred quality scores under 20 (corresponding to a base miscalling rate of >1%) were changed to 'N,' and only reads with at least half the called bases above Q20 were retained for subsequent analysis.

Tn-seq libraries with the *mariner* transposon were prepared as for the *V. cholerae* transposon, but with the following changes. Transformation reactions contained BL21(DE3) cells and a single pDonor plasmid, which encodes a KanR-containing *mariner* transposon with MmeI restriction sites on both ends, and a separate expression cassette for the Himar1C9 transposase controlled by a *lac* promoter. Transformed cells were recovered at 37 °C for 1 h before being plated on bioassay dishes containing 100 μg ml^{-1} carbenicillin, yielding on the order of 5×10^4 colonies. Cells were resuspended in 20 ml fresh LB medium after a single 16-h overnight growth, and the equivalent of 2×10^9 cells were used for genomic DNA (gDNA) extraction. NGS libraries were prepared as described above, except the restriction enzyme digestion reactions to deplete pDonor contained 20 U of BamHI and KpnI instead of HindIII.

Tn-seq data visualization and analysis. The software application Geneious Prime was used to further filter reads based on three criteria: that read lengths correspond to the expected products resulting from MmeI cleavage and adaptor ligation to genomically integrated transposons (112–113 bp for the *V. cholerae* transposon and 87–88 bp for *mariner*); that each read contain the expected transposon end sequence (allowing for one mismatch); and that the transposon-flanking sequence (trimmed to 17 bp for the *V. cholerae* transposon and 14 bp for *mariner*) map perfectly to the reference genome. Mapping to the *E. coli* BL21(DE3) genome (GenBank accession CP001509) was done using the function 'Map to reference' and the following settings: Mapper: Geneious; Fine tuning: None (fast / read mapping); Word length: 17; Maximum mismatches: 0%; Maximum Ambiguity: 1. The 'Map multiple best matches' setting was set to either 'none,' effectively excluding any reads except those that map uniquely to a single site (which we will refer to as 'uniquely mapping reads'), or to 'all,' which allows reads to map to one or multiple sites on the *E. coli* genome (which we will refer to as 'processed mapping reads'). Both sets of reads were exported as fastq files and used for downstream analysis using custom Python scripts. We note that many reads removed in this process perfectly mapped to the donor plasmid (Supplementary Table 4), revealing that HindIII or BamHI/KpnI cleavage was insufficient to completely remove contaminating pDonor-derived sequences. Coverage data for 'processed mapping reads' were exported to generate Fig. 4f.

To visualize the genome-wide integration site distribution for a given sample, 'uniquely mapping reads' were mapped to the same *E. coli* reference genome with custom Python scripts. We define the integration site for each read as the genomic coordinate (with respect to the reference genome) corresponding to the 3′ edge of the mapped read. For visualization purposes, integration events within 5-kb bins were computed and plotted as genome-wide histograms in Fig. 4c, g and Extended Data Fig. 9a, b. Plots were generated using the Matplotlib graphical library. The sequence logo in Fig. 4d was generated using WebLogo 3.

Plots comparing integration sites among biological replicates (Extended Data Fig. 8d–h) were generated by binning the genome-wide histograms based on gene annotations (*mariner*) using GenBank accession CP001509, or into 100-bp bins (*V. cholerae* transposon). For the *V. cholerae* transposon, the bins were shifted so that the 3′ end of the Cascade target site for each sample would correspond to the start of its corresponding 100-bp bin. Linear regression and bivariate analysis for the *mariner* plot (Extended Data Fig. 8d) was performed using the SciPy statistical package.

To analyse the primary integration site for each sample, custom Python scripts were used to map 'processed mapping reads' to a 600-bp genomic window surrounding the corresponding genomic target site. For reads mapping to the opposite strand as the target (that is, for the T-LR orientation, in which integration places the 'left' transposon end closest to the Cascade-binding site), the integration site was shifted 5 bp from the 3′ edge of the target site in order to account for the 5-bp target-site duplication. We define the primary integration site within this 600-bp window by the largest number of mapped reads, while we arbitrarily designate 100 bp centred at the primary integration site as the 'on-target' window. The percentage of on-target integration for each sample is calculated as the number of reads resulting from transposition within the 100-bp window, divided by the total number of reads mapping to the genome. We also determined the ratio of integration in one orientation versus the other; this parameter only utilizes on-target reads, and is calculated as the number of reads resulting from integration of the transposon 'right' end closest to the Cascade-binding site (T-RL), divided by the number reads resulting from integration of the transposon left end closest to the Cascade target site (T-LR). The distribution of integration around the primary site was plotted for both orientations for each sample, and was used to generate Fig. 4e and Extended Data Fig. 9c.

We note that these analyses are susceptible to potential biases from differential efficiencies in the ligation of 3′-terminal NN overhang adaptors, which are not taken into account in our analyses.

Statistics and reproducibility. Analytical PCRs resolved by agarose gel electrophoresis gave similar results in three independent replicates (Figs. 1d, e, i, 2a, 4a) or were analysed by gel electrophoresis once (Fig. 2e and Extended Data Figs. 1d, 2b, d and f) but verified with qPCR for three independent replicates (Fig. 2e). Sanger sequencing and next-generation sequencing of PCR amplicons was performed once (Figs. 1f, g, 3e, g, 4e and Extended Data Figs. 1e, 2a, e, 7). SDS–PAGE experiments were performed for two or more different preparations of the same protein complexes and yielded similar results (Fig. 2b and Extended Data Fig. 3b). Protein binding reactions were performed and analysed by SDS–PAGE once (Extended Data Fig. 3e). Nucleic acid extraction from purified protein preparations and urea–PAGE analysis of samples with and without RNase or DNase treatment was performed twice, with similar results (Fig. 2c and Extended Data Fig. 3d). RNA sequencing was performed once (Fig. 2d).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Next-generation sequencing data are available in the National Center for Biotechnology Information Sequence Read Archive (BioProject Accession: PRJNA546035). Custom Python scripts used for the described data analyses are available online via GitHub (https://github.com/sternberglab/Klompe_etal_2019).

58. Robert, X. & Gouet, P. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* **42**, W320–W324 (2014).
59. Biswas, A., Gagnon, J. N., Brouns, S. J. J., Fineran, P. C. & Brown, C. M. CRISPRTarget: bioinformatic prediction and analysis of crRNA targets. *RNA Biol.* **10**, 817–827 (2013).
60. Shevchenko, A., Tomas, H., Havlis, J., Olsen, J. V. & Mann, M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protocols* **1**, 2856–2860 (2006).
61. Heidrich, N., Dugar, G., Vogel, J. & Sharma, C. M. Investigating CRISPR RNA biogenesis and function using RNA-seq. *Methods Mol. Biol.* **1311**, 1–21 (2015).
62. Reiter, W. D., Palm, P. & Yeats, S. Transfer RNA genes frequently serve as integration sites for prokaryotic genetic elements. *Nucleic Acids Res.* **17**, 1907–1914 (1989).
63. Boyd, E. F., Almagro-Moreno, S. & Parent, M. A. Genomic islands are dynamic, ancient integrative elements in bacterial evolution. *Trends Microbiol.* **17**, 47–53 (2009).

Acknowledgements We thank M. I. Hogan for laboratory support, S. P. Chen and H. H. Wang for discussions, S. J. Resnick and A. Chavez for assistance with NGS experiments, R. Neme for assistance with NGS data analysis, L. F. Landweber for qPCR instrument access, the Department of Microbiology & Immunology for facilities and equipment support, the JP Sulzberger Columbia Genome Center for NGS support, and R. K. Soni and the Herbert Irving Comprehensive Cancer Center for proteomics support. Funding was provided by a generous start-up package from the Columbia University Irving Medical Center Dean's Office and the Vagelos Precision Medicine Fund.

Author contributions S.E.K. and S.H.S. conceived of and designed the project. S.E.K. performed most transposition experiments, generated NGS libraries, and analysed the data. P.L.H.V. helped with cloning and transposition experiments, and performed computational analyses. T.S.H.-H. performed biochemical experiments. S.H.S., S.E.K. and all other authors discussed the data and wrote the manuscript.

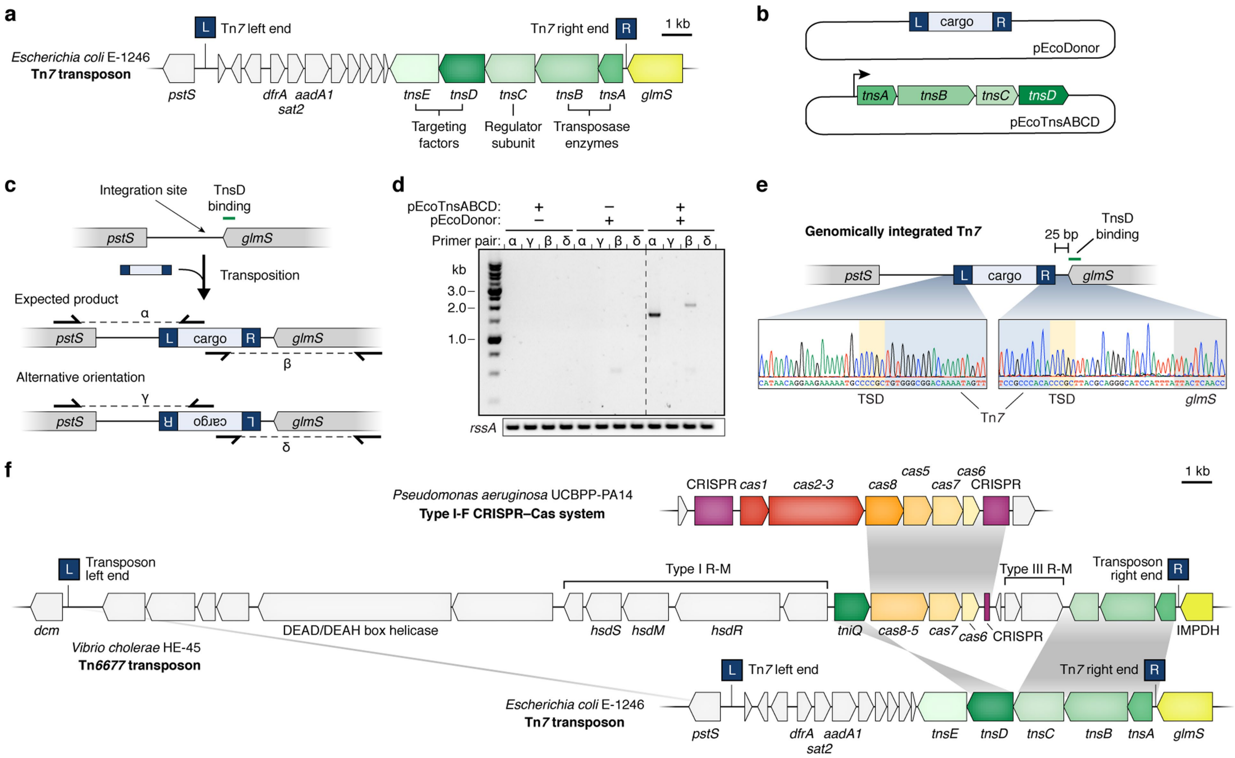
Competing interests Columbia University has filed a patent application related to this work for which S.E.K. and S.H.S. are inventors. S.E.K. and S.H.S. are inventors on other patents and patent applications related to CRISPR–Cas systems and uses thereof. S.H.S. is a co-founder and scientific advisor to Dahlia Biosciences, and an equity holder in Dahlia Biosciences and Caribou Biosciences.

Additional information

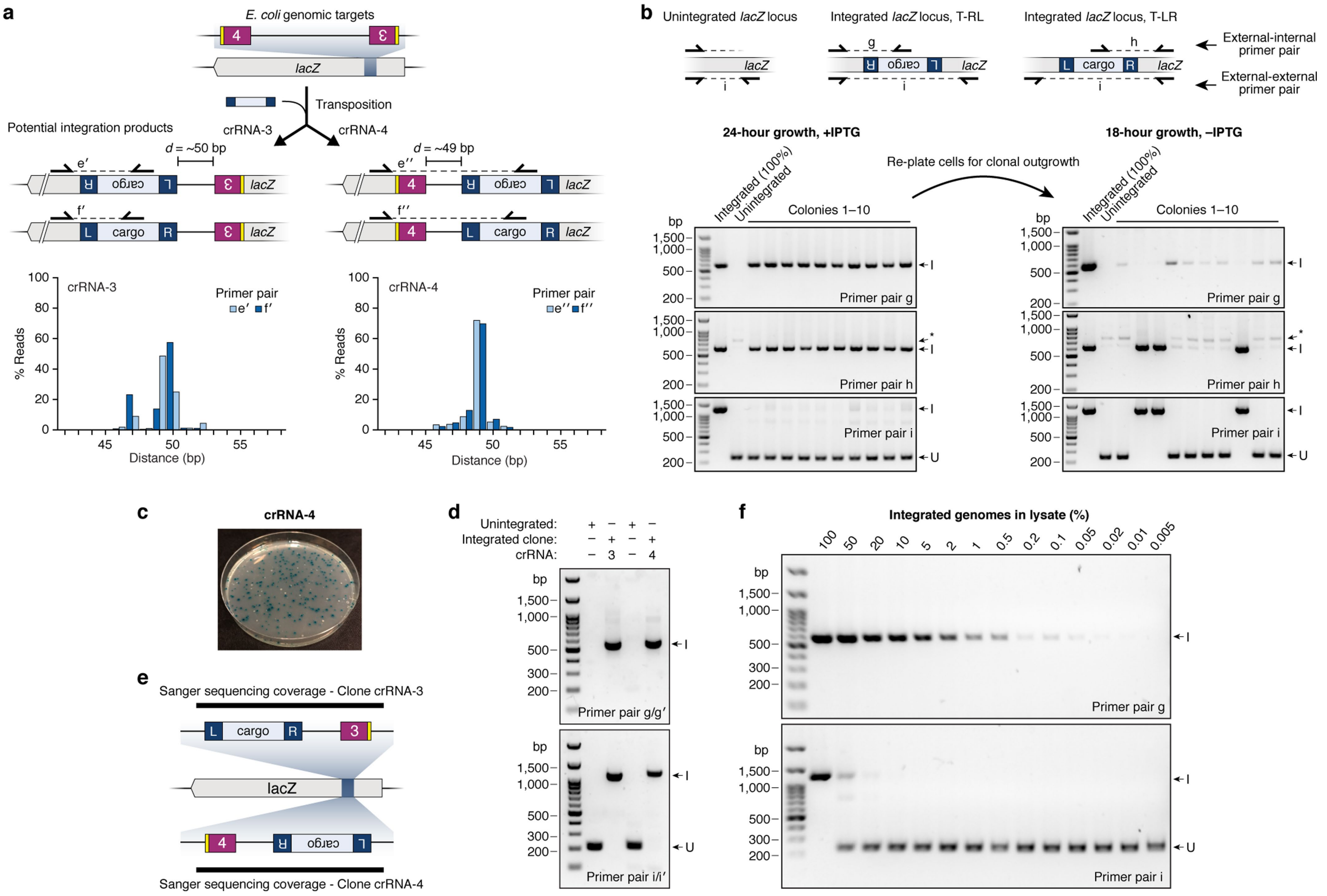
Supplementary information. is available for this paper at https://doi.org/10.1038/s41586-019-1323-z.

Correspondence and requests for materials should be addressed to S.H.S.

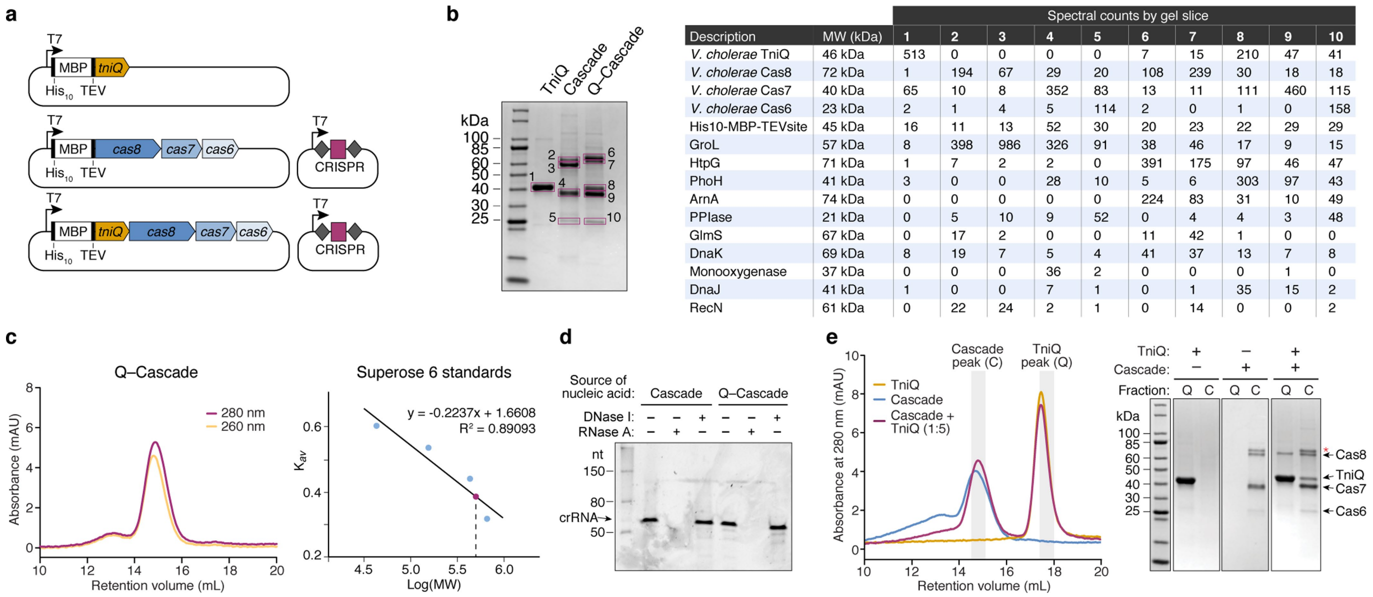
Reprints and permissions information is available at http://www.nature.com/reprints.



Extended Data Fig. 1 | Transposition of the *E. coli* Tn*7* transposon and genetic architecture of the Tn*6677* transposon from *V. cholerae*. **a**, Genomic organization of the native *E. coli* Tn*7* transposon adjacent to its known attachment site (*attTn7*) within the *glmS* gene. **b**, Expression plasmid and donor plasmid for Tn*7* transposition experiments. **c**, Genomic locus containing the conserved TnsD-binding site (*attTn7*), including the expected and alternative orientation Tn*7* transposition products and PCR primer pairs to selectively amplify them. **d**, PCR analysis of Tn*7* transposition, resolved by agarose gel electrophoresis. Amplification of *rssA* serves as a loading control; gel source data may be found in Supplementary Fig. 1. **e**, Sanger sequencing chromatograms of both upstream and downstream junctions of genomically integrated Tn*7*. **f**, Genomic organization of the native *V. cholerae* strain HE-45 Tn*6677* transposon. Genes that are conserved between Tn*6677* and the *E. coli* Tn*7* transposon, and between Tn*6677* and a canonical type I-F CRISPR–Cas system from *P. aeruginosa*[28], are highlighted. The *cas1* and *cas2-3* genes, which mediate spacer acquisition and DNA degradation during the adaptation and interference stages of adaptive immunity, respectively, are missing from CRISPR–Cas systems encoded by Tn*7*-like transposons. Similarly, the *tnsE* gene, which facilitates non-sequence-specific transposition, is absent. The *V. cholerae* HE-45 genome contains another Tn*7*-like transposon (located within GenBank accession ALED01000025.1), which lacks an encoded CRISPR–Cas system and exhibits low sequence similarity to the Tn*6677* transposon investigated in this study.

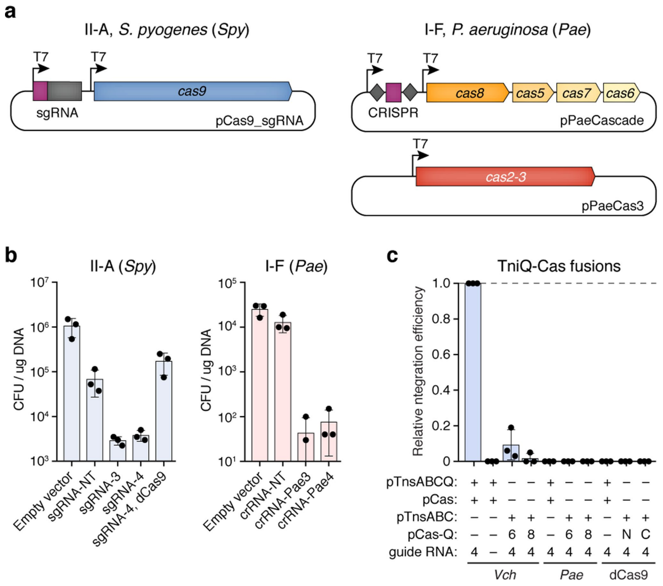


Extended Data Fig. 2 | Analysis of *E. coli* cultures and strain isolates containing *lacZ*-integrated transposons. a, Top, genomic locus targeted by crRNA-3 and crRNA-4, including both potential transposition products and the PCR primer pairs to selectively amplify them. Bottom, NGS analysis of the distance between the Cascade target site and transposon insertion site for crRNA-3 (left) and crRNA-4 (right), determined with two alternative primer pairs. **b**, Top, schematic of the *lacZ* locus with or without integrated transposon after transposition experiments with crRNA-4. T-LR and T-RL denote transposition products in which the transposon left end and right end are proximal to the target site, respectively. Primer pairs g and h (external–internal) selectively amplify the integrated locus, whereas primer pair i (external–external) amplifies both unintegrated and integrated loci. Bottom, PCR analysis of 10 colonies after 24-h growth on +IPTG plates (left) indicates that all colonies contain integration events in both orientations (primer pairs g and h), but with efficiencies sufficiently low that the unintegrated product predominates after amplification with primer pair i. After resuspending cells, allowing for an additional 18 h of clonal growth on −IPTG plates, and performing the same PCR analysis on 10 colonies (right), 3 out of 10 colonies now exhibit clonal integration in the T-LR orientation (compare primer pairs h and i). The remaining colonies show low-level integration in both orientations, which presumably occurred during the additional 18-h growth owing to leaky expression. These analyses indicate that colonies are genetically heterogeneous after growth on +IPTG plates, and that RNA-guided DNA integration only occurs in a proportion of cells within growing colonies. I, integrated product; U, unintegrated product. Asterisk denotes mispriming product also present in the negative (unintegrated) control. **c**, Photograph of LB-agar plate used for blue–white colony screening. Cells from IPTG-containing plates were replated on X-gal-containing plates, and white colonies expected to contain *lacZ*-inactivating transposon insertions were selected for further characterization. **d**, PCR analysis of *E. coli* strains identified by blue–white colony screening that contain clonally integrated transposons, as in **b**. **e**, Schematic of Sanger sequencing coverage across the *lacZ* locus for strains shown in **d**. **f**, PCR analysis of transposition experiment with crRNA-4 after serially diluting lysate from a clonally integrated strain with lysate from a control strain to simulate variable integration efficiencies, as in **b**. These experiments demonstrate that transposition products can be reliably detected by PCR with an external–internal primer pair at efficiencies above 0.5%, but that PCR bias leads to preferential amplification of the unintegrated product using the external-external primer pair at any efficiency substantially below 100%. For gel source data, see Supplementary Fig. 1.

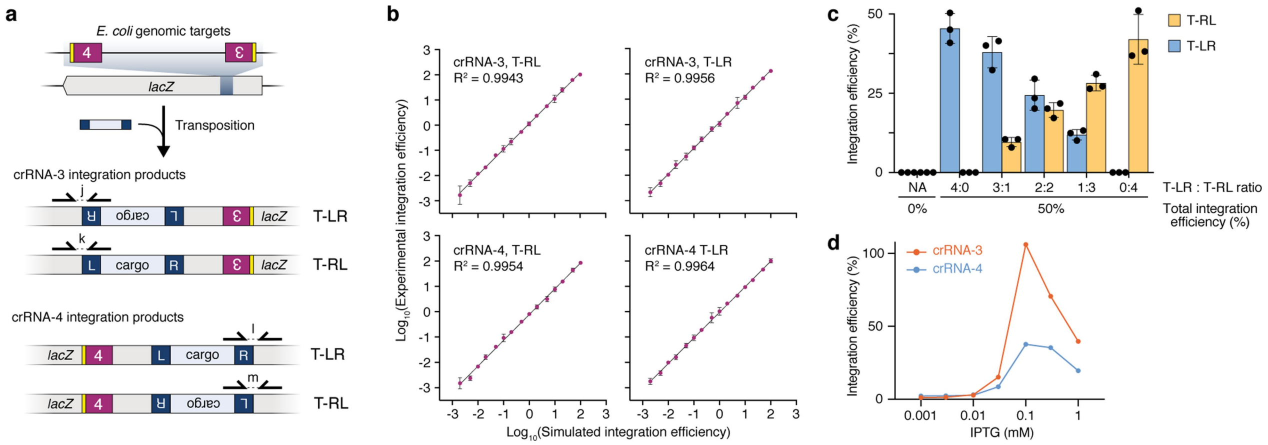


		Spectral counts by gel slice									
Description	MW (kDa)	1	2	3	4	5	6	7	8	9	10
V. cholerae TniQ	46 kDa	513	0	0	0	0	7	15	210	47	41
V. cholerae Cas8	72 kDa	1	194	67	29	20	108	239	30	18	18
V. cholerae Cas7	40 kDa	65	10	8	352	83	13	11	111	460	115
V. cholerae Cas6	23 kDa	2	1	4	5	114	2	0	1	0	158
His10-MBP-TEVsite	45 kDa	16	11	13	52	30	20	23	22	29	29
GroL	57 kDa	8	398	986	326	91	38	46	17	9	15
HtpG	71 kDa	1	7	2	2	0	391	175	97	46	47
PhoH	41 kDa	3	0	0	28	10	5	6	303	97	43
ArnA	74 kDa	0	0	0	0	0	224	83	31	10	49
PPIase	21 kDa	0	5	10	9	52	0	4	4	3	48
GlmS	67 kDa	0	17	2	0	0	11	42	1	0	0
DnaK	69 kDa	8	19	7	5	4	41	37	13	7	8
Monooxygenase	37 kDa	0	0	0	36	2	0	0	0	1	0
DnaJ	41 kDa	1	0	0	7	1	0	1	35	15	2
RecN	61 kDa	0	22	24	2	1	0	14	0	0	2

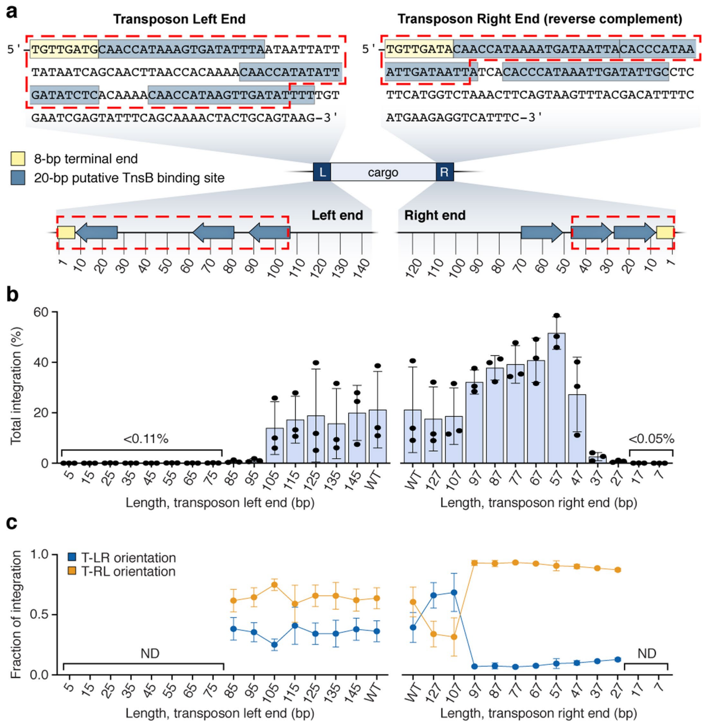
Extended Data Fig. 3 | Analysis of *V. cholerae* Cascade and TniQ–Cascade complexes. a, Expression vectors for recombinant protein or ribonucleoprotein complex purification. **b**, Left, SDS–PAGE analysis of purified TniQ, Cascade and TniQ–Cascade complexes, highlighting protein bands excised for in-gel trypsin digestion and mass spectrometry analysis. Right, table listing *E. coli* and recombinant proteins identified from these data, and spectral counts of their associated peptides. Note that Cascade and TniQ–Cascade samples used for this analysis are distinct from the samples presented in Fig. 2. **c**, Size-exclusion chromatogram of the TniQ–Cascade co-complex on a Superose 6 10/300 column (left), and a calibration curve generated using protein standards (right). The measured retention time of TniQ–Cascade (maroon) is consistent with a complex having a molecular mass of approximately 440 kDa. **d**, RNase A and DNase I sensitivity of nucleic acids that co-purified with Cascade and TniQ–Cascade, resolved by denaturing urea–PAGE. **e**, TniQ, Cascade and a Cascade + TniQ binding reaction were resolved by size-exclusion chromatography (left), and indicated fractions were analysed by SDS–PAGE (right). Asterisk denotes an HtpG contaminant. For gel source data, see Supplementary Fig. 1.



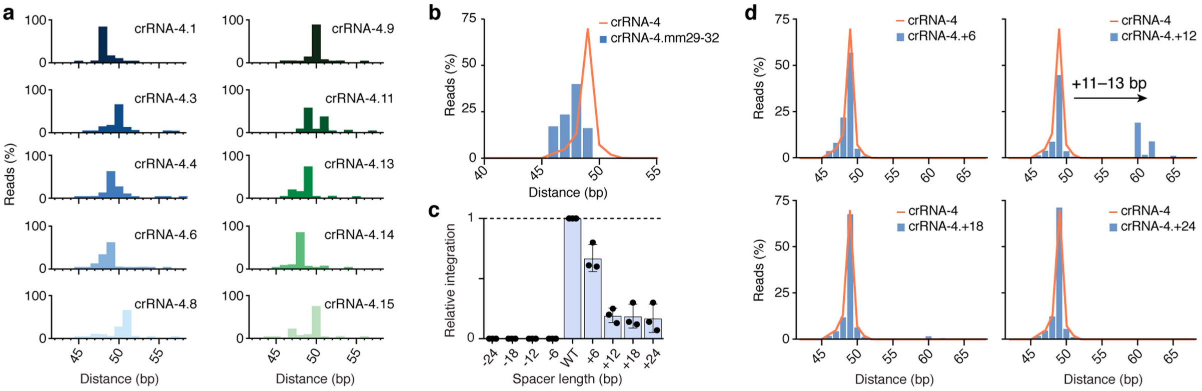
Extended Data Fig. 4 | Control experiments demonstrating efficient DNA targeting with Cas9 and *P. aeruginosa* Cascade. a, Plasmid expression system for *S. pyogenes* (*Spy*) Cas9-sgRNA (type II-A, left) and *P. aeruginosa* Cascade (*Pae*Cascade) and Cas2-3 (type I-F, right). The Cas2-3 expression plasmid was omitted from experiments described in Fig. 2e. **b**, Cell killing experiments using *S. pyogenes* Cas9-sgRNA (left) or *Pae*Cascade and Cas2-3 (right), monitored by determining colony-forming units (CFU) after plasmid transformation. Complexes were programmed with guide RNAs that target the same genomic *lacZ* sites as with *V. cholerae* crRNA-3 and crRNA-4, such that efficient DNA targeting and degradation results in lethality and thus a drop in transformation efficiency. **c**, qPCR-based quantification of transposition efficiency from experiments using the *V. cholerae* transposon donor and TnsA-TnsB-TnsC, together with DNA targeting components comprising *V. cholerae* Cascade (*Vch*), *P. aeruginosa* Cascade (*Pae*) or *S. pyogenes* dCas9–RNA (dCas9). TniQ was expressed either on its own from pTnsABCQ or as a fusion to the targeting complex (pCas-Q) at the Cas6 C terminus (6), Cas8 N terminus (8), or dCas9 N or C terminus. The same sample lysates as in Fig. 2e were used. Data in **b** and **c** are shown as mean ± s.d. for $n = 3$ biologically independent samples.



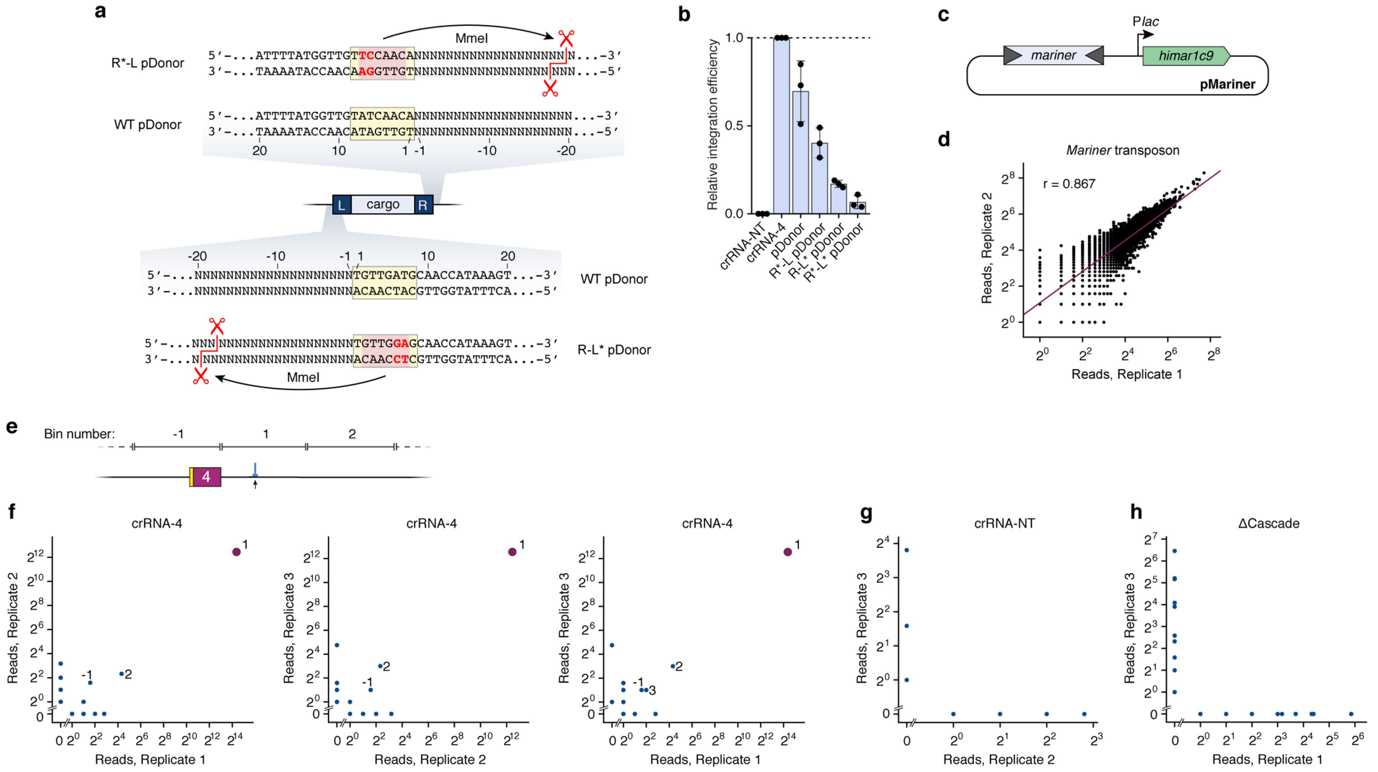
Extended Data Fig. 5 | qPCR-based quantification of RNA-guided DNA integration efficiencies. a, Potential *lacZ* transposition products in either orientation for both crRNA-3 and crRNA-4, and qPCR primer pairs to selectively amplify them. **b**, Comparison of simulated integration efficiencies for T-LR and T-RL orientations, generated by mixing clonally integrated and unintegrated lysates in known ratios, versus experimentally determined integration efficiencies measured by qPCR. **c**, Comparison of simulated mixtures of bidirectional integration efficiencies for crRNA-4, generated by mixing clonally integrated and unintegrated lysates in known ratios, versus experimentally determined integration efficiencies measured by qPCR. **d**, RNA-guided DNA integration efficiency as a function of IPTG concentration for crRNA-3 and crRNA-4, measured by qPCR. Data in **b** and **c** are shown as mean ± s.d. for $n = 3$ biologically independent samples.



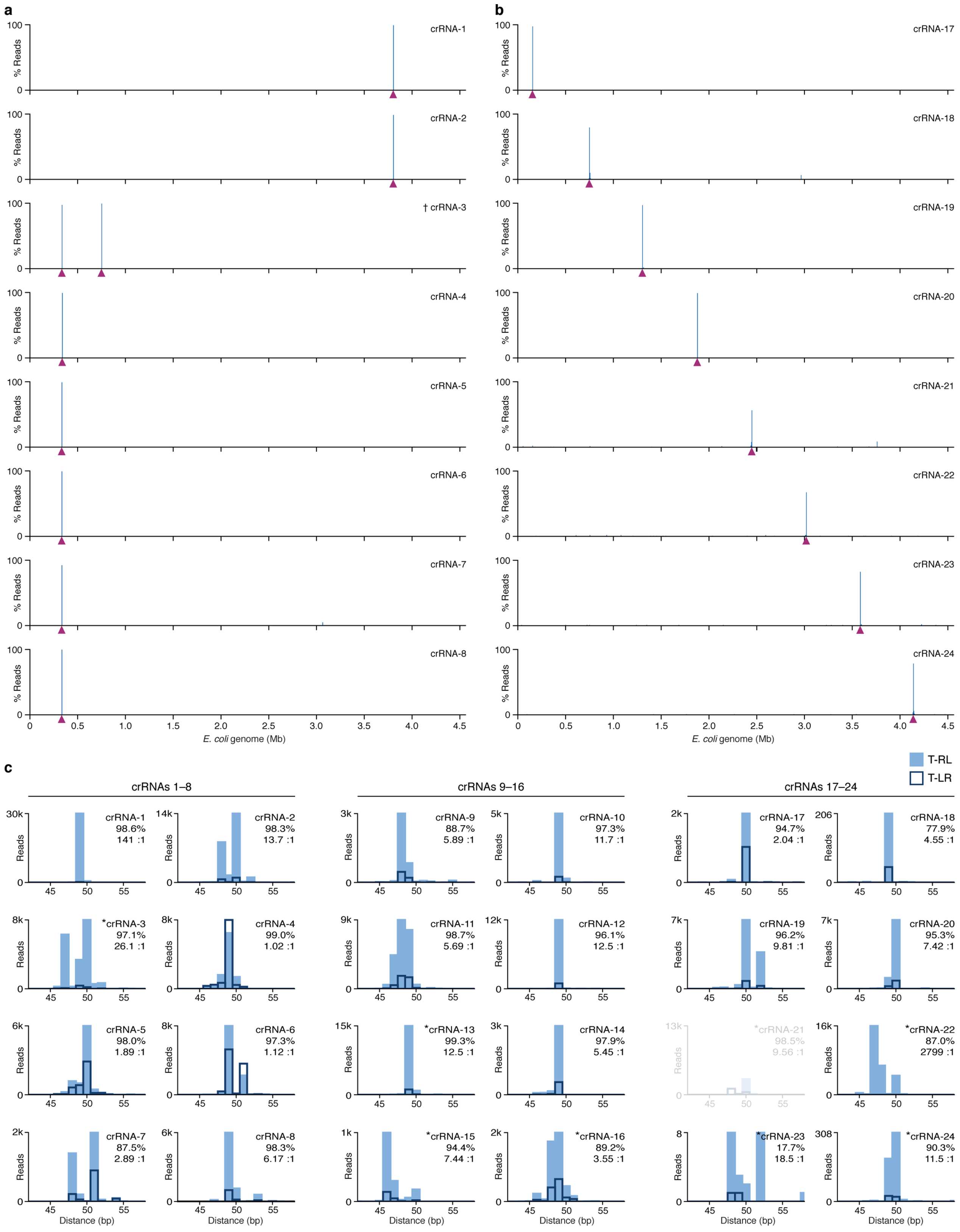
Extended Data Fig. 6 | Influence of transposon end sequences on RNA-guided DNA integration. a, Sequence (top) and schematic (bottom) of *V. cholerae* Tn*6677* left- and right-end sequences. The putative TnsB-binding sites (blue) were determined based on sequence similarity to the TnsBbinding sites previously described[14]. The 8-bp terminal ends are shown in yellow, and the empirically determined minimum end sequences required for transposition are denoted by red dashed boxes. **b**, Integration efficiency with crRNA-4 as a function of transposon end length, as determined by qPCR. **c**, The relative fraction of both integration orientations as a function of transposon end length, determined by qPCR. ND, not determined. Data in **b** and **c** are shown as mean ± s.d. for $n = 3$ biologically independent samples.



Extended Data Fig. 7 | Analysis of RNA-guided DNA integration for PAM-tiled crRNAs and extended spacer length crRNAs.
a, Integration site distribution for all crRNAs described in Fig. 3d, e having a normalized transposition efficiency more than 20%, determined by NGS. **b**, Integration site distribution for a crRNA containing mismatches at positions 29–32, compared with the distribution with crRNA-4, determined by NGS. **c**, The crRNA-4 spacer length was shortened or lengthened by 6-nucleotide increments, and the resulting integration efficiencies were determined by qPCR. Data are normalized to crRNA-4 and are shown as mean ± s.d. for $n = 3$ biologically independent samples. **d**, Integration site distribution for extended length crRNAs compared with the distribution with crRNA-4, determined by NGS.

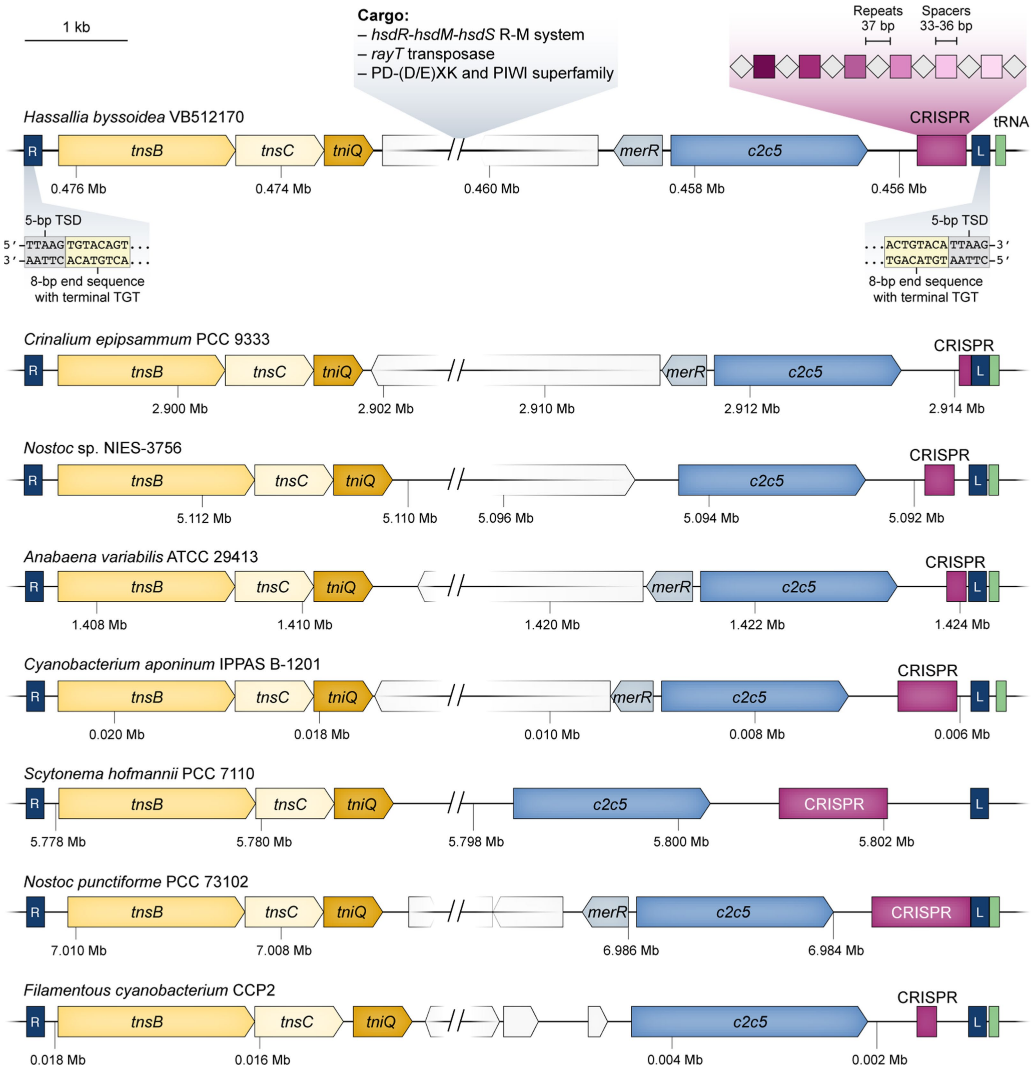


Extended Data Fig. 8 | Development and analysis of Tn-seq. **a**, Schematic of the *V. cholerae* transposon end sequences. The 8-bp terminal sequence of the transposon is boxed and highlighted in light yellow. Mutations generated to introduce MmeI recognition sites are shown in red letters, and the resulting recognition site is highlighted in red. Cleavage by MmeI occurs 17–19 bp away from the transposon end, generating a 2-bp overhang. **b**, Comparison of integration efficiencies for the wild-type and MmeI-containing transposon donors, determined by qPCR. Labels on the *x* axis denote which plasmid was transformed last; we reproducibly observed higher integration efficiencies when pQCascade was transformed last (crRNA-4) than when pDonor was transformed last. The transposon containing an MmeI site in the transposon 'right' end (R∗-L pDonor) was used for all Tn-seq experiments. Data are mean ± s.d. for $n = 3$ biologically independent samples. **c**, Plasmid expression system for Himar1C9 and the *mariner* transposon. **d**, Scatter plot showing correlation between two biological replicates of Tn-seq experiments with the *mariner* transposon. Reads were binned by *E. coli* gene annotations, and a linear regression fit and Pearson linear correlation coefficient (*r*) are shown. **e**, Schematic of 100-bp binning approach used for Tn-seq analysis of transposition experiments with the *V. cholerae* transposon, in which bin 1 is defined as the first 100 bp immediately downstream (PAM-distal) of the Cascade target site. **f**, Scatter plots showing correlation between biological replicates of Tn-seq experiments with the *V. cholerae* transposon programmed with crRNA-4. All highly sampled reads fall within bin 1, but we also observed low-level but reproducible, long-range integration into 100-bp bins just upstream and downstream of the primary integration site (bins −1, 2 and 3). **g**, Scatter plot showing correlation between biological replicates of Tn-seq experiments with the *V. cholerae* transposon programmed with a non-targeting crRNA (crRNA-NT). **h**, Scatter plot showing correlation between biological replicates of Tn-seq experiments with the *V. cholerae* transposon expressing TnsA-TnsB-TnsC-TniQ but not Cascade. For **f**–**h**, bins are only plotted when they contain at least one read in either dataset.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Tn-seq data for additional crRNAs tested. **a**, **b**, Genome-wide distribution of genome-mapping Tn-seq reads from transposition experiments with the *V. cholerae* transposon programmed with crRNAs 1–8 (**a**) and crRNAs 17–24 (**b**). The location of each target site is denoted by a maroon triangle. Dagger symbol indicates that the *lacZ* target site for crRNA-3 is duplicated within the λ DE3 prophage, as is the transposon integration site; Tn-seq reads for this dataset were mapped to both genomic loci for visualization purposes only, although we are unable to determine from which locus they derive. **c**, Analysis of integration site distributions for crRNAs 1–24 determined from the Tn-seq data; the distance between the Cascade target site and transposon insertion site is shown. Data for both integration orientations are superimposed, with filled blue bars representing the T-RL orientation and the dark outlines representing the T-LR orientation. Values in the top-right corner of each graph give the on-target specificity (%), calculated as the percentage of reads resulting from integration within 100 bp of the primary integration site, as compared with the total number of reads aligning to the genome; and the orientation bias (*X*:*Y*), calculated as the ratio of reads for the T-RL orientation to reads for the T-LR orientation. Most crRNAs favour integration in the T-RL orientation 49–50 bp downstream of the Cascade target site. crRNA-21 is greyed out because the expected primary integration site is present in a repetitive stretch of DNA that does not allow us to map the reads confidently. Asterisks denote samples for which more than 1% of the genome-mapping reads could not be uniquely mapped.



Extended Data Fig. 10 | Bacterial transposons also contain type V-U5 CRISPR–Cas systems encoding C2c5. Representative genomic loci from various bacterial species containing identifiable transposon left and right ends (blue boxes, L and R), genes with homology to *tnsB-tnsC-tniQ* (shades of yellow), CRISPR arrays (maroon), and the CRISPR-associated gene *c2c5* (blue). The example from *Hassallia byssoidea* (top) highlights the target-site duplication and terminal repeats, as well as genes found within the cargo portion of the transposon. As with the type I CRISPR–Cas system-containing Tn*7*-like transposons, type V CRISPR–Cas system-containing transposons appear to preferentially contain genes associated with innate immune system functions, such as restriction-modification systems. *c2c5* genes are frequently flanked by the predicted transcriptional regulator, *merR* (light blue), and the C2c5-containing transposons appear to usually fall just upstream of tRNA genes (green), a phenomenon that has also been observed for other prokaryotic integrative elements[62,63]. Analysis of 50 spacers from the 8 CRISPR arrays shown with CRISPRTarget[59] revealed 6 spacers with imperfectly matching targets (average of 6 mismatches), none of which mapped to bacteriophages, plasmids, or to the same bacterial genome containing the transposon itself. Whether C2c5 also mediates RNA-guided DNA integration awaits future experimentation.

natureresearch

Corresponding author(s): Samuel H. Sternberg

Last updated by author(s): May 27, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	□	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
☒	□	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	□	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Next-generation sequencing data utilized the Illumina platform (Basespace), including automated de-multiplexing and adapter trimming.
Data analysis	Next-generation sequencing data were analyzed using either Geneious Prime (version 2019.0.4) and/or custom Python scripts (available on GitHub). Mass spectrometry data were analyzed using Proteome Discoverer 2.2 and Scaffold (Proteome Software). Multiple sequence alignments were made using Clustal Omega and visualized with ESPript 3.0. Analysis of spacers was performed using CRISPRTarget. Sequence logos were generated using WebLogo 3.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Next-generation sequencing data will become available in the National Center for Biotechnology Information Sequence Read Archive. The data sets generated and/or analyzed during the current study, as well as custom scripts used for the described data analyses, are available from the corresponding author upon reasonable request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences　□ Behavioural & social sciences　□ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes are reported in the figure legends. Generally experiments were done individually for three biological replicates.
Data exclusions	No data were excluded.
Replication	All data could be reproduced, and most experiments and analyses presented were the result of three independent biological replicates.
Randomization	Almost all analyses were performed on the entire heterogeneous population that was grown on solid media to prevent growth biases, therefore randomization is not applicable.
Blinding	Samples were prepared unblinded but in parallel transformation/incubation/harvesting. Mapping of reads from Tn-seq was done without prior knowledge of which site was targeted and was only introduced later to analyze on-target specificity.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology
☒	□ Animals and other organisms
☒	□ Human research participants
☒	□ Clinical data

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging

Appendix B

Halpin-Healy *et al*, 2020

Structural basis of DNA targeting by a transposon-encoded CRISPR–Cas system

https://doi.org/10.1038/s41586-019-1849-0

Received: 16 July 2019

Accepted: 25 October 2019

Published online: 18 December 2019

Tyler S. Halpin-Healy[1], Sanne E. Klompe[1], Samuel H. Sternberg[1]* & Israel S. Fernández[1]*

Bacteria use adaptive immune systems encoded by CRISPR and Cas genes to maintain genomic integrity when challenged by pathogens and mobile genetic elements[1–3]. Type I CRISPR–Cas systems typically target foreign DNA for degradation via joint action of the ribonucleoprotein complex Cascade and the helicase–nuclease Cas3[4,5], but nuclease-deficient type I systems lacking Cas3 have been repurposed for RNA-guided transposition by bacterial Tn*7*-like transposons[6,7]. How CRISPR- and transposon-associated machineries collaborate during DNA targeting and insertion remains unknown. Here we describe structures of a TniQ–Cascade complex encoded by the *Vibrio cholerae* Tn*6677* transposon using cryo-electron microscopy, revealing the mechanistic basis of this functional coupling. The cryo-electron microscopy maps enabled de novo modelling and refinement of the transposition protein TniQ, which binds to the Cascade complex as a dimer in a head-to-tail configuration, at the interface formed by Cas6 and Cas7 near the 3′ end of the CRISPR RNA (crRNA). The natural Cas8–Cas5 fusion protein binds the 5′ crRNA handle and contacts the TniQ dimer via a flexible insertion domain. A target DNA-bound structure reveals critical interactions necessary for protospacer-adjacent motif recognition and R-loop formation. This work lays the foundation for a structural understanding of how DNA targeting by TniQ–Cascade leads to downstream recruitment of additional transposase proteins, and will guide protein engineering efforts to leverage this system for programmable DNA insertions in genome-engineering applications.

We previously demonstrated that a transposon derived from *V. cholerae* Tn*6677* undergoes programmable transposition in *Escherichia coli* directed by a crRNA, and that this activity requires four transposon- and three CRISPR-associated genes in addition to a CRISPR array[7] (Fig. 1a). Whereas TnsA, TnsB and TnsC exhibit functions that are consistent with their homologues from the related and well-studied cut-and-paste DNA transposon *E. coli* Tn*7*[8], we showed that TniQ, a homologue of *E. coli* TnsD, forms a co-complex with the Cascade ribonucleoprotein complex encoded by the type I-F variant CRISPR–Cas system. This finding suggested an alternative role for TniQ, compared with the role of *E. coli* TnsD in identifying target sites during Tn*7* transposition. We proposed that RNA-guided DNA targeting by Cascade could deliver TniQ to DNA in a manner compatible with downstream transpososome formation, and that TniQ might interact with Cascade near the 3′ end of the crRNA, consistent with RNA-guided DNA insertion occurring around 49 bp downstream of the protospacer-adjacent motif (PAM)-distal edge of the target site. To determine this unambiguously, we purified the *V. cholerae* TniQ–Cascade complex loaded with a native crRNA and determined its structure by cryo-electron microscopy (cryo-EM) (Supplementary Table 1).

Cryo-EM structure of TniQ–Cascade complex

The overall complex adopts a helical architecture with protuberances at both ends (Fig. 1, Extended Data Figs. 1, 2). The global architecture is similar to previously determined structures of Cascade from I-E and I-F systems[9–12] (Extended Data Fig. 3), with the exception of a large mass of additional density attributable to TniQ (see below). Maximum likelihood classification methods implemented in Relion3[13] enabled us to identify marked dynamics in the entire complex, which appears to ‘breathe’, widening and narrowing the distance between the two protuberances (Extended Data Fig. 1d, Supplementary Video 1). The large subunit encoded by a natural Cas8–Cas5 fusion protein (hereafter referred to simply as Cas8) forms one protuberance and recognizes the 5′ end of the crRNA via base- and backbone-specific contacts (Extended Data Figs. 4, 5a–c, 6a), similar to the canonical roles of Cas8 and Cas5 (Extended Data Fig. 3). Cas8 contains two primary subdomains formed mainly by α-helices and a third domain of approximately 100 residues (residues 277 to 385) that is predicted to form three α-helices but could not be built in our maps owing to its intrinsic flexibility (Fig. 1c). However, low-pass-filtered maps revealed that this flexible domain connects with the TniQ protuberance at the opposite end of the crescent-shaped complex (Extended Data Fig. 2e). Additionally, there seemed to be a loose coupling between the Cas8 flexible domain and overall breathing of the complex, as stronger density for that domain could be observed in the closed state (Extended Data Fig. 1d, Supplementary Video 1).

Six Cas7 subunits protect much of the crRNA by forming a helical filament along its length (Fig. 1b, d), similar to other type I Cascade complexes[9–12] (Extended Data Fig. 3). A ‘finger’ motif in Cas7 clamps the crRNA at regular intervals, causing every sixth nucleotide (nt) of the 32-nt spacer to flip out while leaving the flanking nucleotides

[1]Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY, USA. *e-mail: shsternberg@gmail.com; isf2106@cumc.columbia.edu

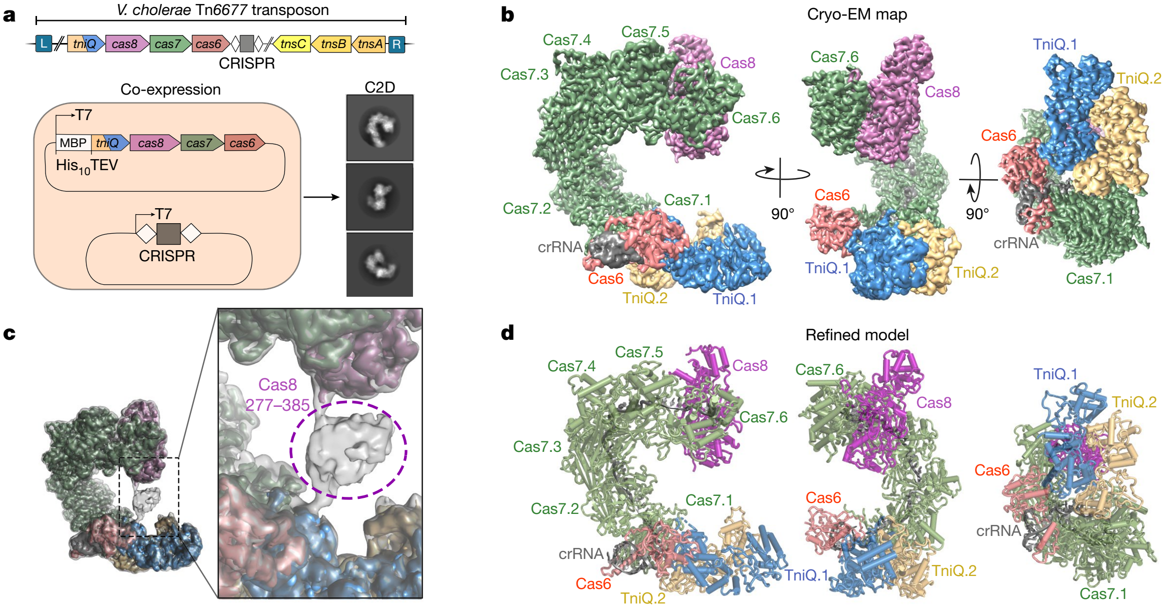


Fig. 1 | Overall architecture of the *V. cholerae* TniQ–Cascade complex. **a**, Genetic architecture of the Tn*6677* transposon (top), and plasmid constructs used to express and purify the TniQ–Cascade complex. Right, selected cryo-EM reference-free two-dimensional class averages in multiple orientations. **b**, Orthogonal views of the cryo-EM map of the TniQ–Cascade complex, showing Cas8 (purple), six Cas7 monomers (green), Cas6 (salmon), crRNA (grey) and TniQ monomers (blue, yellow). The complex adopts a helical architecture with protuberances at both ends. **c**, A flexible domain in Cas8 comprising residues 277–385 (grey) could only be visualized in low-pass-filtered maps. The unsharpened map is shown as semi-transparent, grey map overlaid on the post-processed map segmented and coloured as in **a**. **d**, Refined model for the TniQ–Cascade complex derived from the cryo-EM maps shown in **b**.

available for DNA recognition (Extended Data Figs. 4f, 6). These bases are pre-ordered in short helical segments, with a conserved phenylalanine stacking below the first base of every segment. Cas7.1, the monomer furthest away from Cas8, interacts with Cas6 (also known as Csy4), which is the RNase responsible for processing of the precursor RNA transcript derived from the CRISPR locus. The Cas6–Cas7.1 interaction is mediated by a β-sheet formed by the contribution of a β-strand from Cas6 and the two β-strands that form the finger of Cas7.1 (Extended Data Fig. 5f). Cas6 also forms extensive interactions with the conserved stem-loop in the repeat-derived 3′ crRNA handle (Fig. 1, Extended Data Fig. 5d, e), with an arginine-rich α-helix (residues 110 to 128) docked in the major groove, positioning multiple basic residues within interaction distance of the negatively charged RNA backbone.

The interaction established between Cas6 and Cas7.1 forms a continuous surface on which TniQ is docked, forming the other protuberance of the crescent. The intrinsic flexibility of the complex resulted in lower local resolutions in this area of the maps, which we overcame using local alignments masking the area comprising TniQ, Cas6, Cas7.1 and the crRNA handle (Extended Data Fig. 7). The enhanced maps enabled de novo modelling and refinement of TniQ, for which no previous structure or homology model has been reported, to our knowledge (Fig. 2). Notably, TniQ binds to Cascade as a dimer with head-to-tail configuration (Fig. 2), a surprising result given the expectation that *E. coli* TnsD functions as a monomer during Tn*7* transposition[14].

TniQ binds to Cascade as a dimer

TniQ is composed of two domains: an N-terminal domain of approximately 100 residues formed by three short α-helices and a second, larger domain of approximately 300 residues with a signature sequence for the TniQ family. A DALI search[15] using the refined TniQ model as a probe yielded marked structural similarity of the N-terminal domain to proteins containing helix–turn–helix (HTH) domains. This domain is often involved in nucleic acid recognition; however, there are examples where it has been re-purposed for protein–protein interactions[16]. The remaining C-terminal TniQ domain is formed by ten α-helices of variable length and is predicted to contain two tandem zinc finger motifs, although this region was poorly defined in the maps (Fig. 2). Overall, the double domain composition of TniQ results in an elongated structure, bent at the junction of the HTH and the TniQ domain (Fig. 2). The HTH domain of one monomer engages the TniQ domain of the other monomer via interactions between α-helix 3 (H3) and α-helix 12 (H12), respectively, in a tight protein–protein interaction (Fig. 2c). This reciprocal interaction is complemented by multiple interactions established between the TniQ domains from both monomers (up to 45 non-covalent interactions as reported by PISA[17]).

Tethering of the TniQ dimer to Cascade is accomplished by specific interactions established with both Cas6 and Cas7.1 (Fig. 3). One monomer of TniQ interacts with Cas6 via its C-terminal TniQ domain, whereas the other TniQ monomer contacts Cas7.1 through its N-terminal HTH domain (Figs. 2b, 3). The loop connecting α-helices H7 and H8 of the TniQ domain of the first TniQ monomer is inserted in a hydrophobic cavity formed at the interface of two α-helices of Cas6 (Fig. 3b, d). The TniQ histidine residue 265 is involved in rearranging the hydrophobic loop connecting H7 and H8 (Fig. 3d), which is inserted in the hydrophobic pocket of Cas6 formed by residues L20, Y74, M78, Y83 and F84. The buried surface in the Cas6–TniQ.1 interaction interface has an area of 420 Å^2. The HTH domain of the other TniQ monomer interacts with Cas7.1 through a network of interactions established mainly by α-helix H2 and the linker connecting H2 and H3, burying a surface area of 595 Å^2 (Fig. 3c, e). Thus, the HTH domain and the TniQ domain exert dual roles to drive TniQ dimerization and dock onto Cascade. The aggregate buried surface area for the TniQ–Cascade interaction is 1,015 Å^2, significantly smaller than other Cascade–effector interactions such as with the nuclease Cas3, in which 2,433 Å^2 is buried[18]. This difference is not surprising given the flexibility observed for the TniQ dimer in its association with Cascade (Supplementary Video 1).

Structure of the DNA-bound TniQ–Cascade complex

To investigate the structural determinants of DNA recognition by the TniQ–Cascade complex, we determined the structure of the complex

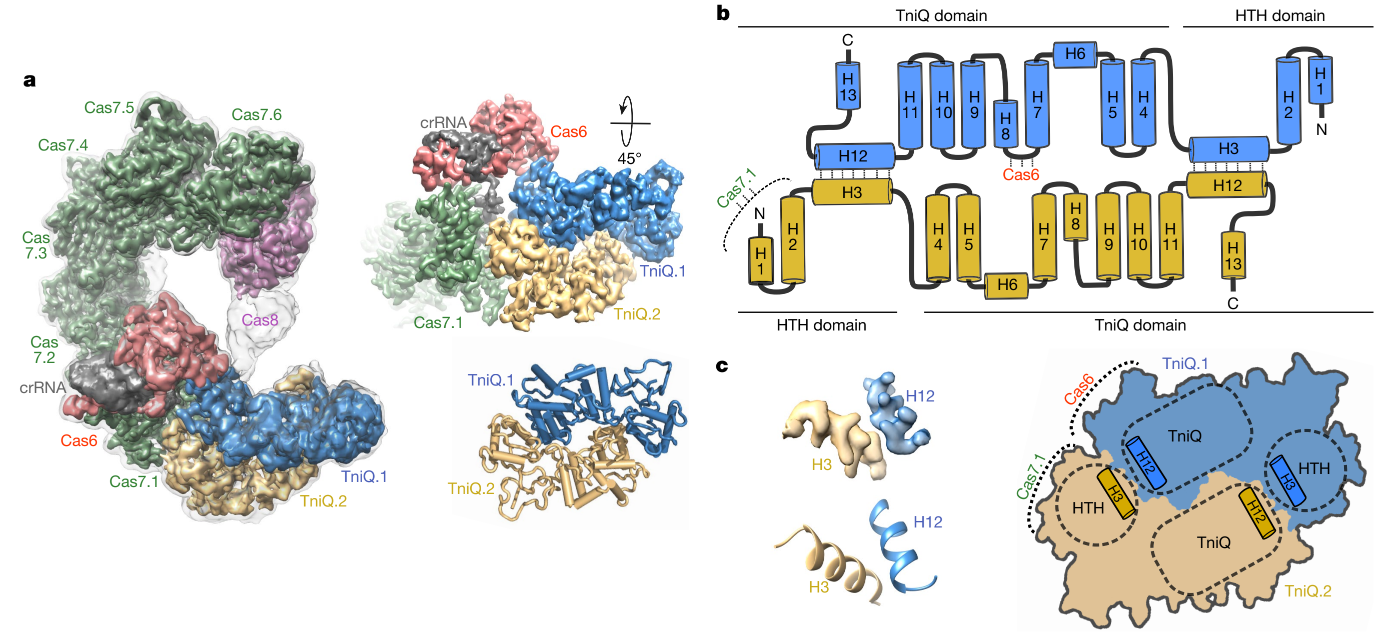


Fig. 2 | TniQ binds Cascade in a dimeric, head-to-tail configuration. a, Left, overall view of the TniQ–Cascade cryo-EM unsharpened map (grey) overlaid on the post-processed map segmented and coloured as in Fig. 1. Right, cryo-EM map (top) and refined model (bottom) of the TniQ dimer. The two monomers interact with each other in a head-to-tail configuration and are anchored to Cascade by Cas6 and Cas7.1. **b**, Secondary structure diagram of the TniQ dimer: thirteen α-helices are organized into an N-terminal HTH domain and a C-terminal TniQ domain. Dimer interactions between H3 and H12 are indicated, as are interaction sites with Cas6 and Cas7.1. **c**, Cryo-EM density for the H3–H12 interaction shows clear side-chain features (top), allowing accurate modelling of the interaction (bottom). **d**, Schematic of the dimer interaction, showing the important dimerization interface between the HTH and TniQ domain.

bound to a double-stranded DNA (dsDNA) substrate containing the 32-bp target sequence, 5′-CC-3′ PAM, and 20 bp of flanking dsDNA on both ends (Fig. 4, Extended Data Fig. 8). Density for 28 nucleotides of the target strand and 8 nucleotides of the non-target strand could be confidently assigned in the reconstructed maps (Fig. 4c). As with previous I-F Cascade structures, Cas8 recognizes the double-stranded PAM within the minor groove[11] (Extended Data Fig. 9), and an arginine residue (R246) establishes a stacking interaction with a guanine nucleotide on the target strand, which acts as a wedge to separate the double-stranded PAM from the neighbouring unwound DNA where base-pairing with the crRNA begins (Fig. 4c).

Twenty-two nucleotides of the target strand within the 32-bp target showed clear density, but surprisingly, the terminal nine nucleotides were not ordered. The target-strand base pairs with the spacer region of the crRNA in short, discontinuous helical segments, as observed previously for I-E and I-F DNA-bound Cascade complexes[11,12], with every sixth base flipped out of the heteroduplex by the insertion of a Cas7 finger (Extended Data Fig. 6b). The observed 22-bp heteroduplex is stabilized by the four Cas7 monomers proximal to the PAM (Cas7.6–Cas7.3), but even after local masked refinements, no density could be observed for any target strand nucleotides that would base-pair with the 3′ end of the crRNA spacer bound by Cas7.2 and Cas7.1. These two Cas7 monomers are proximal to Cas6 and in the region previously described to exhibit dynamics owing to the interaction of the Cas8 flexible domain with the inner face of the TniQ dimer. In addition, the disordered nucleotides also correspond to positions 25–28 of the target site where RNA–DNA mismatches

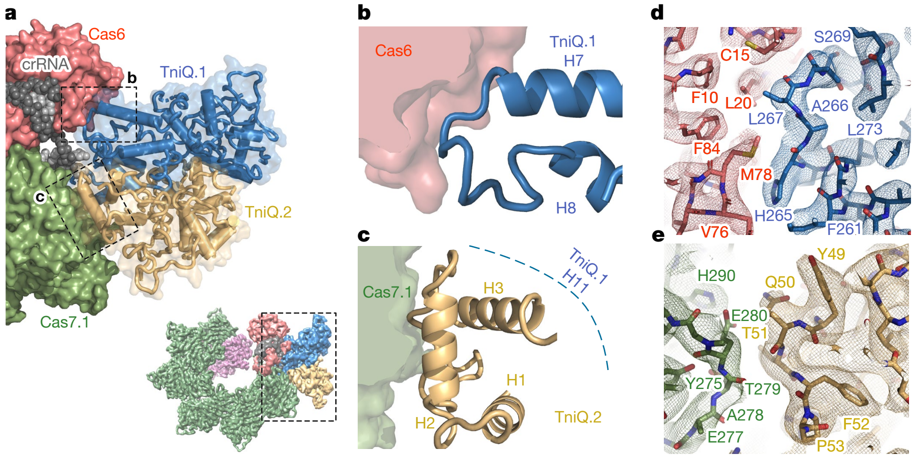


Fig. 3 | Cas6 and Cas7.1 form a binding platform for TniQ. a, Top, magnified area showing the interaction site of Cascade and the TniQ dimer. Cas6 and Cas7.1 are displayed as molecular Van der Waals surfaces, the crRNA is shown as grey spheres and the TniQ monomers are shown as ribbons. **b**, The loop connecting TniQ.1 α-helices H7 and H8 (blue) binds within a hydrophobic cavity of Cas6. **c**, Cas7.1 interacts with the HTH domain of the TniQ.2 monomer (yellow), mainly through H2 and the loop connecting H2 and H3. **d**, **e**, Experimental cryo-EM densities observed for the TniQ–Cas6 (**d**) and TniQ–Cas7.1 (**e**) interactions.

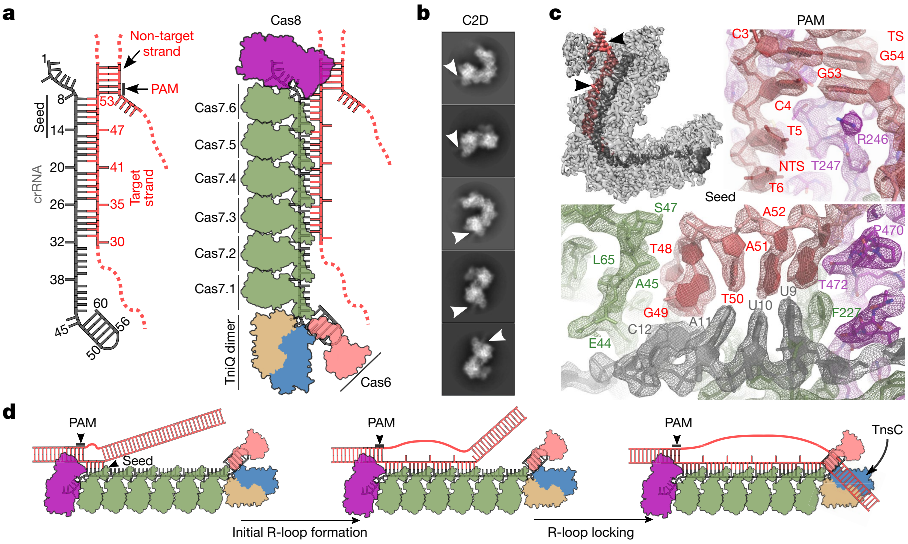


Fig. 4 | DNA-bound structure of the TniQ–Cascade complex. a, Schematic of crRNA and the portion of the dsDNA substrate that was experimentally observed within the electron density map for DNA-bound TniQ–Cascade. The target strand, non-target strand, PAM and seed regions are indicated (left); protein components are shown on the right. **b**, Selected cryo-EM reference-free two-dimensional class averages for DNA-bound TniQ–Cascade; density corresponding to dsDNA could be directly observed protruding from the Cas8 component in the two-dimensional class averages (white arrows). **c**, Cryo-EM map for DNA-bound TniQ–Cascade. The crRNA is shown in dark grey and the DNA is shown in red. Right bottom, detailed views of the PAM and seed-recognition regions of the map, with refined models represented as sticks within the electron density. Cas8 is shown in purple, Cas7 is shown in green, the crRNA is in grey and DNA is shown in red. NTS, non-target strand. **d**, The *V. cholerae* transposon encodes a TniQ–Cascade co-complex that uses the sequence content of the crRNA to bind complementary DNA target sites. We propose that the incomplete R-loop observed in our structure (middle) represents an intermediate state that may precede a downstream 'locking' step involving proofreading of the RNA–DNA complementarity. TniQ is positioned at the PAM-distal end of the DNA-bound Cascade complex, where it probably interacts with TnsC during downstream steps of RNA-guided DNA insertion.

are detrimental for RNA-guided DNA integration[7]. Thus, we propose that the partial R-loop structure that we observed could represent an intermediate conformation refractory to integration, and that further structural rearrangements may be critical for further stabilization of an open conformation, possibly driven by recruitment of the TnsC ATPase.

Here we present cryo-EM structures of a CRISPR–Cas effector complex bound to the transposition protein TniQ, with and without target DNA. These structures reveal the unexpected presence of TniQ as a dimer that forms bipartite interactions with Cas6 and Cas7.1 within the Cascade complex, forming a probable recruitment platform for downstream-acting transposition proteins[19] (Fig. 4d). Our structures further reveal a possible fidelity checkpoint, whereby formation of a complete R-loop requires conformational rearrangements that may depend on extensive RNA–DNA complementarity and/or downstream factor recruitment; this proofreading step could account for the highly specific RNA-guided DNA integration that we previously reported for the *V. cholerae* transposon[7]. In light of recent work demonstrating exaptation of type V-K CRISPR–Cas systems by similar Tn7-like transposons that also encode TniQ[20,21], it will be informative to determine whether tethering of TniQ to evolutionarily distinct crRNA effector complexes—Cascade or Cas12k—is a general theme of RNA-guided transposition.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1849-0.

1. Dy, R. L., Richter, C., Salmond, G. P. C. & Fineran, P. C. Remarkable mechanisms in microbes to resist phage infections. *Annu. Rev. Virol*. **1**, 307–331 (2014).
2. Hille, F. et al. The biology of CRISPR–Cas: backward and forward. *Cell* **172**, 1239–1259 (2018).
3. Doron, S. et al. Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* **359**, eaar4120 (2018).
4. Sinkunas, T. et al. In vitro reconstitution of Cascade-mediated CRISPR immunity in *Streptococcus thermophilus*. *EMBO J*. **32**, 385–394 (2013).
5. Redding, S. et al. Surveillance and processing of foreign DNA by the *Escherichia coli* CRISPR–Cas system. *Cell* **163**, 854–865 (2015).
6. Peters, J. E., Makarova, K. S., Shmakov, S. & Koonin, E. V. Recruitment of CRISPR–Cas systems by Tn7-like transposons. *Proc. Natl Acad. Sci. USA* **114**, E7358–E7366 (2017).
7. Klompe, S. E., Vo, P. L. H., Halpin-Healy, T. S. & Sternberg, S. H. Transposon-encoded CRISPR–Cas systems direct RNA-guided DNA integration. *Nature* **571**, 219–225 (2019).
8. Peters, J. E. Tn7. *Microbiol. Spectr*. **2**, https://doi.org/10.1128/microbiolspec.MDNA3-0010-2014 (2014).
9. Jackson, R. N. et al. Crystal structure of the CRISPR RNA-guided surveillance complex from *Escherichia coli*. *Science* **345**, 1473–1479 (2014).
10. Chowdhury, S. et al. Structure reveals mechanisms of viral suppressors that intercept a CRISPR RNA-guided surveillance complex. *Cell* **169**, 47–57 (2017).
11. Guo, T. W. et al. Cryo-EM structures reveal mechanism and inhibition of DNA targeting by a CRISPR–Cas surveillance complex. *Cell* **171**, 414–426 (2017).
12. Mulepati, S., Héroux, A. & Bailey, S. Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target. *Science* **345**, 1479–1484 (2014).
13. Zivanov, J. et al. New tools for automated high-resolution cryo-EM structure determination in RELION-3. *eLife* **7**, 163 (2018).
14. Holder, J. W. & Craig, N. L. Architecture of the Tn7 posttransposition complex: an elaborate nucleoprotein structure. *J. Mol. Biol*. **401**, 167–181 (2010).
15. Holm, L. & Laakso, L. M. Dali server update. *Nucleic Acids Res*. **44** (W1), W351–W355 (2016).
16. Aravind, L., Anantharaman, V., Balaji, S., Babu, M. M. & Iyer, L. M. The many faces of the helix–turn–helix domain: transcription regulation and beyond. *FEMS Microbiol. Rev*. **29**, 231–262 (2005).
17. Krissinel, E. Stock-based detection of protein oligomeric states in jsPISA. *Nucleic Acids Res*. **43** (W1), W314–W319 (2015).
18. Xiao, Y., Luo, M., Dolan, A. E., Liao, M. & Ke, A. Structure basis for RNA-guided DNA degradation by Cascade and Cas3. *Science* **361**, eaat0839 (2018).
19. Choi, K. Y., Spencer, J. M. & Craig, N. L. The Tn7 transposition regulator TnsC interacts with the transposase subunit TnsB and target selector TnsD. *Proc. Natl Acad. Sci. USA* **111**, E2858–E2865 (2014).
20. Faure, G. et al. CRISPR–Cas in mobile genetic elements: counter-defence and beyond. *Nat. Rev. Microbiol*. **17**, 513–525 (2019).
21. Strecker, J. et al. RNA-guided DNA insertion with CRISPR-associated transposases. *Science* **365**, 48–53 (2019).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

TniQ–Cascade purification

Protein components of TniQ–Cascade were expressed from a pET-derivative vector containing the native *V. cholerae tniQ-cas8-cas7-cas6* operon with an N-terminal His_{10}-MBP-TEV site fusion on TniQ. The crRNA was expressed separately from a pACYC-derivative vector containing a minimal repeat–spacer–repeat CRISPR array encoding a spacer from the endogenous *V. cholerae* CRISPR array. The TniQ–Cascade complex was overexpressed and purified as described previously[7], and was stored in Cascade storage buffer (20 mM Tris-Cl, pH 7.5, 200 mM NaCl, 1 mM DTT, 5% glycerol).

Sample preparation for electron microscopy

For negative staining, 3 µl of purified TniQ–Cascade ranging from 100 nM to 2 µM was incubated with plasma treated (H_2/O_2 gas mix, Gatan Solarus) CF400 carbon-coated grids (EMS) for 1 min. Excess solution was blotted and 3 µl of 0.75% uranyl formate was added for an additional minute. Excess stain was blotted away and grids were air-dried overnight. Grid screening for both negative staining and cryo conditions was performed on a Tecnai-F20 microscope (FEI) operated at 200 KeV and equipped with a Gatan K2-Summit direct detector. Microscope operation and data collection were carried out using the Leginon/Appion software. Initial negative staining grid screening allowed determination of a suitable concentration range for cryo conditions. Several grid geometries were tested in the 1–4 µM concentration range for cryo conditions using a Vitrobot Mark-II operated at 4 °C, 100% humidity, blot force 3, drain time 0, waiting time 15 s, and blotting times ranging from 3–5 s. The best ice distribution and particle density was obtained with 0.6/1 UltrAuFoil grids (Quantifoil).

Electron microscopy

A preliminary dataset of 300 images in cryo was collected with the Tecnai-F20 microscope using a pixel size of 1.22 Å/pixel with illumination conditions adjusted to 8 e^-/pixel/second with a frame window of 200 ms. Preprocessing and image processing were integrally done in Relion3[13] with ctf estimation integrated via a wrapper to Gctf[22]. An initial model computed using the SGD algorithm[23] implemented in Relion3 was used as initial reference for a refine three-dimensional job that generated a sub-nanometric reconstruction with approximately 10,000 selected particles. Clear secondary structure features in the two-dimensional averages and the three-dimensional reconstruction could be identified.

For the DNA-bound TniQ–Cascade complex containing DNA, we pre-incubated two complementary 74-nt oligonucleotides: (NTS: 5′-TTCATCAAGCCATTGGACCGCCTTACAGGACGCTTTGGCTTCATTGCTTTTCAGCTTCGCCTTGACGGCCAAAA-3′, TS: 5′-TTTTGGCCGTCAAGGCGAAGCTGAAAAGCAATGAAGCCAAAGCGTCCTGTAAGGCGGTCCAATGGCTTGATGAA-3′) for 5 min at 95 °C in hybridization buffer (20 mM Tris-Cl, pH 7.5, 100 mM KCl, 5 mM $MgCl_2$) to form dsDNA, which was subsequently aliquoted and flash-frozen. Complex formation was performed by incubating a 3× molar excess of dsDNA with TniQ–Cascade at 37 °C for 5 min before vitrification, which followed the conditions optimized for the apo complex (defined as TniQ–Cascade with crRNA but no DNA ligand).

High-resolution data for the apo complex were collected in a Tecnai-Polara-F30 microscope operated at 300 KeV equipped with a K3 direct detector (Gatan). A 30-µm C2 aperture was used with a pixel size of 0.95 Å/pixel and illumination conditions in microprobe mode adjusted to a fluence of 16 e^-/pixel/second. Four-second images with a frame width of 100 ms (1.77 e^-/Å^2/frame) were collected in counting mode.

For the DNA-bound complex, high-resolution data were collected in a Titan Krios microscope (FEI) equipped with an energy filter (20 eV slit width) and a K2 direct detector (Gatan) operated at 300 KeV. A 50-µm C2 aperture was used with a pixel size of 1.06 Å/pixel and illumination conditions adjusted in nanoprobe mode to a fluence of 8 e^-/pixel/second. Eight-second images with a frame width of 200 ms (1.42 e^-/Å^2/frame) were collected in counting mode.

Image processing

Motion correction was performed for every micrograph applying the algorithm described for Motioncor2[24] implemented in Relion3 with 5 by 5 patches for the K2 data and 7 by 5 patches for the K3 data. Parameters of the contrast transfer function for each motion-corrected micrograph were obtained using Gctf[22] integrated in Relion3. Initial particle picking of a subset of 200 images randomly chosen was performed with the Laplacian tool of the Auto-picking module of Relion3, using an estimated size for the complex of 200 Å. Then, 15,000 particles were extracted in a 300-pixel box size and binned 3 times for an initial two-dimensional classification job. Selected two-dimensional averages from this job were used as templates for Auto-picking of the full dataset. The full dataset of binned particles was subjected to a two-dimensional classification job to identify particles able to generate averages with clear secondary structure features. The selected subgroup of binned particles after the two-dimensional classification selection was refined against a three-dimensional volume obtained by SGD with the F20 data. This consensus volume was inspected to localize areas of heterogeneity that were clearly identified at both ends of the crescent shape characteristic of this complex. Both ends were then individually masked using soft masks of around 20 pixels that were subsequently used in classification jobs without alignments in Relion3. The T parameter used for this classification job was 6 and the total number of classes was 10. This strategy allowed us to identify two main population of particles which correspond to an open and closed state of the complex. Particles from both subgroups were separately re-extracted to obtain unbinned data sets for further refinement. New features implemented in Relion3, namely Bayesian polishing and ctf parameters refinement, allowed the extension of the resolution to 3.4, 3.5 and 2.9 Å for the two apo and the DNA-bound complexes, respectively. Post-processing was performed with a soft-mask of 5 pixels being the B-factor estimated automatically in Relion3 following standard practice. A final set of local refinements was performed with the masks used for classification. The locally aligned maps exhibit very good quality for the ends of the C-shape. These maps were used for de novo modelling and initial model refinement.

Model building and refinement

For the Cas7 and Cas6 monomers, the *E. coli* homologues (PDB accession code 4TVX) were initially docked with Chimera[25] and transformed to poly-alanine models. Substantial rearrangement of the finger region of Cas7 monomers, as well as other secondary structure elements of Cas6, were performed manually in COOT[26] before amino acid substitution of the poly-alanine model. Well-defined bulky side chains of aromatic residues allowed a confident assignment of the register. The crRNA was also well defined in the maps and was traced de novo with COOT. For Cas8 and TniQ in particular, no structural similarity was found in the published structures that was able to explain our densities. Locally refined maps using soft masks at both ends of the crescent-shaped complex rendered well-defined maps below 3.5 Å resolution. These maps were used for manual de novo tracing of a poly-alanine model in COOT that was subsequently mutated to the *V. cholerae* sequences. Bulky side chains for aromatic residues showed excellent density and were used as landmarks to adjust the register of the sequence.

For refinement, an initial step of real-space refinement against the cryo-EM maps was performed with the phenix.real_space refinement tool of the Phenix package[27], with secondary structure restraints activated. A second step of reciprocal space refinement was performed in Refmac5[28], with secondary restraints calculated with Prosmart[29] and LibG[30]. Weight of the geometry term versus the experimental term was adjusted to avoid overfitting of the model into cryo-EM map, as previously reported[31]. Model validation was performed in Molprobity[32].

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Maps and models have been deposited in the Electron Microscopy Data Bank with accession codes EMD-20349, EMD-20350 and EMD-20351 and the Protein Data Bank with accession codes 6PIF, 6PIG and 6PIJ.

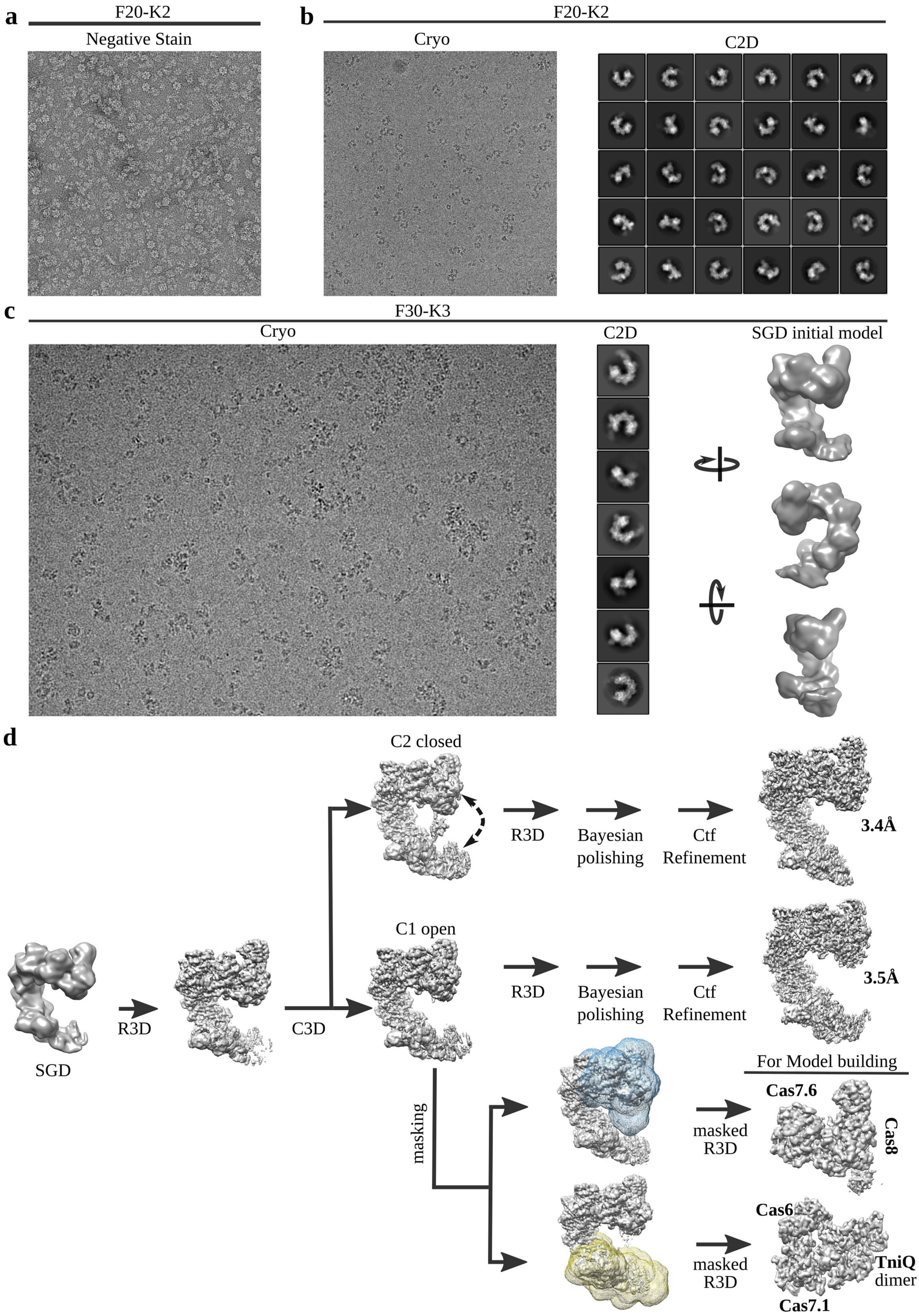
22. Zhang, K. Gctf: Real-time CTF determination and correction. *J. Struct. Biol*. **193**, 1–12 (2016).
23. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).
24. Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
25. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem*. **25**, 1605–1612 (2004).
26. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
27. Afonine, P. V. et al. Real-space refinement in PHENIX for cryo-EM and crystallography. *Acta Crystallogr. D* **74**, 531–544 (2018).
28. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53**, 240–255 (1997).
29. Nicholls, R. A., Fischer, M., McNicholas, S. & Murshudov, G. N. Conformation-independent structural comparison of macromolecules with ProSMART. *Acta Crystallogr. D* **70**, 2487–2499 (2014).
30. Brown, A. et al. Tools for macromolecular model building and refinement into electron cryo-microscopy reconstructions. *Acta Crystallogr. D* **71**, 136–153 (2015).
31. Fernández, I. S., Bai, X.-C., Murshudov, G., Scheres, S. H. W. & Ramakrishnan, V. Initiation of translation by cricket paralysis virus IRES requires its translocation in the ribosome. *Cell* **157**, 823–831 (2014).
32. Williams, C. J. et al. MolProbity: more and better reference data for improved all-atom structure validation. *Protein Sci*. **27**, 293–315 (2018).
33. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* **11**, 63–65 (2014).

Acknowledgements We thank B. Grassucci and Z. Zhang for technical assistance with cryo-EM data acquisition. Part of this work was performed at the Simons Electron Microscopy Center and National Resource for Automated Molecular Microscopy located at the New York Structural Biology Center, supported by grants from the Simons Foundation (SF349247), NYSTAR and the NIH National Institute of General Medical Sciences (GM103310).

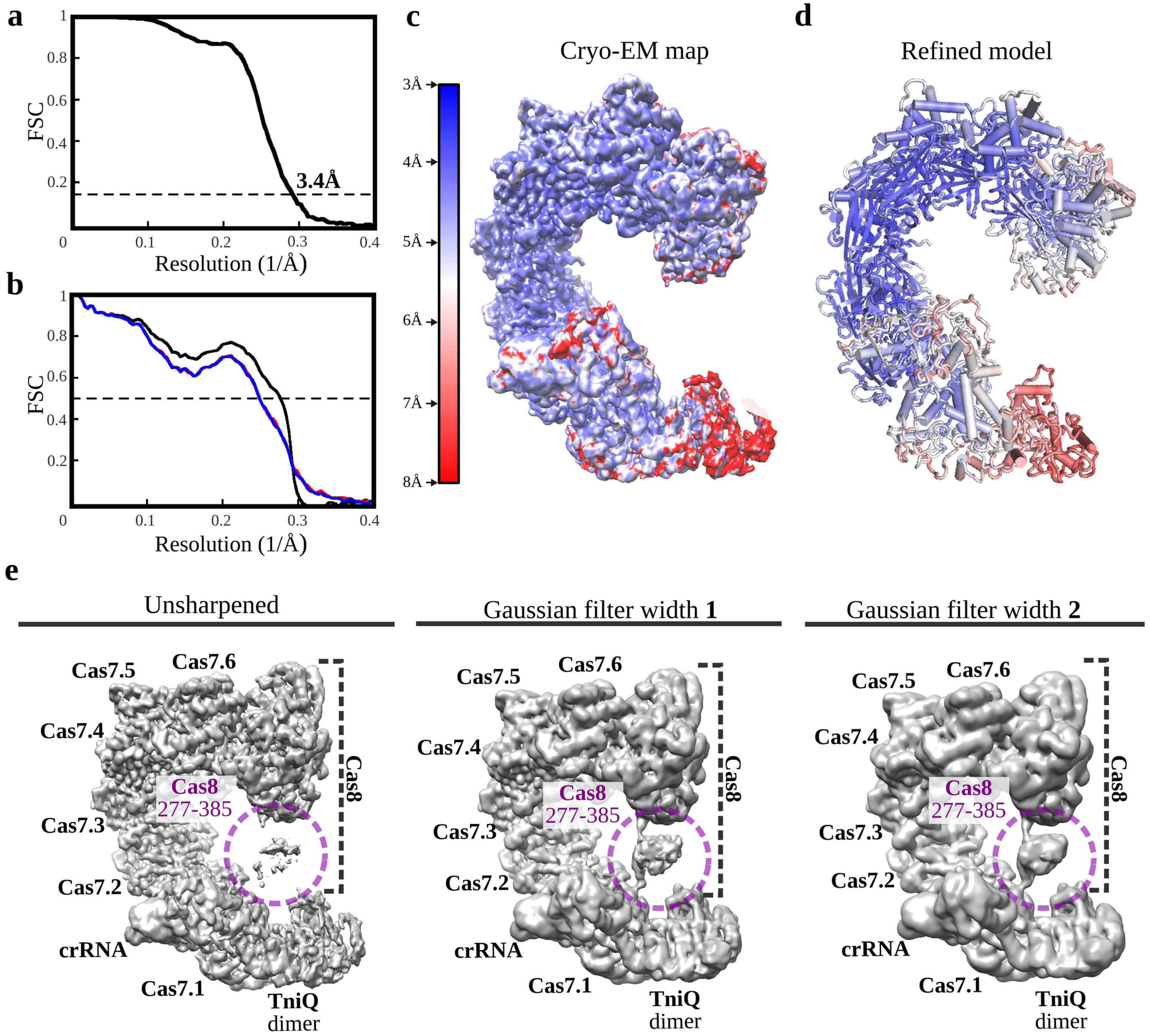
Author contributions All authors conceived and designed the project. T.S.H.-H. purified ribonucleoprotein complexes and assisted in cryo-EM data acquisition. I.S.F. collected EM data and solved the structures. I.S.F., S.H.S. and the other authors discussed the data and wrote the manuscript.

Competing interests Columbia University has filed a patent application related to this work. S.E.K. and S.H.S. are inventors on other patents and patent applications related to CRISPR–Cas systems and uses thereof. S.H.S. is a co-founder and scientific advisor to Dahlia Biosciences and an equity holder in Dahlia Biosciences and Caribou Biosciences.

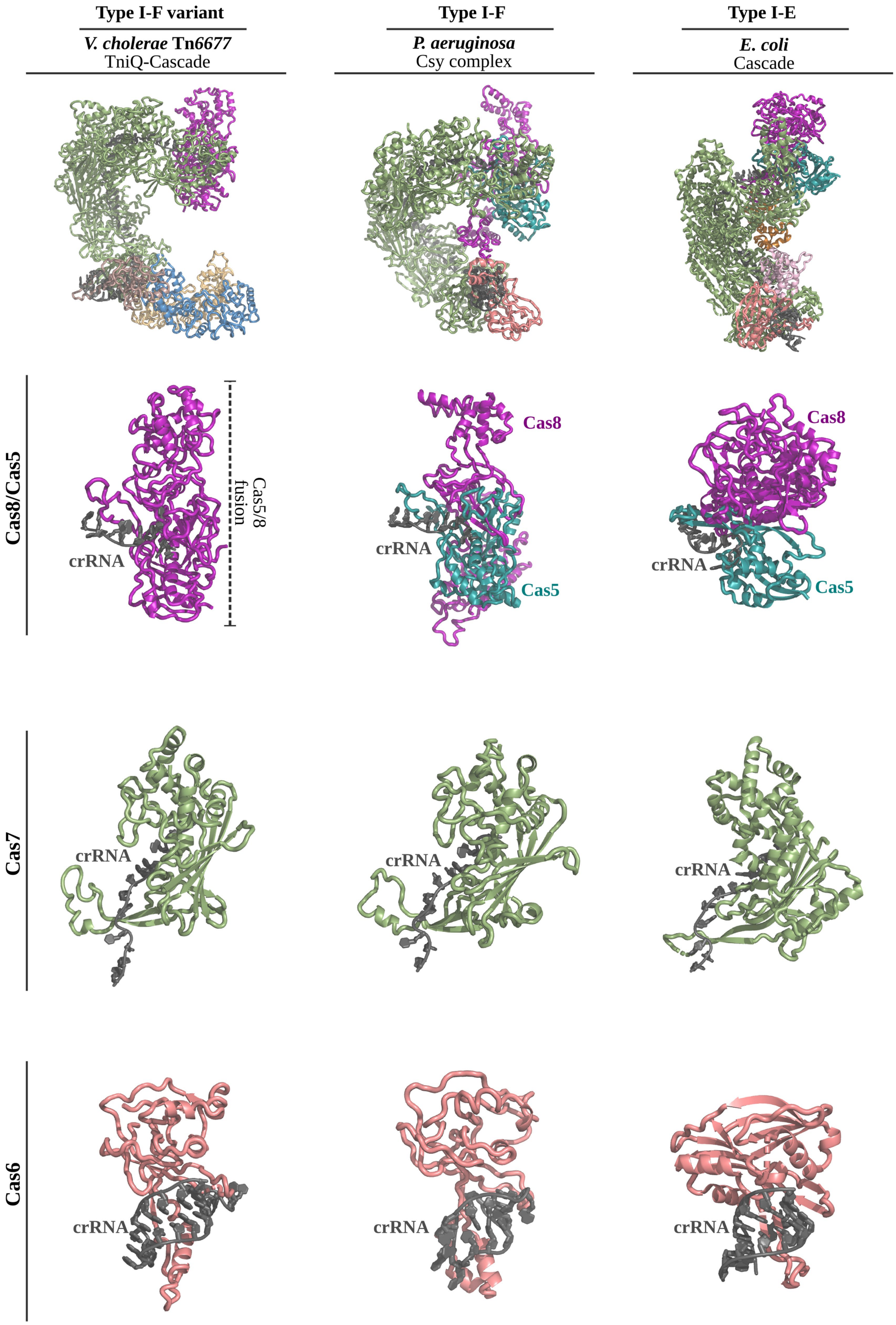
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1849-0.
Correspondence and requests for materials should be addressed to S.H.S. or I.S.F.
Peer review information *Nature* thanks Scott Bailey, Ronald Chalmers, David Taylor and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



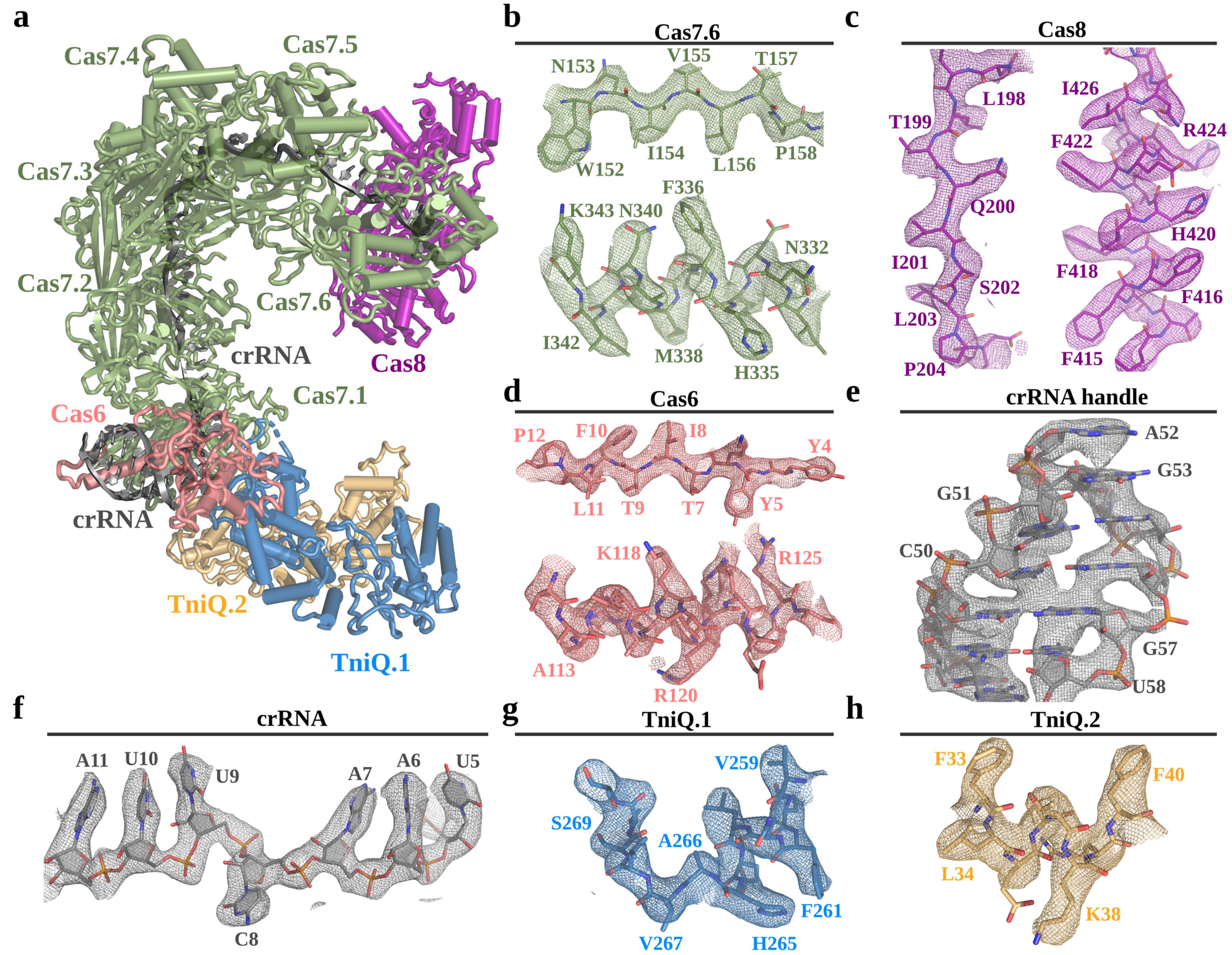
Extended Data Fig. 1 | Cryo-EM sample optimization and image processing workflow. **a**, Representative negatively stained micrograph for 500 nM TniQ–Cascade. **b**, Left, representative cryo-EM image for 2 µM TniQ–Cascade. A small dataset of 200 images was collected in a Tecnai-F20 microscope equipped with a Gatan K2 camera. Right, reference-free two-dimensional class averages for this initial cryo-EM dataset. **c**, Left, representative image from a large dataset collected in a Tecnai Polara microscope equipped with a Gatan K3 detector. Middle, detailed two-dimensional class averages were obtained that were used for initial model generation using the SGD algorithm[23] implemented in Relion3[13] (right). **d**, Image processing workflow used to identify the two main classes of the TniQ–cascade complex in open and closed conformations. Local refinements with soft masks were used to improve the quality of the map within the terminal protuberances of the complex. These maps were instrumental for de novo modelling and initial model refinement.



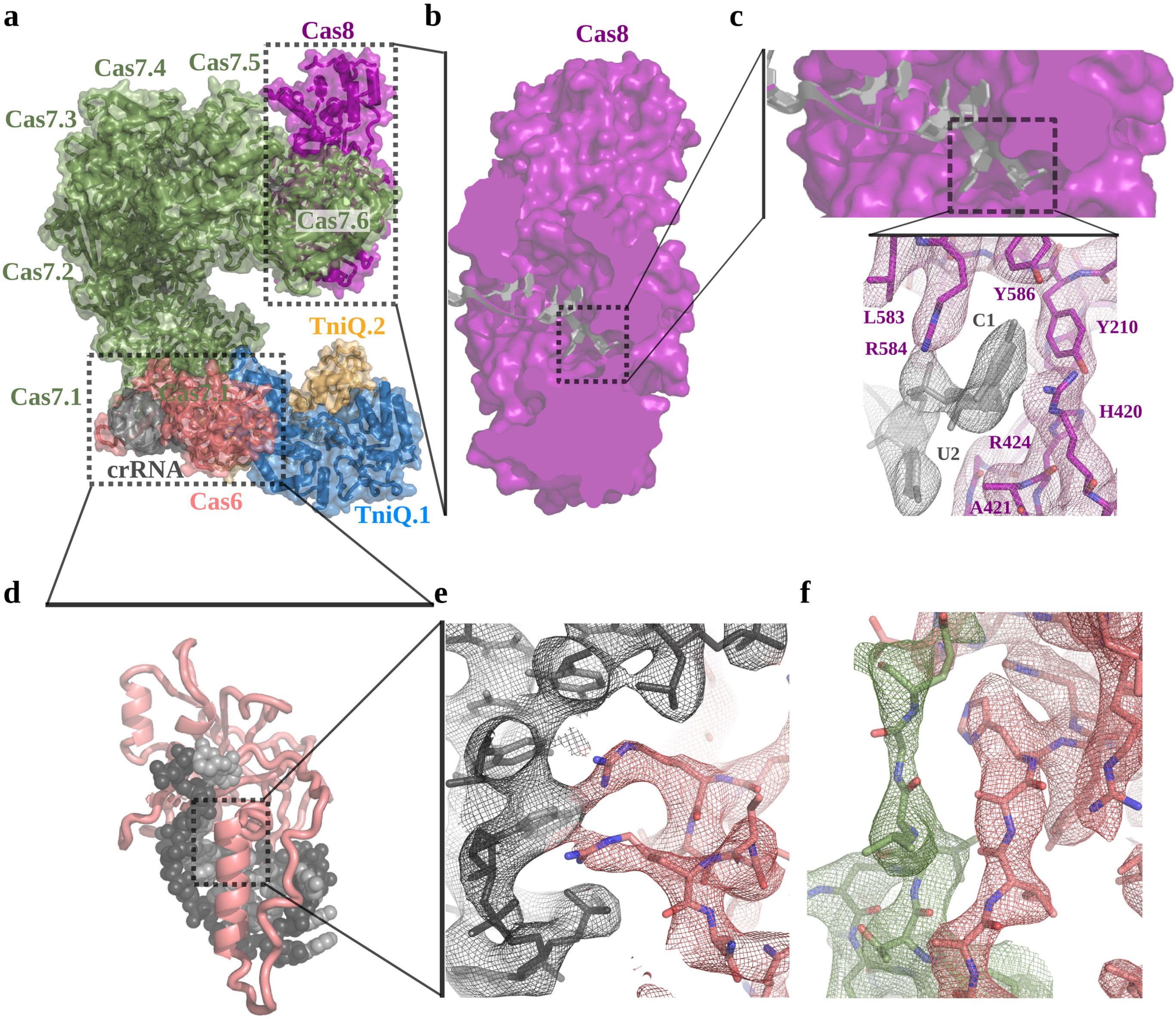
Extended Data Fig. 2 | Fourier shell correlation curves, local resolution, and unsharpened filter maps for the TniQ–Cascade complex in closed conformation. **a**, Gold-standard Fourier shell correlation (FSC) curve using half maps; the global resolution estimation is 3.4 Å by the FSC 0.143 criterion. **b**, Cross-validation model-versus-map FSC. Blue curve, FSC between the shacked model refined against half map 1; red curve, FSC against half map 2, not included in the refinement; black curve, FSC between final model against the final map. The overlap observed between the blue and red curves guarantees a non-overfitted model. **c**, Unsharpened map coloured according to local resolutions, as reported by RESMAP[33]. **d**, Final model coloured according to B-factors calculated by REFMAC[28]. **e**, A flexible Cas8 domain encompassing residues 277–385 contacts the TniQ dimer at the other side of the crescent shape. Applying a Gaussian filter of increasing width to the unsharpened map allows for a better visualization of this flexible region.



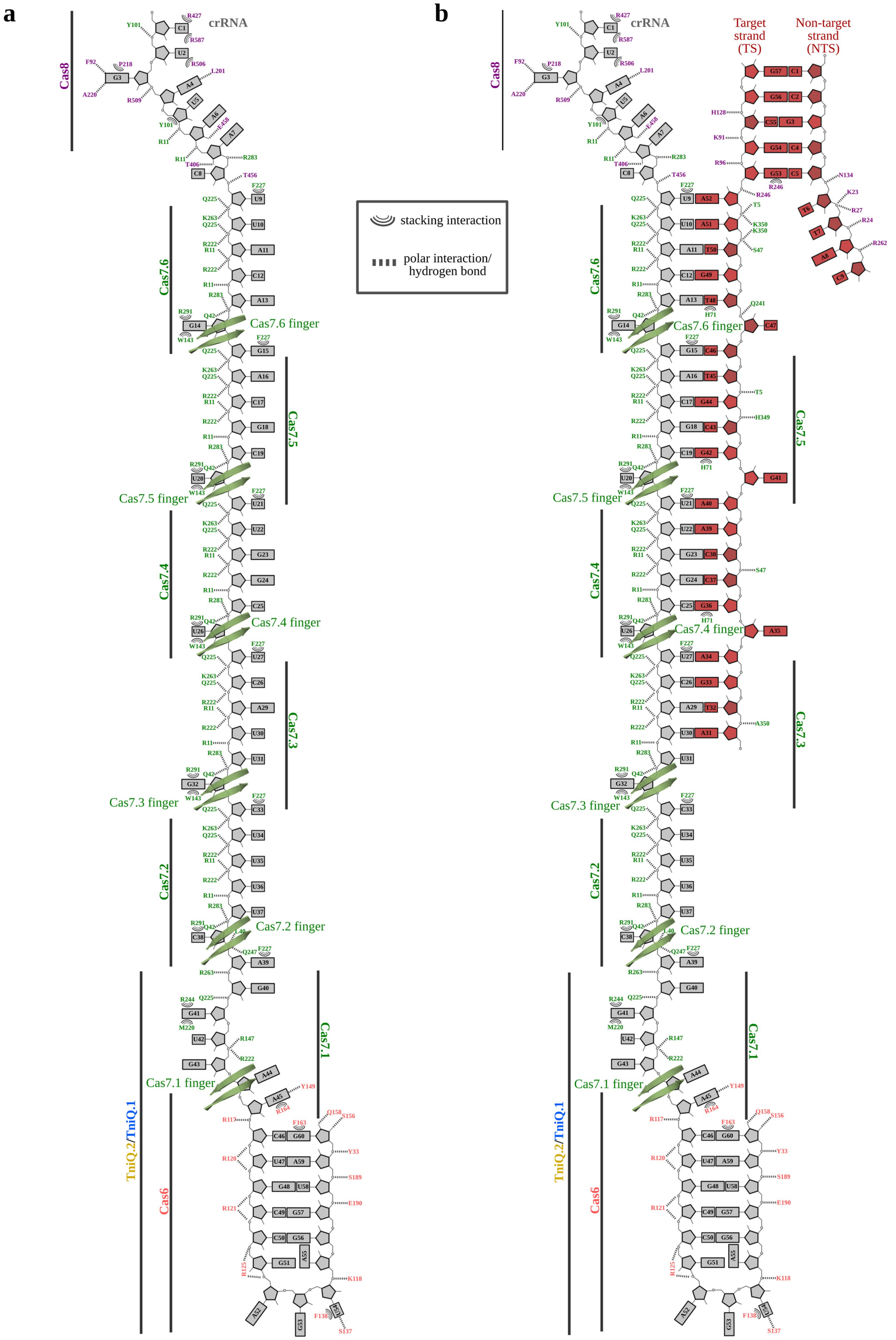
Extended Data Fig. 3 | Alignment of TniQ–Cascade with structurally similar Cascade complexes. The *V. cholerae* I-F variant TniQ–Cascade complex (left) was superposed with *Pseudomonas aeruginosa* I-F Cascade[11] (also known as Csy complex; middle, PDB ID: 6B45) and *E. coli* I-E Cascade[9] (right, PDB ID: 4TVX). Shown are alignments of the entire complex (top), the Cas8 and Cas5 subunits with the 5′ crRNA handle (second from top), the Cas7 subunit with a fragment of crRNA (second from bottom) and the Cas6 subunit with the 3′ crRNA handle (bottom).



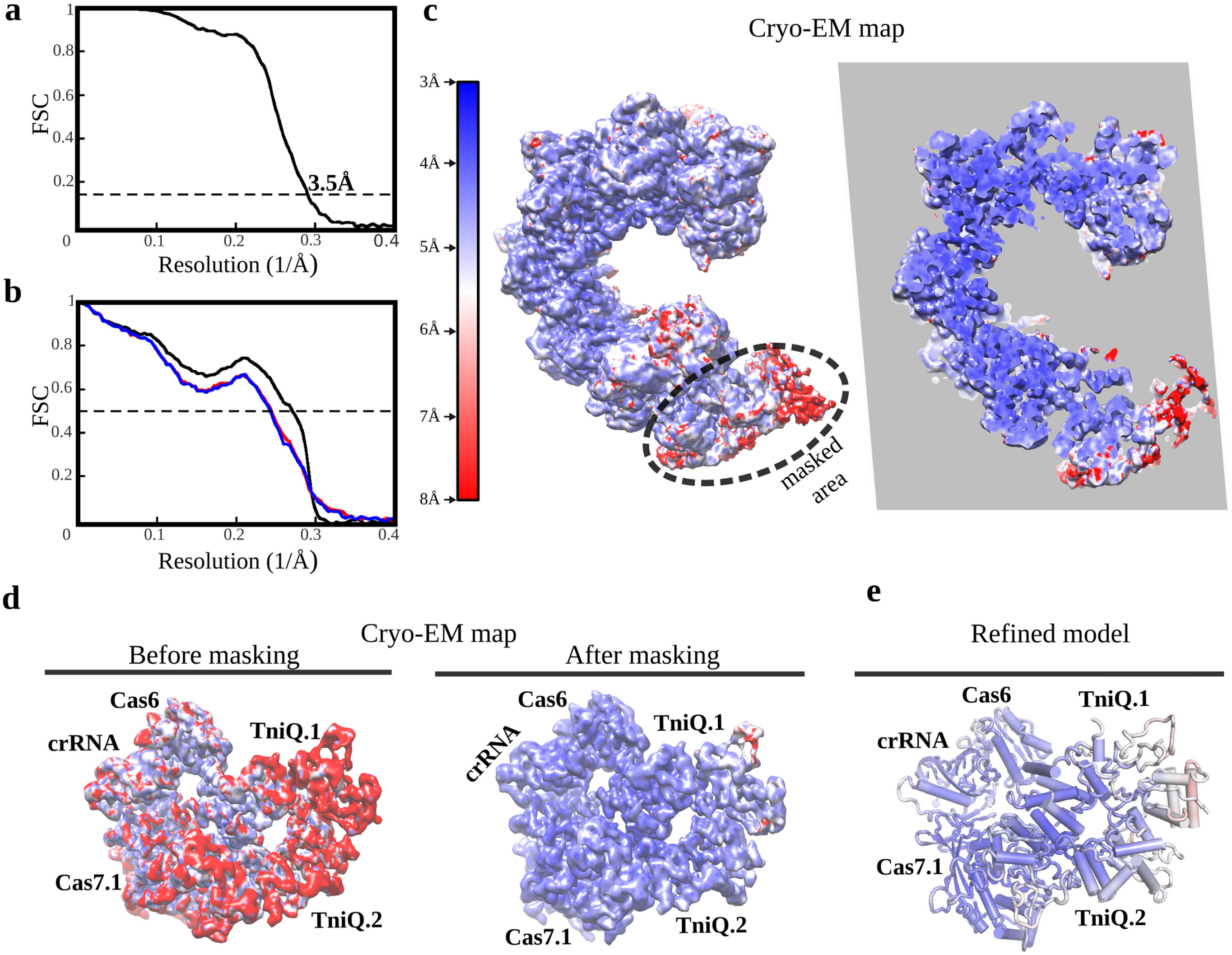
Extended Data Fig. 4 | Representative cryo-EM densities for all the components of the TniQ–Cascade complex in closed conformation. a, Final refined model of TniQ–Cascade, with Cas8 in purple, Cas7 monomers in green, Cas6 in salmon, the TniQ monomers in blue and yellow, and the crRNA in grey. **b**–**h**, Final refined model inserted in the final cryo-EM density for select regions of all the molecular components of the TniQ–Cascade complex. Residues are numbered.



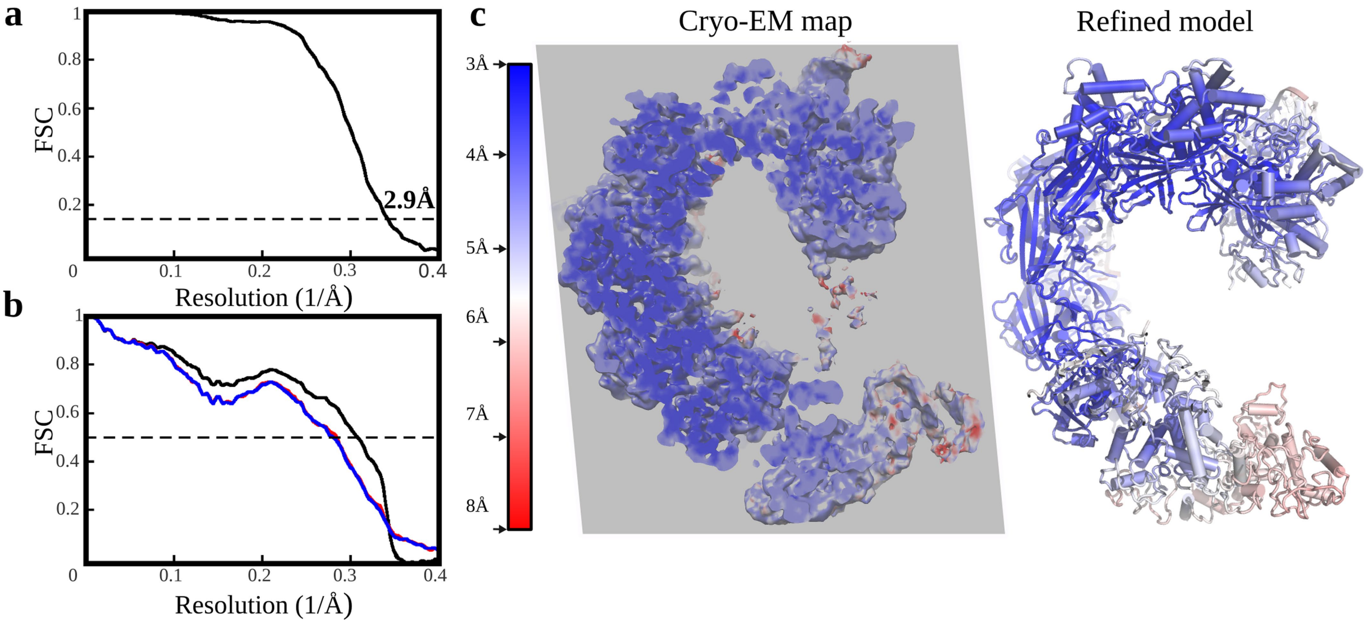
Extended Data Fig. 5 | Cas8 and Cas6 interaction with the crRNA. a, Refined model for the TniQ–Cascade shown as ribbons inserted in the semi-transparent Van der Waals surface, coloured as in Fig. 1. **b**, **c**, Magnified view of Cas8, which interacts with the 5′ end of the crRNA. The inset shows electron density for the highlighted region, where the base of nucleotide C1 is stabilized by stacking interactions with arginine residues R584 and R424. **d**, Cas6 interacts with the 3′ end of the crRNA 'handle' (nucleotides 45–60). **e**, An arginine-rich α-helix is deeply inserted within the major groove of the terminal stem–loop. This interaction is mediated by electrostatic interactions between basic residues of Cas6 and the negatively charged phosphate backbone of the crRNA. **f**, Cas6 (salmon) also interacts with Cas7.1 (green), establishing a β-sheet formed by β-strands contributed from both proteins.



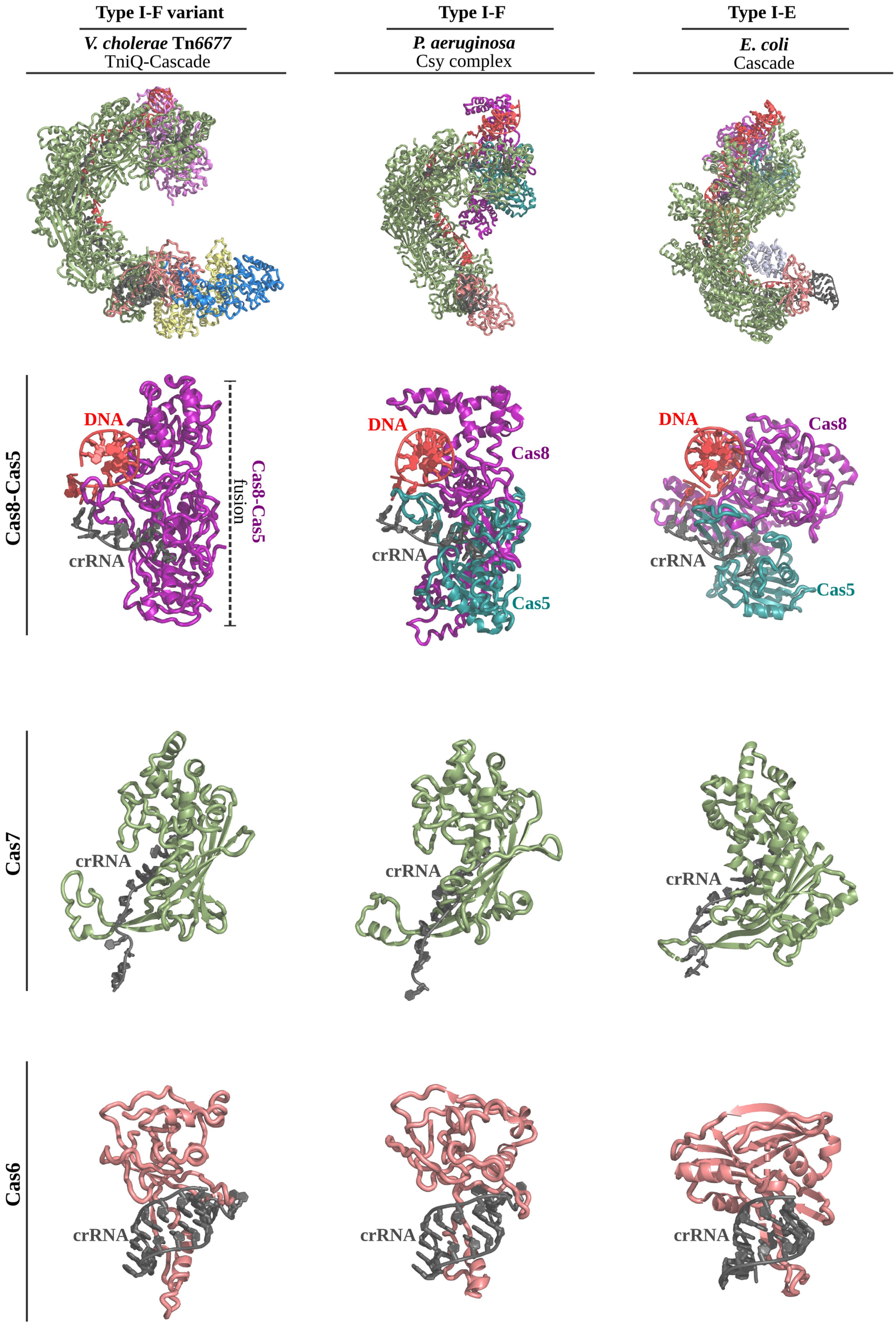
Extended Data Fig. 6 | Schematic representation of crRNA and target DNA recognition by TniQ–Cascade. a, TniQ–Cascade residues that interact with the crRNA are indicated. Approximate location for all protein components of the complex are also shown, as well as the position of each Cas7 finger. **b**, TniQ–Cascade residues that interact with crRNA and target DNA, shown as in **a**.



Extended Data Fig. 7 | FSC curves, local resolution, and local refined maps for the TniQ–Cascade complex in open conformation. **a**, Gold-standard FSC curve using half maps; the global resolution estimation is 3.5 Å by the FSC 0.143 criterion. **b**, Cross-validation model-versus-map FSC. Blue curve, FSC between shacked model refined against half map 1; red curve, FSC against half map 2, not included in the refinement; black curve, FSC between final model against the final map. The overlapping between the blue and red curves guarantees a non-overfitted model. **c**, Unsharpened map coloured according to local resolutions, as reported by RESMAP[33]. Right, slice through the map shown on the left. **d**, Local refinements with soft masks improved the maps in flexible regions. Shown is the region of the map corresponding to the TniQ dimer. Unsharpened maps coloured according to the local resolution estimations are shown before (left) and after (right) masked refinements. **e**, Final model for the TniQ dimer region, coloured according to the local B-factors calculated by REFMAC[28].



Extended Data Fig. 8 | FSC curves, local resolution, and unsharpened filter maps for the DNA-bound TniQ–Cascade complex complex. a, Gold-standard FSC curve using half maps; the global resolution estimation is 2.9 Å by the FSC 0.143 criterion. **b**, Cross-validation model-versus-map FSC. Blue curve, FSC between the shacked model refined against half map 1; red curve, FSC against half map 2, not included in the refinement; black curve, FSC between final model against the final map. The overlap observed between the blue and red curves guarantees a non-overfitted model. **c**, Left, unsharpened map coloured according to local resolutions, as reported by RESMAP[33]. dsDNA is visible at the top right projecting outside of the complex. Right, final model coloured according to B-factors calculated by REFMAC[28].



Extended Data Fig. 9 | Alignment of DNA-bound TniQ–Cascade with structurally similar Cascade complexes. The DNA-bound structure of *V. cholerae* I-F variant TniQ–Cascade complex (left) was superposed with DNA-bound structures of *P. aeruginosa* I-F Cascade[11] (also known as Csy complex; middle, PDB ID: 6B44) and *E. coli* I-E Cascade[9] (right, PDB ID: 5H9F). Shown are alignments of the entire complex (top), the Cas8 and Cas5 subunits with the 5′ crRNA handle and double-stranded PAM DNA (middle top), the Cas7 subunit with a fragment of crRNA (middle bottom), and the Cas6 subunit with the 3′ crRNA handle (bottom).

nature research

Corresponding author(s): Israel S. Fernandez

Last updated by author(s): Aug 13, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☒	□	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	□	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	□	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
☒	□	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	□	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	□	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	For cryoEM data collection we have used the Leginon/Appion package.
Data analysis	Ctf estimation was performed with Gctf v1.06 and Image processin with Relion v3.0-beta. Figures were generated with Chimera and PyMol.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

PDBs and cryoEM maps are deposited in the PDB and EMDB public databases with accession codes: 6PIF, 6PIG and 6PIJ.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

Describe how sample size was determined, detailing any statistical methods used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.

Data exclusions

Describe any data exclusions. If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.

Replication

Describe the measures taken to verify the reproducibility of the experimental findings. If all attempts at replication were successful, confirm this OR if there are any findings that were not replicated or cannot be reproduced, note this and describe why.

Randomization

Describe how samples/organisms/participants were allocated into experimental groups. If allocation was not random, describe how covariates were controlled OR if this is not relevant to your study, explain why.

Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging