ARTICLE

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 Tn*7*-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 (*attTn7*) 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

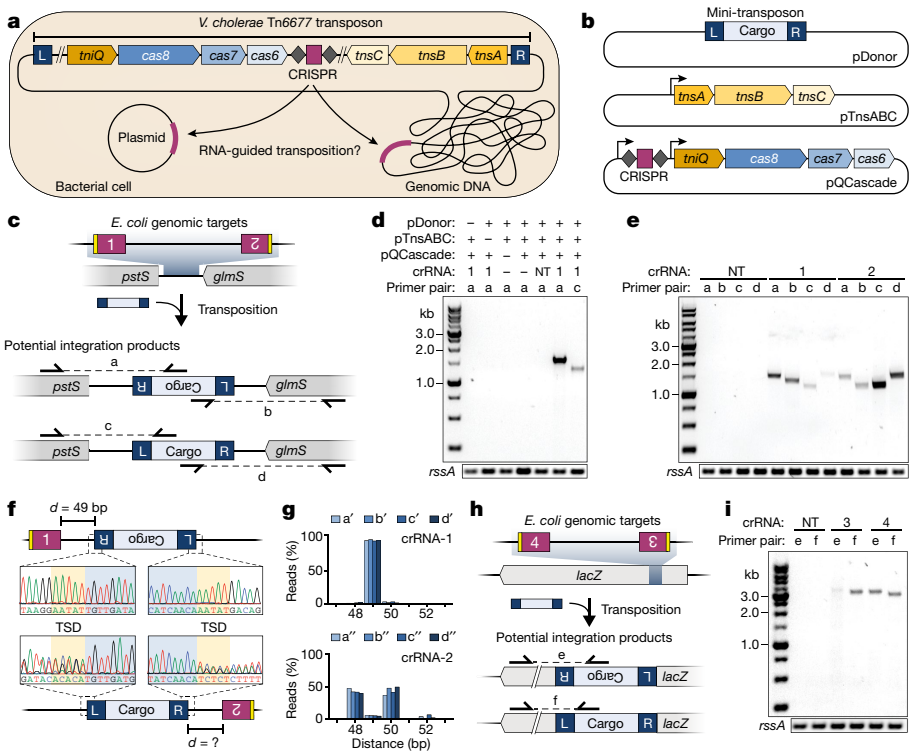


Fig. 1 | RNA-guided DNA integration with a *V. cholerae* transposon. **a**, Hypothetical scenario for Tn*6677* transposition into plasmid or genomic target sites complementary to a crRNA. **b**, Plasmid schematics for transposition experiments in which a mini-transposon on pDonor is mobilized in *trans*. 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.

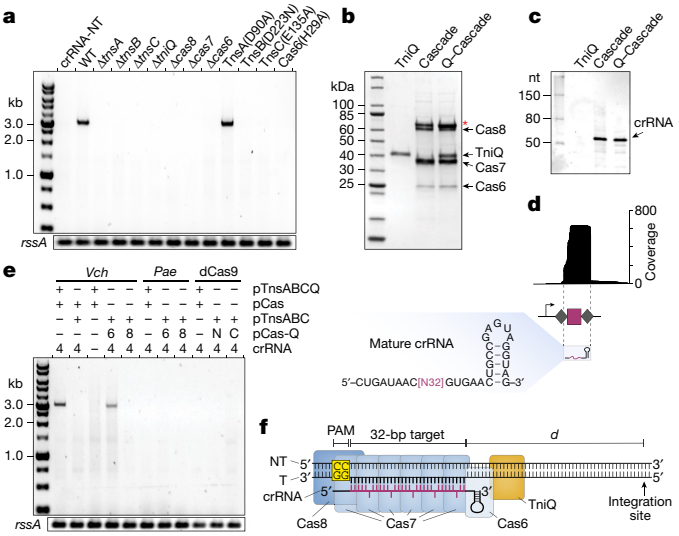


Fig. 2 | TniQ forms a complex with Cascade and is necessary for RNA-guided DNA integration. a, PCR analysis of transposition with crRNA-4 and a panel of gene deletions or point mutations, resolved by agarose gel electrophoresis. **b**, SDS–PAGE analysis of purified TniQ, Cascade and a TniQ–Cascade (Q–Cascade) 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

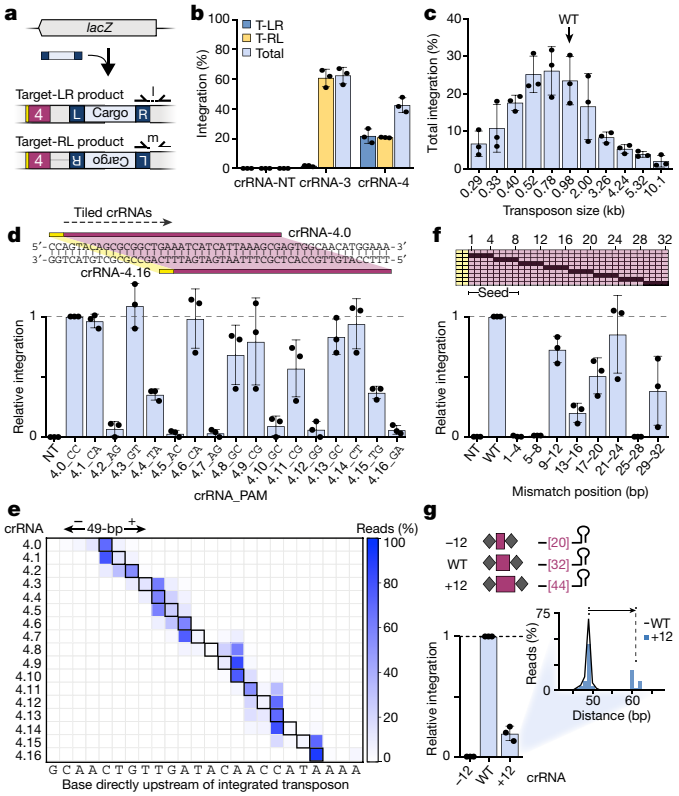


Fig. 3 | Influence of cargo size, PAM sequence, and crRNA mismatches on RNA-guided DNA integration. a, Schematic of alternative integration orientations and the primer pairs to selectively detect them by qPCR. **b**, qPCR-based quantification of transposition efficiency in both orientations with crRNA-NT, 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).

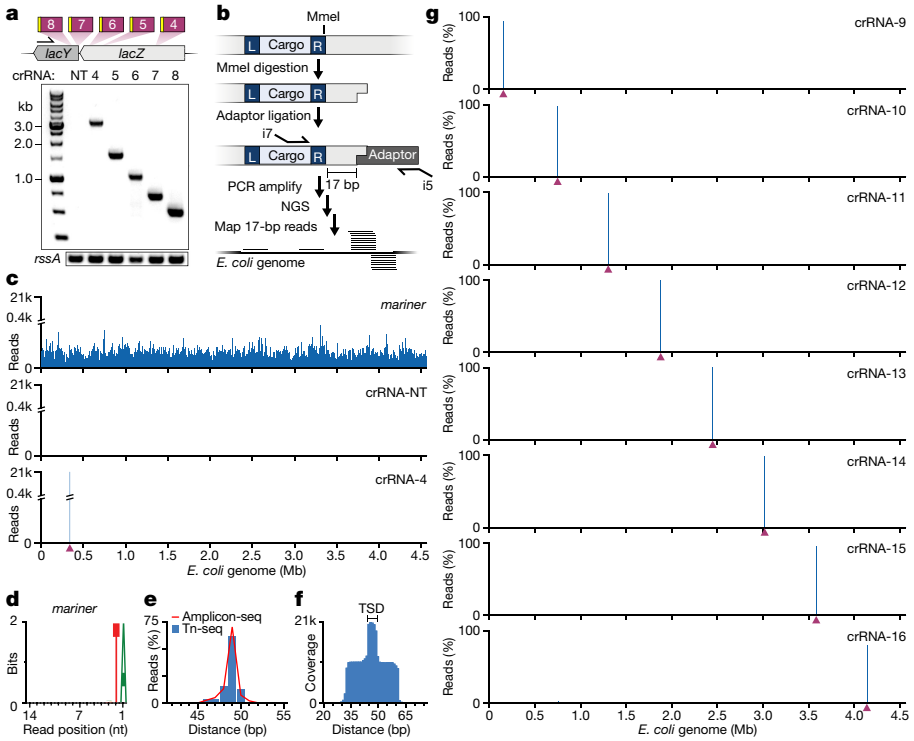


Fig. 4 | Genome-wide analysis of programmable RNA-guided DNA integration. a, Genomic locus targeted by crRNAs 4–8 (top), and PCR analysis of transposition resolved by agarose gel electrophoresis (bottom). Amplification of *rssA* serves as a loading control; gel source data may be found in Supplementary 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

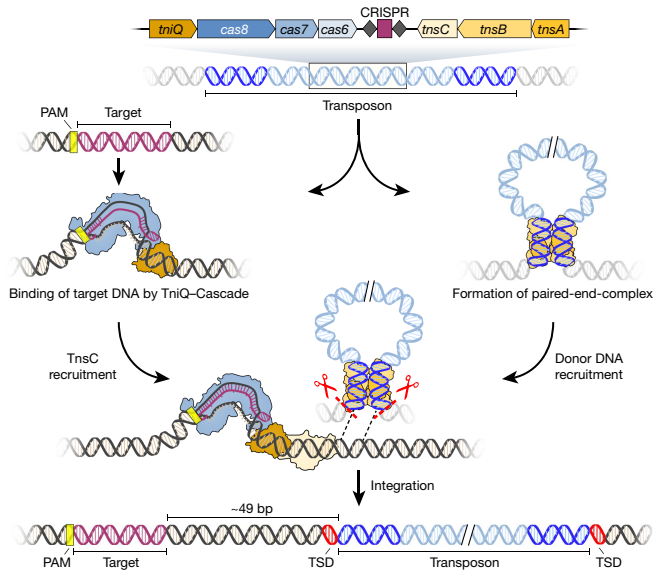


Fig. 5 | Proposed model for RNA-guided DNA integration by Tn*7*-like transposons encoding CRISPR–Cas systems. The *V. cholerae* Tn*6677* transposon encodes a programmable RNA-guided DNA-binding complex called Cascade, which we have shown forms a co-complex with TniQ. We propose that TniQ–Cascade complexes 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).

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 $OD_{600} = 2.0$) 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 $2^{\Delta Cq}$, 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.

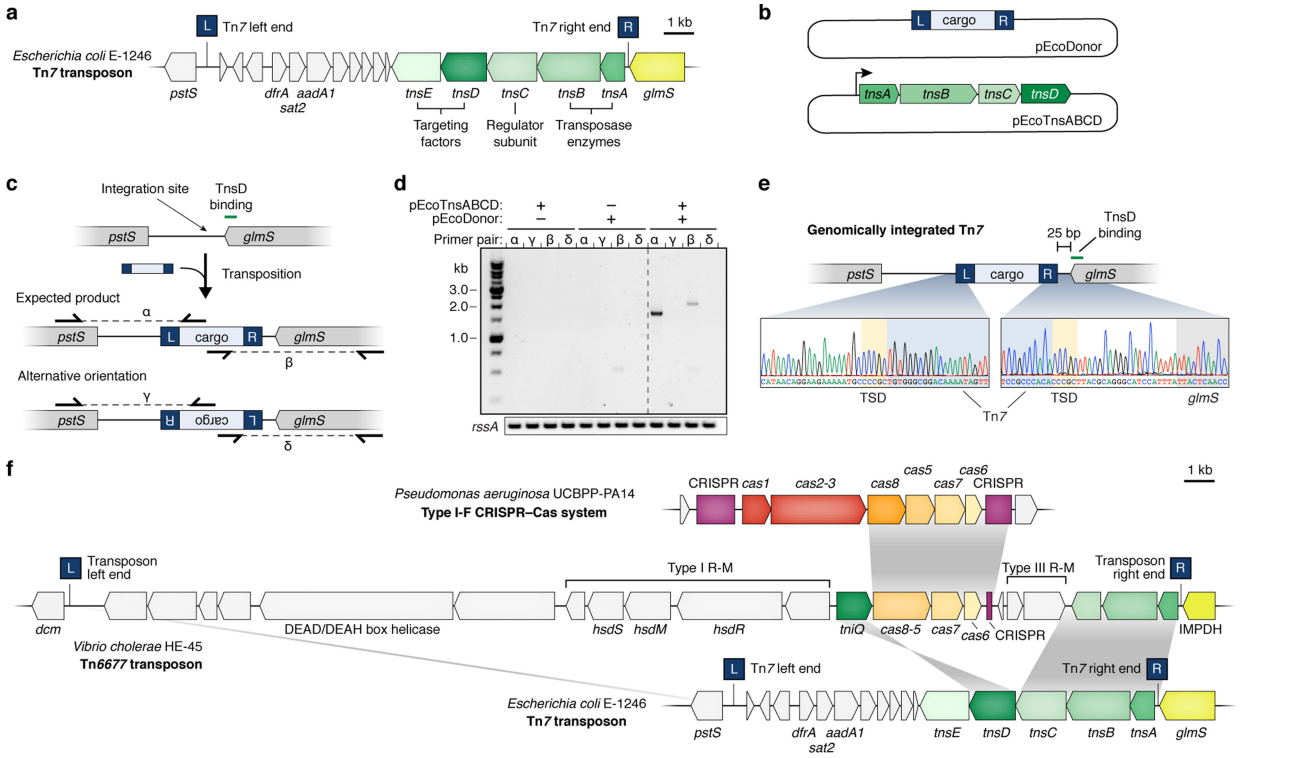
Author contributions S.E.K. and S.H.S. conceived of and designed the project. S.E.K. performed most transposition experiments, generated NGS libraries, and analysed the data. P.L.H.V. helped with cloning and transposition experiments, and performed computational analyses. T.S.H.-H. performed biochemical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

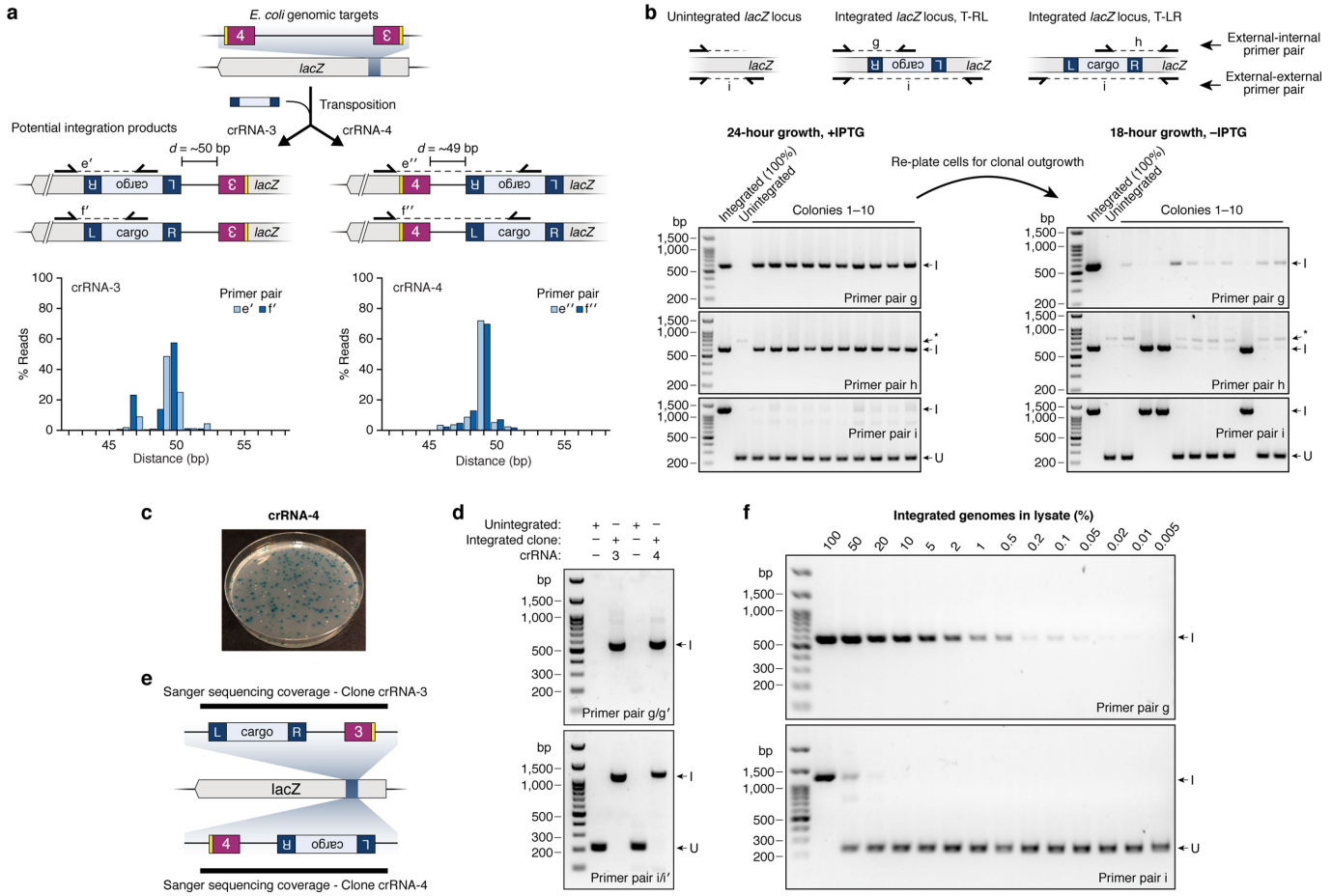
Additional information

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

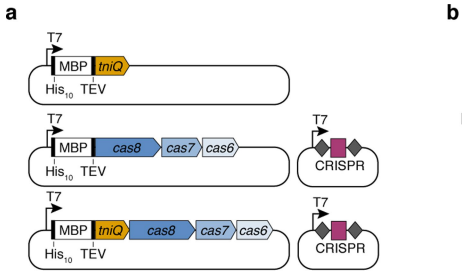
Correspondence and requests for materials should be addressed to S.H.S.

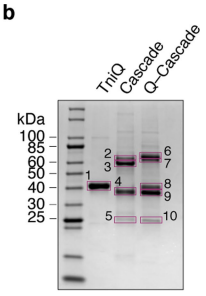


Extended Data Fig. 1 | Transposition of the *E. coli* Tn*7* transposon and genetic architecture of the Tn*6677* transposon from *V. cholerae*. **a**, Genomic organization of the native *E. coli* Tn*7* transposon adjacent to its known attachment site (*attTn7*) within the *glmS* gene. **b**, Expression 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.

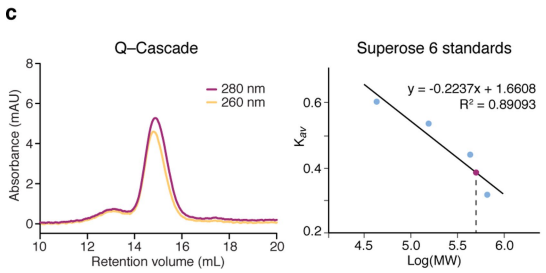


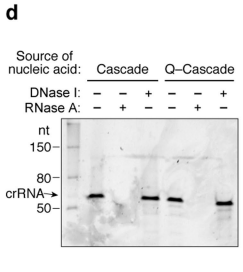
Extended Data Fig. 2 | Analysis of *E. coli* cultures and strain isolates containing *lacZ*-integrated transposons. a, Top, genomic locus targeted by crRNA-3 and crRNA-4, including both potential transposition products and the PCR primer pairs to selectively amplify them. Bottom, NGS analysis of the 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.

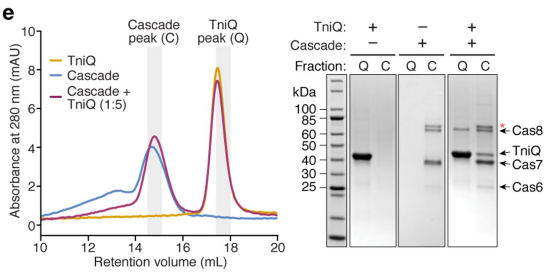




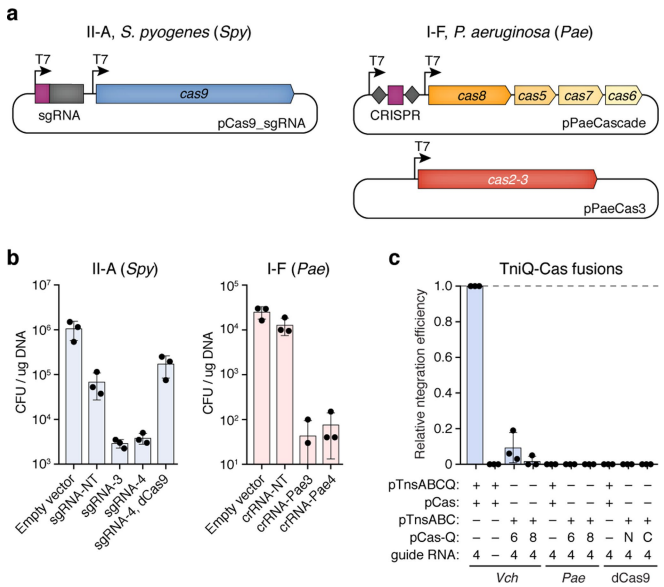
		Spectral counts by gel slice									
Description	MW (kDa)	1	2	3	4	5	6	7	8	9	10
V. cholerae TniQ	46 kDa	513	0	0	0	0	7	15	210	47	41
V. cholerae Cas8	72 kDa	1	194	67	29	20	108	239	30	18	18
V. cholerae Cas7	40 kDa	65	10	8	352	83	13	11	111	460	115
V. cholerae Cas6	23 kDa	2	1	4	5	114	2	0	1	0	158
His10-MBP-TEVsite	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



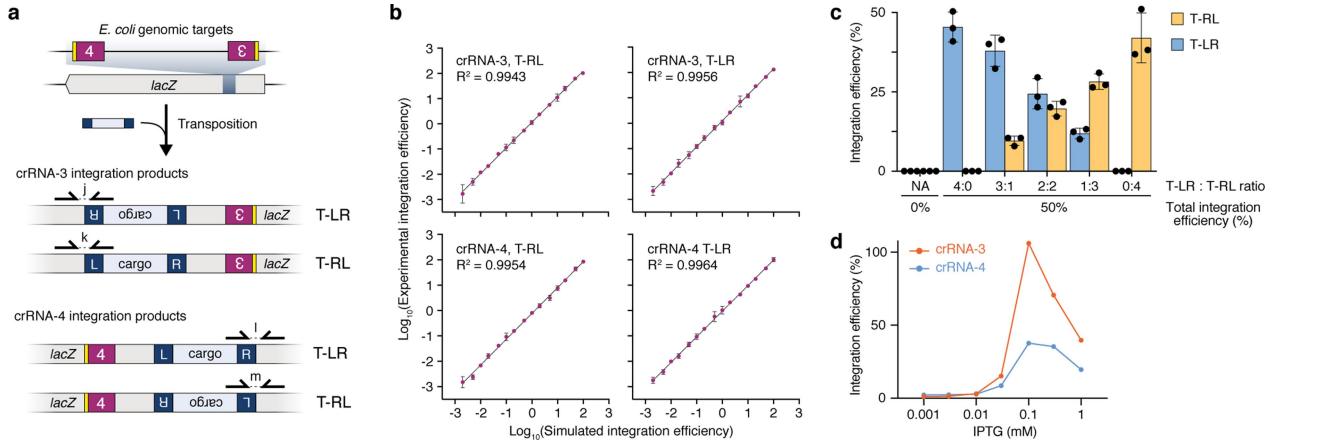




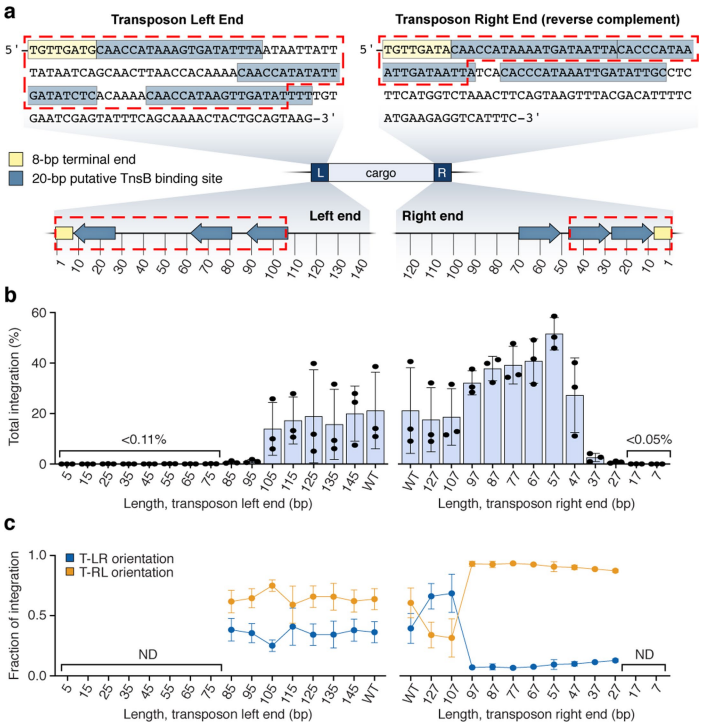
Extended Data Fig. 3 | Analysis of *V. cholerae* Cascade and TniQ–Cascade complexes. a, Expression vectors for recombinant protein or ribonucleoprotein complex purification. **b**, Left, SDS–PAGE analysis of purified TniQ, Cascade and TniQ–Cascade complexes, highlighting protein bands excised for in-gel 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.



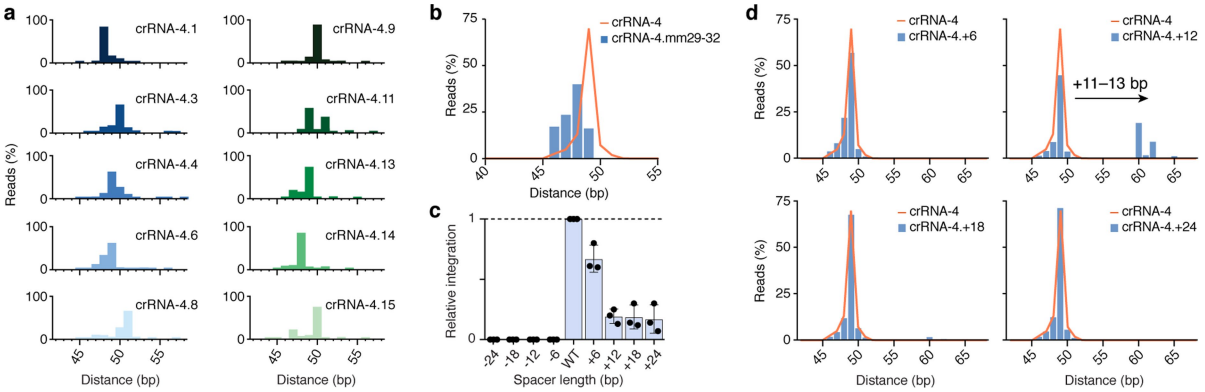
Extended Data Fig. 4 | Control experiments demonstrating efficient DNA targeting with Cas9 and *P. aeruginosa* Cascade. a, Plasmid expression system for *S. pyogenes* (*Spy*) Cas9-sgRNA (type II-A, left) and *P. aeruginosa* Cascade (*Pae*Cascade) and Cas2-3 (type I-F, right). The Cas2-3 expression plasmid 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.



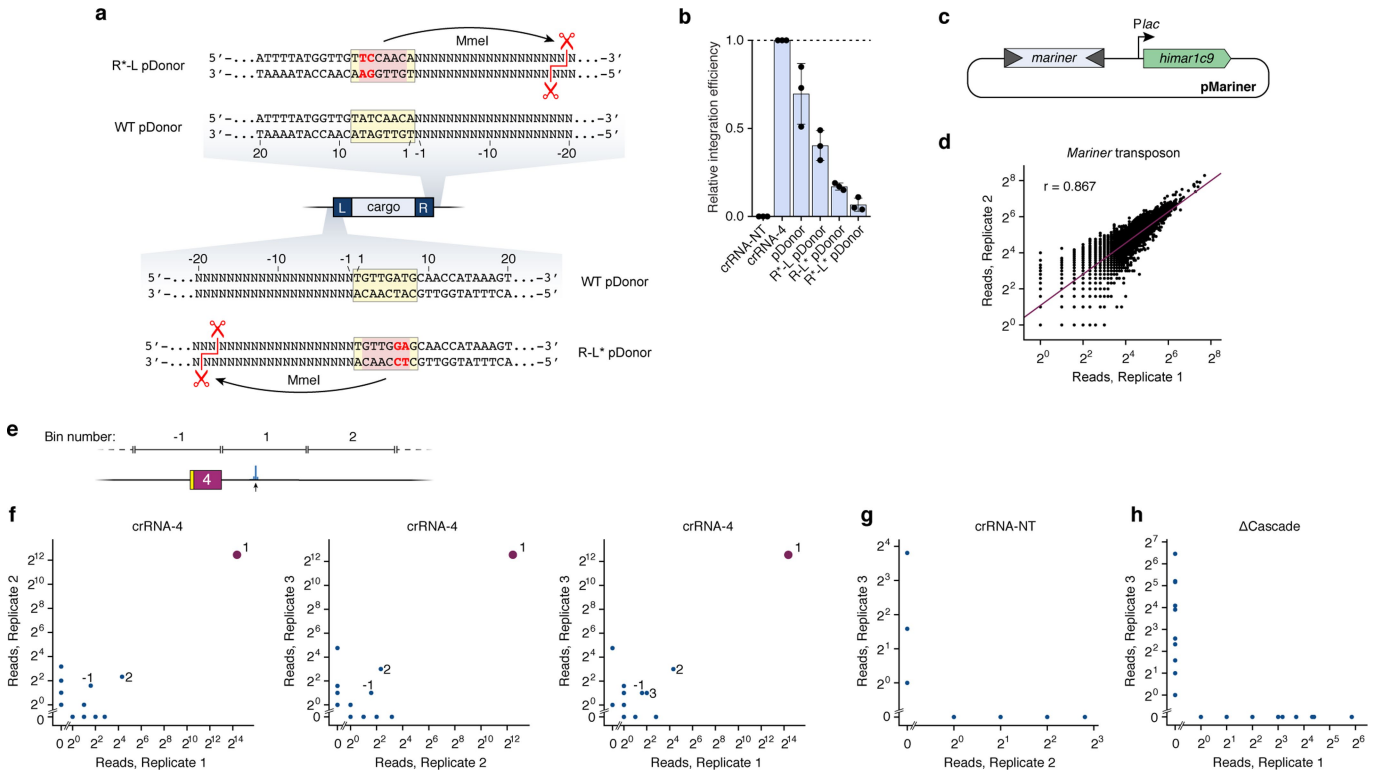
Extended Data Fig. 5 | qPCR-based quantification of RNA-guided DNA integration efficiencies. a, Potential *lacZ* transposition products in either orientation for both crRNA-3 and crRNA-4, and qPCR primer pairs to selectively amplify them. **b**, Comparison of simulated integration efficiencies for T-LR 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.



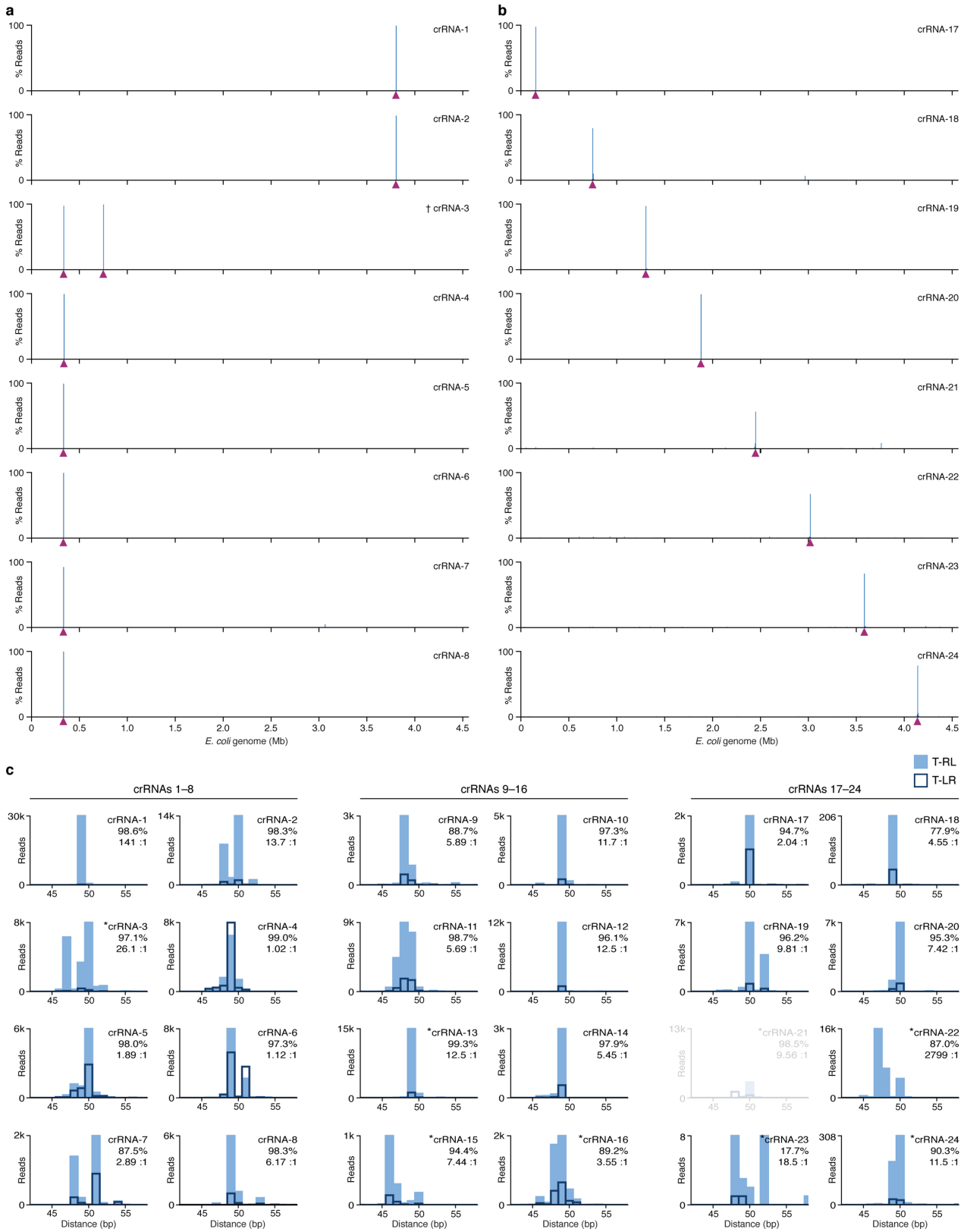
Extended Data Fig. 6 | Influence of transposon end sequences on RNA-guided DNA integration. a, Sequence (top) and schematic (bottom) of *V. cholerae* Tn*6677* left- and right-end sequences. The putative TnsB-binding sites (blue) were determined based on sequence similarity to the TnsBbinding sites previously 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.

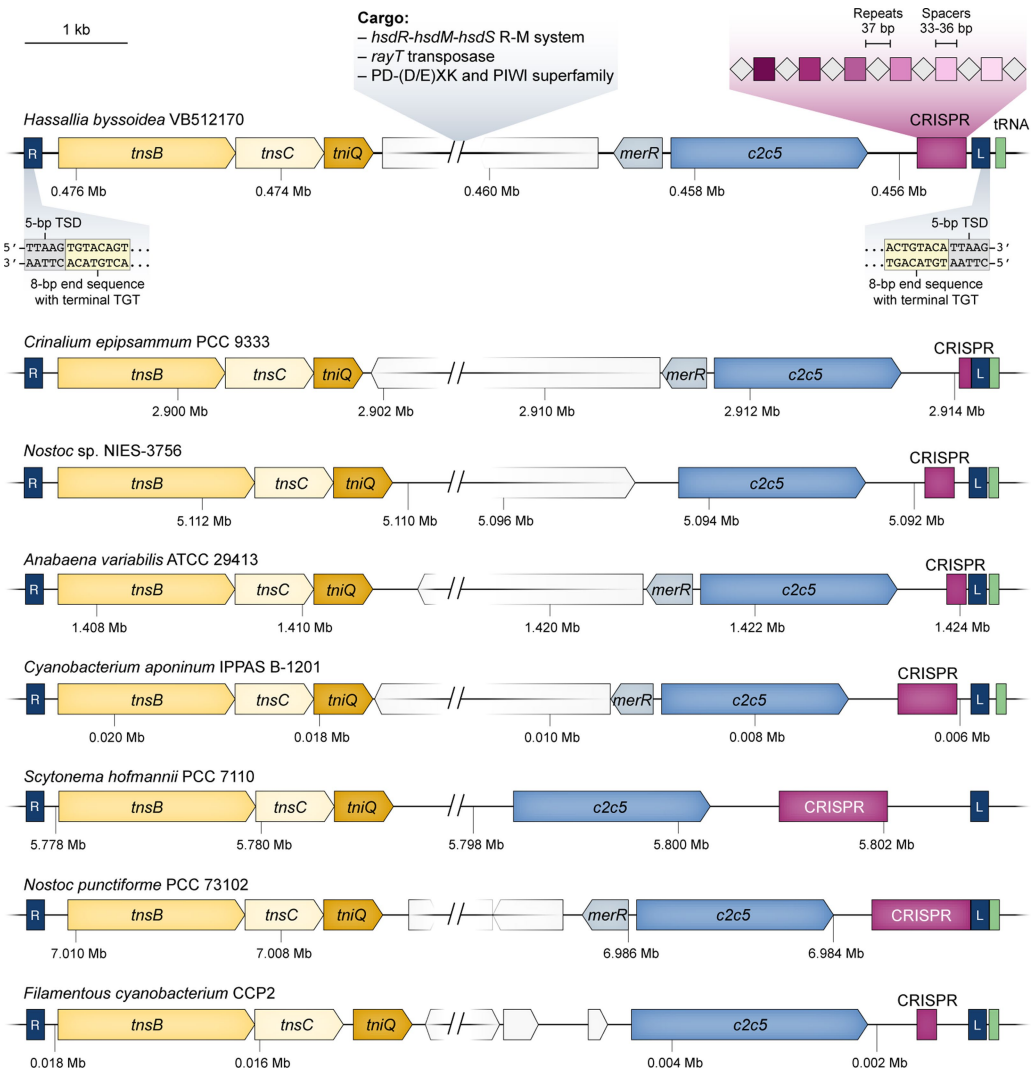


Extended Data Fig. 8 | Development and analysis of Tn-seq. **a**, Schematic of the *V. cholerae* transposon end sequences. The 8-bp terminal sequence of the transposon is boxed and highlighted in light yellow. Mutations generated to introduce MmeI recognition sites are shown in red letters, and the resulting 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.

nature research

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