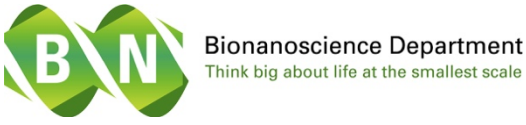
A CRISPR-associated transposase presents null cargo integration efficiency when targeting a transcriptionally highly active region

By
Rodrigo González Linares
Student number: 4770587

In fulfillment of the requirements for the course
Research Project (NB5900)
30-09-2019 to 17-07-2020

Supervisors:
Dr. Ana Rita Costa
Dr. Cristóbal Almendros Romero
Dr. Stan J. J. Brouns

Thesis committee:
Dr. Stan J. J. Brouns
Dr. Chirlmin Joo
Dr. Christophe Danelon



Abstract

CRISPR-Cas effectors (e.g. Cas9) have been widely used to perform genetic knock-outs. Performing knock-ins however, remains challenging due to the inefficiency of the endogenous pathway cells use to integrate a donor genetic cargo into its genome (homology directed repair) when compared to other repair pathways like non-homologous end joining. CRISPR-associated transposases are complexes formed by a catalysis-deficient effector and a transposase. These complexes are able to sequester a transposon, localize a genomic target specified by a CRISPR RNA (crRNA), and integrate the transposon near the targeted site; thereby bypassing homology directed repair. In this study we aimed at developing a screening method using a CRISPR-associated transposase known as CAST, to detect integration events based on the disruption of *lacZ*. During the development, we found that CAST is unable to integrate a cargo in this highly active gene, most likely due to RNA polymerase-mediated dislodgment of the complex, and physical impediment for transposition proteins to reach the target DNA.

Table of contents

1. Introduction

1.1. The CRISPR-Cas system

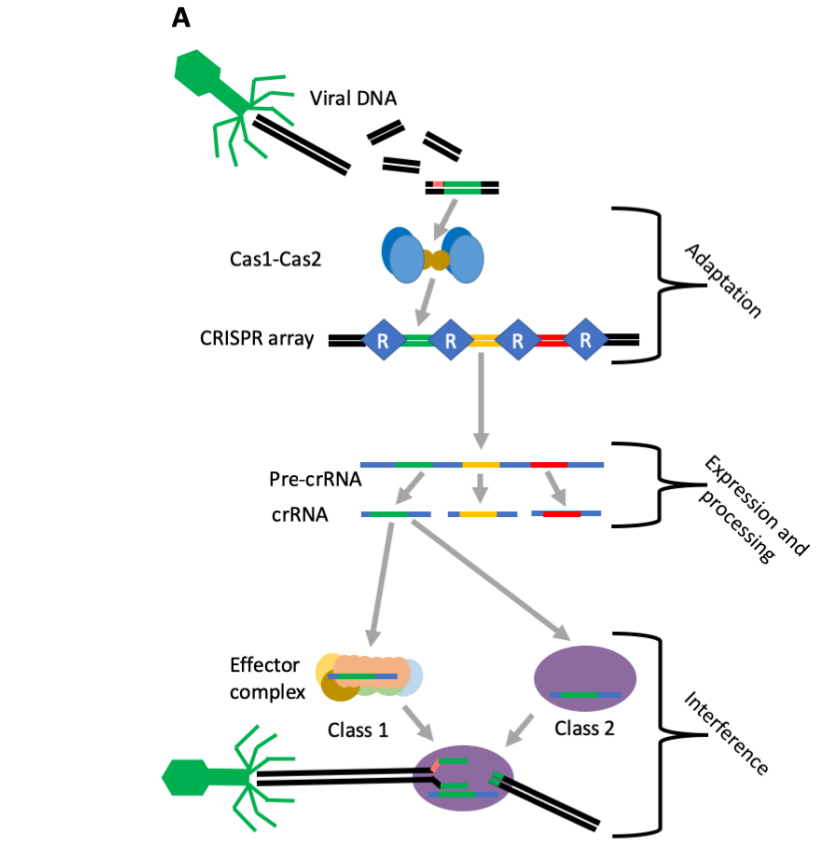
The seminal papers of Francisco Mojica and collaborators that described repeating palindromic sequences interspersed by fragments of foreign origin (Mojica, Juez, & Rodríguez-Valera, 1993) and the later recognition of their involvement in an adaptive immune system (Mojica, Díez-Villaseñor, García-Martínez, & Soria, 2005) revolutionized the understanding of prokaryotic defense systems against mobile genetic elements (MGEs). The viral fragments located in those clusters of regularly interspersed short palindromic repeats (CRISPR) and the CRISPR associated (Cas) proteins form the base of these antiviral defense systems. CRISPR-Cas systems evolved in the context of a continuous arms race between prokaryotes and the viruses infecting them (so-called bacteriophages). This viral pressure, together with contributions of other MGEs such as plasmids, gradually shaped archaic components into primitive adaptive defense systems, which finally diversified into a rich set of unique CRISPR-Cas systems (Koonin, Makarova, & Zhang, 2017). The high abundance of CRISPR-Cas systems in prokaryotic genomes (Westra, Dowling, Broniewski, & Houte, 2016) suggests an essential contribution to an increased host fitness in natural environments.

A CRISPR locus in a prokaryotic genome is composed of a CRISPR array and a *cas* gene operon. The CRISPR array contains two basic elements: spacers and repeats. Spacers are sequences of foreign origin (the memory of the CRISPR system), which are interspaced by identical repeat units. The *cas* operon contains all the genes coding for Cas proteins that form the machinery for spacer acquisition, processing of CRISPR array transcripts and interference against MGEs (Jiang & Doudna, 2017).

CRISPR immunity is conferred via three stages: I) adaptation, II) expression and processing, and III) interference (figure 1A). During the adaptation stage, parts of the invading genetic material are captured and integrated into the CRISPR array as a new spacer by the Cas1-Cas2 complex (for detailed review of CRISPR adaptation see Jackson et al., 2017). Functional spacers are derived from protospacers (via intermediates called prespacers) flanked by a protospacer adjacent motif (PAM) in DNA-targeting systems or a PAM analog in RNA-targeting systems (e.g. an RNA-based PAM [rPAM] for type III systems and a protospacer-flanking sequence [PFS] for type VI systems). A PAM, like its analogues, is a short CRISPR system specific nucleotide sequence that ensures the targeting of foreign invaders rather than the genomic CRISPR locus. To acquire spacers that are functional during the interference stage, the integration event has to occur in a specific, PAM-compliant orientation. In the expression and processing stage, the CRISPR array serves as a template to synthetize a long precursor CRISPR RNA (crRNA) that is further cleaved into smaller mature crRNAs. During the interference stage, these crRNAs are subsequently loaded into the CRISPR effector complex. The loaded complex patrols the cell, screening for complementary sequences that are flanked by a PAM. Upon recognition, the foreign genetic material is cleaved, which effectively aborts the infection (Mohanraju et al., 2016).

CRISPR-Cas systems are divided into two classes, several types and multiple subtypes (Makarova et al., 2020). Class 1 systems include types I, III and IV, and are characterized by the presence of

a multi-subunit effector complex that is involved in the recognition of invader nucleic acids during the interference stage. Class 2 systems include types II, V and VI and have a single multi-domain effector protein responsible for recognition and cleavage of the foreign sequence (figure 1A). Each CRISPR-Cas type has a so-called signature protein and a corresponding signature gene, which characterizes the type (figure 1B) and is present in the corresponding subtypes, even if these subtypes differ in other proteins they code for. Significant for this thesis are CRISPR-Cas types I and V. Cas3 is the signature protein of type I systems. This nuclease is responsible for target DNA cleavage and, depending on the system subtype, is either directly part of the effector complex known as CRISPR-associated complex for antiviral defense (Cascade), or it associates with Cascade after it has located its target (Van Der Oost, Westra, Jackson, & Wiedenheft, 2014). For type V systems, the signature protein is Cas12 (Makarova et al., 2015). Cas12 and a crRNA form an effector complex which both patrols the cell searching for complementary sequences and cleaves that sequence once it is located.



B

	Class I								
	Cas3	Cas8, Cas10 or Csf1	Cas11	Cas7	Cas5	Cas6	Cas1	Cas2	Cas4
Type I	✓	✓ (Cas8)	✓	✓	✓	✓	✓	✓	✓/✗
Type III	✗	**✓ (Cas10)**	✓	✓	✓	✓/✗	✓	✓	✗
Type IV	✗	**✓ (Csf1)**	✓/✗	✓	✓	✓/✗	✓/✗	✓/✗	✗

	Class II			
	Cas9, Cas12 or Cas13	Cas1	Cas2	Cas4
Type II	**✓ (Cas9)**	✓	✓	✓/✗
Type V	**✓ (Cas12)**	✓/✗	✓/✗	✓/✗
Type VI	**✓ (Cas13)**	✓/✗	✓/✗	✗

Figure 1. The CRISPR-Cas adaptive immune system.
A. Three stages of CRISPR immunity. During adaptation, Cas1-Cas2 sequesters PAM-compliant (PAM is depicted in pale red) viral genetic material and integrates it into the CRISPR array in between repeats (R). In the expression and processing stage a pre-crRNA is transcribed from the array and processed into mature crRNAs. The crRNA is then loaded into a Cas effector nuclease and directed towards a complementary sequence which is then cleaved, providing immunity against the bacteriophage in the interference stage. **B.** Generic class I and class II *cas* operons and the presence (✓), absence (✗) or subtype specific presence or absence (✓/✗) of *cas* genes in different types of systems (Makarova et al., 2020). Genes coding for proteins that form part of the adaptation complex are highlighted in red, the gene coding for the protein involved in the processing of pre-crRNAs is highlighted

in green, and genes coding for proteins that form the effector complex are highlighted in blue. In class II systems, the pre-crRNAs are either processed by RNase III (a non-Cas nuclease; Deltcheva et al., 2011) or by the Cas effector protein (East-seletsky et al., 2016; Fonfara, Richter, Bratovič, Le Rhun, & Charpentier, 2016). The signature gene of each CRISPR type is highlighted by a red tick.

1.2. Genome engineering using Cas effector complexes

After recognition of Cas9 as a nuclease directly involved in the prokaryotic immune response (Bolotin, Quinquis, Sorokin, & Ehrlich, 2005), Marraffini *et al.* suggested its use as a tool to cleave user defined DNA sequences (Marraffini & Sontheimer, 2008). Feng Zhang and colleagues later proved that the technology could be harnessed to genetically engineer mammalian cells (Cong et al., 2013).

Cas nucleases (most prominently Cas9 and Cas12) have now been extensively used to induce double-strand breaks (DSB) in desired loci by resorting to user-defined crRNAs, where gene knock-outs (inactivation of an existing genetic element) or knock-ins (integration of exogenous DNA) can be performed. This technology has proven to be versatile and efficient to knock-out genes, especially when compared to techniques based on zinc-finger nucleases (ZFN; Durai et al., 2005) and transcription activator-like effector nucleases (TALEN; Joung & Sander, 2013), but it still remains inefficient for knock-ins.

Knock-ins rely on an endogenous DNA repair system called homology directed repair (HDR) to integrate the foreign genetic material. In this process, the DNA fragment to be knocked-in is flanked by homology arms, which are sequences homologous to those adjacent to the cleavage site. The HDR cell machinery recognizes these homologous arms and inserts the genetic cargo into the cleavage site (Jiang & Doudna, 2017). The caveat is that homology directed repair competes with the nonhomologous end joining (NHEJ) repair system, in which substitutions or small indels (short for insertion or deletion) are induced at the DSB site (Jiang & Doudna, 2017). Because this competition is heavily skewed in favor of NHEJ (figure 2), and the homology directed repair is only active in dividing cells during the S and G phases of the cell cycle (while non-homologous end joining is always active; Jiang & Doudna, 2017), the knock-in efficiency generally does “not exceed several tenths of a percent, or several percent” (Rozov, Permyakova, & Deineko, 2019).

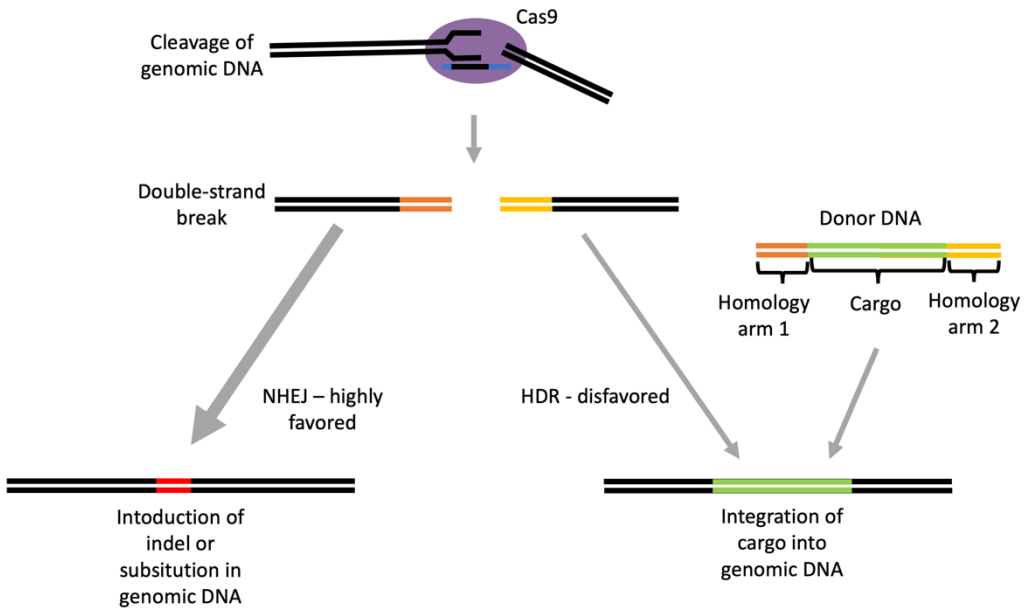


Figure 2. Exploitation of HDR for genetic engineering and competition with NHEJ.
After a DSB is induced in the genomic DNA with the aid of a Cas nuclease, the cellular machinery repairs the damage mainly using one of two pathways: NHEJ or HDR. NHEJ is favored over HDR (as depicted by the use of a thick and a thin arrow, respectively). In the case of NHEJ, a small indel or substitution is introduced at the DSB site to repair the damage. On the other hand, when a donor DNA (composed of a cargo and a couple of arms homologous to the sequences adjacent to the DSB) is provided, the cell might integrate the donor DNA to repair the damage by HDR.

The knock-in efficiency also heavily depends on the size of the cargo. Li *et al.* tested this in mouse embryonic stem cells, observing a knock-in efficiency of 36.3% when using a cargo size of 99 bp, and an efficiency of just 4.3% for a cargo of 720 bp (K. Li, Wang, Andersen, Zhou, & Pu, 2014). When selection markers or enrichment cannot be used for the selection of mutants, a Cas9-based method to knock-in DNA fragments in the order of kilobases becomes implausible or even impossible due to the very limited number of successful integration events.

1.3. Tn7 transposons

DNA transposons (hereafter transposons or transposable elements) are a type of MGE consisting of DNA sequences that can move between genomic loci and even plasmids, without homology constraints. The transposon accomplishes the migration process by means of a self-encoded transposase, which excises the transposon from one locus and integrates it into another (Kazazian Jr., 2004; Peters & Craig, 2001).

The limits of a transposon are defined by two sequences, here referred to as left end (LE) and right end (RE) segments. Any sequence laying between the limits of such segments is part of the transposon (Peters & Craig, 2001).

Of particular importance for the present manuscript is the bacterial Tn7 transposon. The Tn7 transposon is composed of non-identical ~150 bp LE and ~90 bp RE segments. The LE segment contains three transposase-binding sites, while the RE segment contains four, each spanning 22

bp (Peters & Craig, 2001). The Tn7 transposon also contains antibiotic resistance cassettes with genes *dhfrI* (trimethoprim resistance) and *aadA* (streptomycin and spectinomycin resistance), and a defective recombinase (Sundstrom & Sköld, 1990). The essential components are, however, five genes coding for transposition proteins: *tnsA, tnsB, tnsC, tniQ* (also known as *tnsD)* and *tnsE*. Proteins TnsA and TnsB form the TnsAB transposase. In particular TnsA performs the 5' breaks on both transposon ends, while TnsB acts on the 3' ends. Additionally, TnsB has the ability to join the 3' ends of the transposon with the target DNA. TnsC serves as an ATP-dependent transposase activator. It is recruited by TniQ or TnsE to the DNA, inducing the formation of a platform in the minor groove capable of receiving TnsAB, which can then integrate the transposon. TniQ and TnsE are target selectors. When TnsABC interacts with TniQ, the transposition is directed towards an *att*Tn7 sequence, which is highly conserved in bacterial genomes. On the other hand, when the TnsABC complex interacts with TnsE, the transposition to conjugal plasmids is highly favored. Because TnsE interacts with the lagging-strand during replication, genomic transpositions at low frequencies are also observed (Peters & Craig, 2001).

1.4. CRISPR-Cas and transposable elements

Since their discovery, CRISPR-Cas systems have been intimately associated with transposable elements. For example, *cas9* and *cas12* are believed to have evolved from *tnpB* genes present in IS605-like transposons, and *cas1* is thought to originate from a specific type of transposon; a casposon (Faure, Scott, & Peters, 2019). Notably, the type I-E system in *Streptomycetaceae* lacks the adaptation genes *cas1* and *cas2*, and it instead contains the *tnsB* and *tnsC* genes, which normally form part of the molecular machinery required for Tn7-like transposons to accomplish transposition. It has been hypothesized that these two genes could form a novel adaptation module (Faure et al., 2019).

This relationship seems to go both ways, as Tn7-like transposons have captured type I-F and type I-B CRISPR-Cas systems. A type V-K (V-U5) system was also found to be part of a Tn7-like transposon, with all known analogs linked to a transposon (Faure et al., 2019). Notably all the effector nucleases in all these systems are catalysis-deficient, while retaining their ability to recognize and bind target sequences. In type I systems, the Cas3 nuclease is not present, while in the type V-K system accumulation of point mutations in Cas12K (c2c5) striped it from its nuclease activity (Faure et al., 2019).

In particular, two of these systems have been reconstituted and partially characterized in *E. coli*: the type I-F from *Vibrio cholerae* (Klompe, Vo, Halpin-healy, & Sternberg, 2019; figure 3A) and the type V-K from *Scytonema hofmanni* (Strecker, Ladha, Gardner, Schmid-burgk, & Kira, 2019; figure 3B).

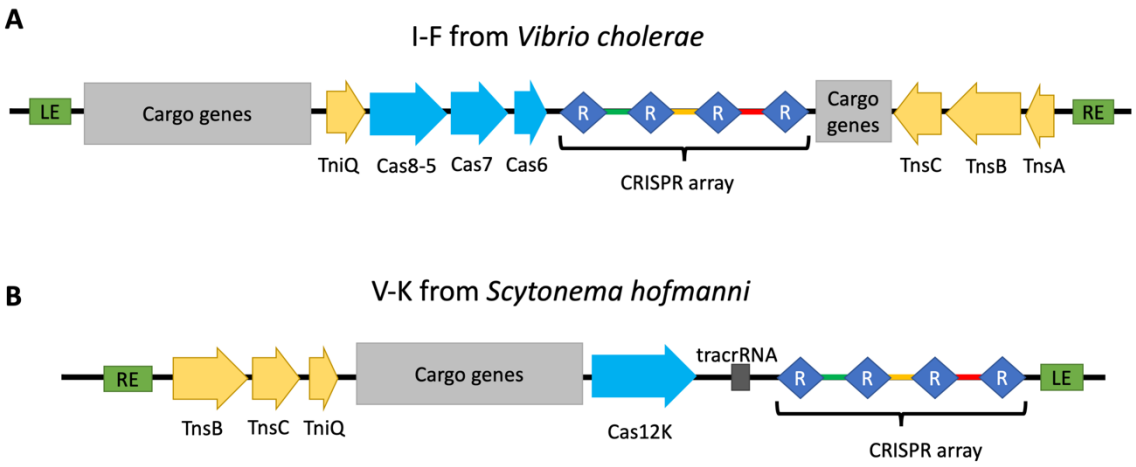


Figure 3. CRISPR-associated transposase loci (5'→3').
A. Tn7-like transposon harboring an I-F CRISPR-Cas system from *Vibrio cholerae* and **B.** Tn7-like transposon harboring a V-K CRISPR-Cas system from *Scytonema hofmanni*. Transposition genes are highlighted in yellow and CRISPR-Cas effector genes are highlighted in blue.

Both systems were shown to be able to integrate transposons near the site targeted by Cascade or Cas12K, respectively (Klompe et al., 2019; Strecker et al., 2019). However, the proposed models for transposition are different.

For the type I-F system of *Vibrio cholerae*, TniQ directly binds the Cascade complex (composed of Cas6, Cas7 and Cas8-Cas5), while TnsAB independently sequesters the transposon. Once the Cascade-TniQ complex has located its target, TnsC is recruited by Cascade, making it possible for TnsAB to perform the transposition (Klompe et al., 2019) (figure 4A). The Cascade-TniQ complex was reconstructed by cryo-electron microscopy (cryo-EM), showing that a TniQ dimer indeed binds Cascade (Halpin-healy, Klompe, Sternberg, & Fernández, 2020).

The model for the type V-K system of *Scytonema hofmanni* is different, as it is proposed that Cas12K, TnsB, TnsC and TniQ directly form a complex referred to as CAST (standing for CRISPR-associated transposase). Within the CAST complex, the transposition proteins first sequester a transposon, Cas12K then locates a target, and finally the transposition proteins integrate the transposon (Strecker et al., 2019; figure 4B). At the moment of writing of this manuscript, no structural reconstruction of the CAST complex has been published.

The V-K system contains a trans-activating CRISPR RNA (tracrRNA); an RNA molecule required by some effector complexes (including Cas12K; Strecker et al., 2019) that forms a duplex with the crRNA (Deltcheva et al., 2011).

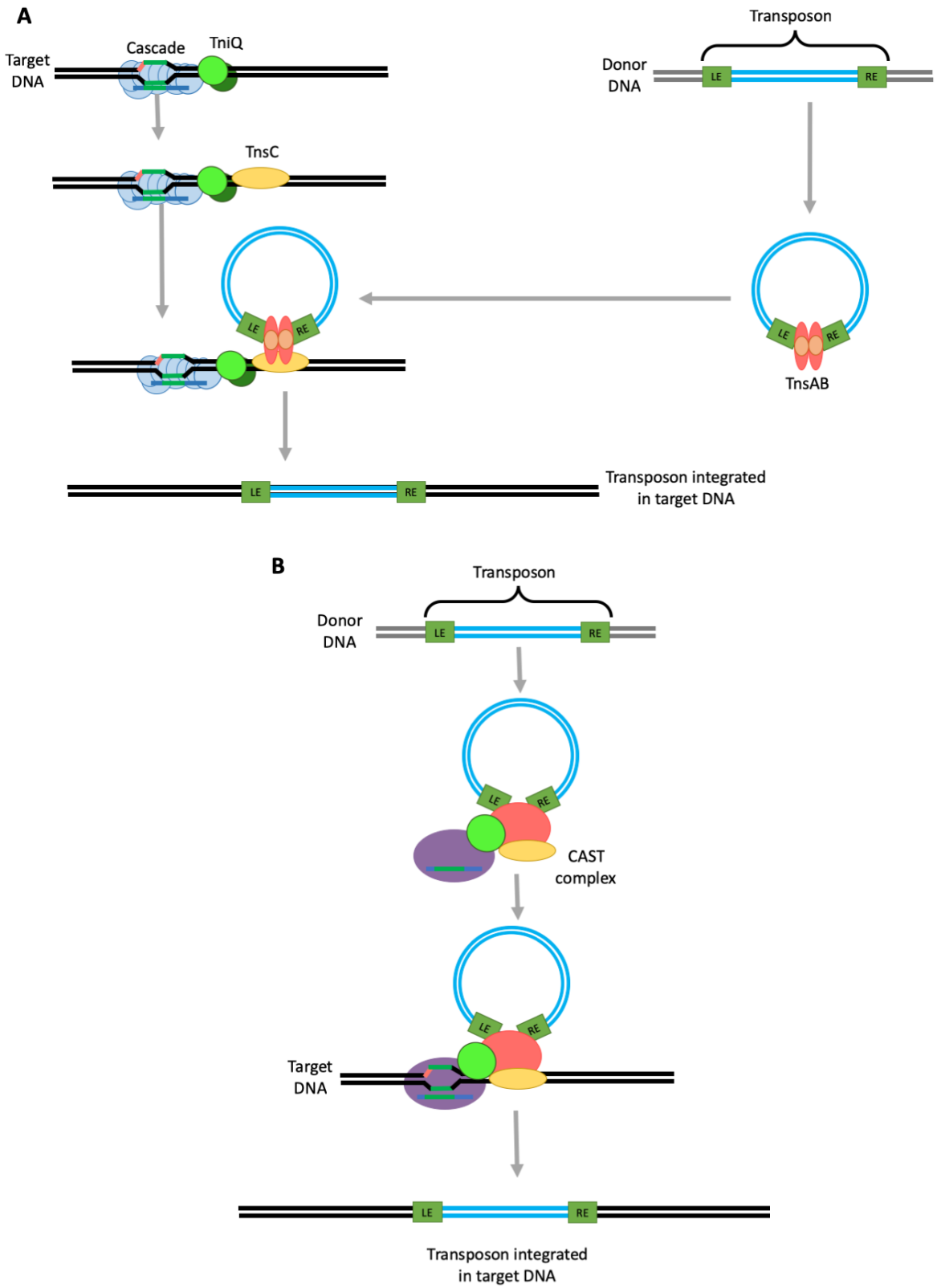


Figure 4. Hypothetical transposition models for the type I-F and type V-K systems.
A. Hypothetical transposition model for the type I-F system. TniQ-Cascade binds the PAM-compliant (pale red) loci specified by the crRNA in the target DNA, after which TnsC is recruited. TnsAB, which sequestered a transposon located in the donor DNA, is

then recruited by TnsC and the transposon is integrated into the target DNA, near the TniQ-Cascade binding site. **B.** Hypothetical transposition model for the type V-K system. The CAST complex sequesters a transposon located in the donor DNA, after which it binds the PAM-compliant loci specified by the crRNA in the target DNA. Next, the transposon is integrated into the target DNA, near the Cas12K binding site.

Looking at the transposition proteins of the type V-K system, it is surprising to see the absence of TnsA (figure 3), as TnsA is one of the two proteins forming the transposase and the protein responsible for inducing 5' breaks in the canonical Tn7 transposon (Klompe et al., 2019; Strecker et al., 2019). Recently, this absence was also noted by Rice *et al.* (Rice, Craig, & Dyda, 2020). Because of this, these authors pointed out that CAST might not be able to perform simple integrations, but rather co-integrations, i.e., the integration of a whole transposon-carrying plasmid (called pDonor in Strecker *et al.* assays) rather than the transposon alone (Rice et al., 2020). The process works as follows: TnsB 3'-nicks the transposon ends and joins them with the target DNA (as previously explained in section 1.3.), but without the 5' break induced by TnsB, one strand of the transposon is joined to the transposon-carrying plasmid while the other is joined to the target DNA, forming a so-called Shapiro intermediate. After replication, the transposon-carrying plasmid and the target DNA are fused, and the transposon is duplicated. Co-integrations can be separated again into independent molecules in a process called resolution, which is driven either by homologous recombination or by transposon-encoded systems. In this case, the final product after resolution would be a plasmid identical to the original transposon-carrying plasmid and a target DNA carrying a transposon (Hickman & Dyda, 2016). However, in the canonical Tn7 system, when TnsA is knocked-out or mutated to be non-functional, the overwhelming majority of end products are unresolved co-integrations (>97% and 82%, respectively; May & Craig, 1996). In response to Rice *et al.*, Strecker *et al.* presented the results of an assay using CAST in which 19.6% of all integrations were co-integrations. Additionally the precise structure of some of these co-integrations was studied by nanopore sequencing. It remains unclear if the co-integrations were resolved into simple integrations by the CAST complex itself, or by an endogenous system of *E. coli*. Strecker *et al.* also proposed that providing a 5' nicked pDonor would prevent co-integration (Strecker, Ladha, Makarova, Koonin, & Zhang, 2020).

The PAM preference, integration directionality and efficiency of both the I-F and the V-K systems have been characterized. The analysis of the type I-F system revealed a CC canonical PAM. The directionality of the integration events was highly dependent on the particular crRNA tested. For example, a particular crRNA yielded integrations in the 5'-RE→LE-3' close to 100% of the times, while other yielded integrations in each direction about half of the times. Integrations occurred 46-55 bp downstream of the PAM. The integration efficiency was shown to be dependent on the transposon size, and seems to follow what resembles a normal distribution. When using a crRNA targeting *lacZ*, a transposon of 0.78 kb yielded the maximal integration efficiency at around 25%. For this system, several crRNAs targeting *lacZ*, *glmS* and eight other arbitrary locations where tested (Klompe et al., 2019).

On the other hand, the type V-K has a NGTN canonical PAM. In this case, integration events were detected in 60.4% of the targeted loci and only in the in 5'-LE→RE-3' direction. The integrations were observed to occur 60-66 bp downstream of the PAM. The integration efficiencies were higher and less strongly dependent on the size of the transposon than those of the type I-F

system. For example, when testing a particular crRNA, the integration efficiency of a 0.5 kb transposon was around 70%, while one of 10 kb was around 45%. It is important to note that all results reported in the study regarding genomic integrations were produced using different crRNAs targeting non-coding regions (Strecker et al., 2019), while many crRNAs targeting coding regions were used for the characterization of the type I-F system (Klompe et al., 2019).

Off-target integrations seem to be rare in the type I-F system (Klompe et al., 2019), and common (47.1% to 51.0%) in the type V-K system due to Cas12K-independent integration events (Strecker et al., 2019).

It is notable that CRISPR-associated transposases have not only been found in nature, but they have also been engineered. Such is the case of a fusion protein composed of the *Hsmar1* transposase derived from the mariner transposon and a dead (catalysis-deficient) Cas9 (dCas9). The system was shown to be functional, but integrations at random loci occurred at high rates (Bhatt & Chalmers, 2019).

1.5. CRISPR-associated transposases in the CRISPR-Cas toolbox

When talking about CRISPR-Cas technologies, Cas9 has certainly been at the spotlight. Some variants of this programmable nuclease have been engineered, for example, to increase its targeting specificity and to perform alternative tasks, some of which extend well beyond the field of genetic engineering (e.g. to regulate transcription, modify the epigenome, image chromatin and modify its topology, among others; Adli, 2018).

Recently, however, two new tools have been incorporated in the CRISPR-Cas toolbox that might outcompete Cas9 for some applications. Table 1 shows how these tools compare to each other.

The first of these new tools are CRISPR-associated transposases, which might replace Cas9 when the integration of large DNA fragments into the genome of an organism is required. The advantage of these systems is that they bypass the inefficient HDR process by performing the integration using their own transposase (Klompe et al., 2019; Strecker et al., 2019). Applications harnessing this technology are starting to emerge. Multicopy chromosomal integration using CRISPR-associated transposases (MUCICAT) uses the type I-F system to control the copy number of gene expression cassettes in a bacterial genome at different loci, rather than on unstable plasmids, as traditionally done. When integration of the desired number of copies is achieved, the donor plasmid is curated to halt further integrations (Zhang et al., 2020). This control is critical to create efficient biocatalyst strains (Zhang et al., 2020).

The second new tool is prime editing, in which a so-called prime editor was engineered by fusing a Cas9 nickase (a version of Cas9 that performs a single-strand break or nick) with a reverse transcriptase domain. The system works by using a prime editing guide RNA (pegRNA; an extended crRNA whose overhang partially primes with the nicked DNA strand) to produce small indels or substitutions after reverse transcription (for a detailed description of the approach see Anzalone et al., 2019). Although the prime editor is based on Cas9, the editing paradigm is completely different from that of traditional Cas9-based methods.

Table 1. Most relevant tools available in the genetic engineering CRISPR toolbox and their characteristics. Cells highlighted in green represent the best tool for each particular application.

	Cas9	CRISPR-associated transposases	Prime editing
Small knock-ins (10^1 bp order)	Yes	Yes	No
Medium size knock-ins (10^2 bp order)	Yes, but inefficient. Can be used with selection markers	Yes	No
Big knock-ins ($\geq 10^3$ bp order)	No, too inefficient	Yes	No
Knock-outs	Yes	Yes, but impractical. E.g. inserting sequence in the middle of a gene	Yes, but impractical. E.g. introducing early stop codon
Single-nucleotide (or few-nucleotides) substitutions	Yes, but impractical. Requires homology directed repair	No	Yes
Is the method scar-less?	Yes, although removing selection markers leaves scar	No, transposon left and right end sites are integrated. Small insertion site is duplicated.	Yes

1.6. Project aim

The project aims at developing a target integration system based on CAST, which is efficient, easy to modify, and allows for the straightforward identification of successful integrations.

CAST was preferred to Cascade because of its apparently superior integration efficiency, especially for fragments in the order of 10^2 and 10^3 bp.

2. Material and Methods

2.1. Bacterial strains and growth conditions

E. coli DH5α was used for plasmid transformation and production, while *E. coli* K-12 MG1655 was used to perform the integration assays. Strains were grown at 37 ºC, except for the production and maintenance of pKD46 and thermosensitive new pDonor V2 (pTU396; see supplementary table 1 in section 8.1 for a full list of plasmids used or created for this project), which were grown at 30 ºC. Ampicillin (AMP; final concentration of 100 µg/ml), chloramphenicol (CHL; final concentration of 30 µg/ml), and kanamycin (KAN; final concentration of 50 µg/ml) were used in the assays.

2.2. Transformation

2.2.1. Chemically competent cells

E. coli DH5α cells were made chemically competent by inoculating them in 10 mL of SOB medium and growing them at 30 ºC under agitation until an optical density at 600 nm (OD_{600}) of 0.3-0.4 was reached. Cells were then pelleted by centrifuging at 3000 x *g* and 4 ºC for 10 min, after which the supernatant was decanted. An additional 1 min centrifugation under the same conditions was carried out to remove additional medium with a pipette. The pellet was resuspended in 10 mL of 1x wash buffer (Zymo Research), and once again centrifuged. The final cell pellet was resuspended in 5 ml of 1x competent buffer (Zymo Research) and 100 µL aliquots were prepared in pre-cooled Eppendorf tubes. The tubes were snap-frozen in liquid nitrogen and stored at -80 ºC.

The chemically competent *E. coli* DH5α cells were transformed with 2 µL of plasmid, by incubating on ice for 30 min, heat-shocking at 42 ºC for 1 min in a water bath, and re-incubating for 2 min on ice, followed by a 37 ºC (or 30 ºC for transforming thermosensitive new pDonor V2 or pKD46) incubation for 1h. Cells were then plated in LB agar containing the appropriate antibiotic at the same temperature.

2.2.2. Electrocompetent cells

E. coli K-12 MG1655 cells were made electrocompetent as follows. The bacteria were inoculated in 20 mL of LB medium and grown at 37 ºC under agitation until an OD_{600} of 0.3-0.5 was reached. Cells were then centrifuged for 4 min at 4 ºC, 3000 x *g*, and the supernatant was discarded. Cells were resuspended in 20 mL of miliQ water, and again centrifuged. The washing step was repeated two additional times, and the pellet was resuspended in 200 µL of ice-cold 10% glycerol. The samples were distributed in aliquots of 50 µL and used immediately or stored at -80 ºC.

To perform the transformation, 2 µL of plasmid were added to the cells and electroporated at 2500 V, 200 Ω and 25 µF in 2 mm-gap cuvettes. 800 µL of LB were added to the cells and the sample was incubated at 37 ºC (or 30 ºC for transforming thermosensitive new pDonor V2 or pKD46) for 1 h. Cells were then plated in LB agar containing the appropriate antibiotic at the same temperature.

2.3. Polymerase chain reaction

A list of all oligonucleotides (oligos) used in this work can be found in the supplementary table 2 (section 8.2.).

2.3.1. Using Q5 DNA polymerase

The Q5 High-Fidelity DNA polymerase (NEB) was used whenever the products were meant to be used for cloning purposes. The reactions were set up as shown in Table 2 and run according to the program in Table 3.

Table 2. Q5 High-Fidelity DNA polymerase PCR reaction components

Component	Volume per reaction	Final concentration
5X Q5 reaction buffer (NEB)	5 µL	1X
10 mM dNTPs	0.5 µL	200 µM
10 µM forward Primer	1.25 µL	0.5 µM
10 µM reverse Primer	1.25 µL	0.5 µM
Template DNA	Variable	Variable
Q5 High-Fidelity DNA Polymerase (NEB)	0.25 µL	0.02 U/µL (0.01 µL per 1 µL of reaction)
MiliQ water	Up to 25 µL	-

Table 3. Q5 DNA polymerase reaction conditions

Step	Temperature (ºC)	Time
Denaturing	98	5 min
30-35 cycles (denaturing, annealing, extension)	98	30 s
	Variable*	30 s
	72	30 s/kb
Extension	72	5 min
Hold	12	Indefinite

*The annealing step was first performed at the temperature indicated by the NEB Tm calculator (available at https://tmcalculator.neb.com/#!/main) and the temperature was increased if the desired product was not observed after running an agarose gel electrophoresis to increase specificity.

2.3.2. Using *Taq* DNA polymerase

The Quick-Load Taq 2X Master Mix from NEB was used for screening purposes. The reaction was set up as shown in Table 4 and run according to the program in Table 5.

Table 4. Quick-Load Taq 2X Master Mix reaction components

Component	Volume per reaction	Final concentration
10 µM forward Primer	0.5 µL	0.2 µM
10 µM reverse Primer	0.5 µL	0.2 µM
Template DNA	Variable	Variable
Quick-Load *Taq* 2X Master Mix (NEB)	12.5 µL	1x
MiliQ water	Up to 25 µL	-

Table 5. Quick-Load Taq 2X Master Mix reaction conditions.

Step	Temperature (ºC)	Time
Denaturing	95	5 min
30-35 cycles (denaturing, annealing, extension)	95	30 s
	Variable*	30 s
	68	1 min/kb
Extension	68	5 min
Hold	10	Indefinite

* The annealing step was performed at the temperature indicated by the NEB Tm calculator.

2.4. Agarose gel electrophoresis

1.2% agarose gels were prepared in 1x TAE buffer (Promega) and stained with 0.1 µL/mL of Sybr Safe stain (Invitrogen). DNA samples were mixed with purple loading buffer (NEB) or directly introduced into the wells when using the Quick-Load Taq 2X Master Mix. Samples were run at 100 V for 45-90 min and visualized using a ChemiDoc XRS+ (Bio-Rad) gel documentation system.

2.5. DNA purification from agarose gels

DNA excised from an agarose gel was purified using the Zymoclean Gel DNA Recovery Kit (Zymo Research) following the manufacturer's protocol.

2.6. Plasmid purification

Plasmids were purified using the GeneJET plasmid Miniprep kit (Thermo Scientific) following the manufacturer's protocol.

2.7. DNA quantification

DNA was quantified by spectrophotometry using a NanoPhotometer NP80 (Implen).

2.8. Cloning

2.8.1. Restriction enzyme digestion and ligation

Approximately 1 µg of DNA was digested with 2 µL of each High Fidelity NEB restriction enzyme and 5 µL of CutSmart buffer 10x (NEB) in a reaction of 50 µL, at the temperature recommended by NEBcloner (available at http://nebcloner.neb.com/#!/redigest) overnight. Reaction products were run in an agarose gel and bands corresponding to digested DNA where excised and purified as described above.

Insert DNA and vector DNA were ligated at a ratio between 3:1 and 7:1, using 1 µL of T4 DNA Ligase and 2 µL of T4 DNA Ligase Buffer 10x (for a total reaction volume of 20 µL). Ligation reactions were incubated at 16 ºC overnight. The ligation product was then transformed into chemically competent or electrocompetent cells, as described above.

2.8.2. Ligation independent cloning (LIC)

LIC cloning was used to create new pDonor (pTU394). The following LIC sites were used for both vector (pACYCDuet-1 backbone) and insert (synthetic transposon [GeneArt; Thermo Scientific]): 5'-TTTAAGAAGGAGATATAGAT-3' and 5'-ATCCCAACTCCATAA-3'.

The vector was prepared by mixing 100 ng of DNA, 2 µL of 25 µM dCTP, 2 µL of buffer 2.1 (NEB), 1 µL of 100 µM DTT and 0.5 µL of T4 DNA polymerase (NEB) in a total reaction volume of 20 µL. The insert was prepared similarly, but using dGTP instead of dCTP. The reactions were incubated for 30 min at room temperature and then inactivated for 20 min at 75 ºC.

5 µL of each reaction (25 ng of insert and 25 ng of vector; ~2.5:1 molar ratio) were combined into one tube and incubated for 30 min at room temperature. 1-2 µL of the mixture were used to transform chemically competent cells.

2.9. Sanger sequencing

500-800 ng of template DNA and 2 µL of a 10 µM primer where mixed in a total volume of 10 µL. The samples were sent to Macrogen for sequencing via their EZ-seq service. Files containing the sequencing chromatograms were provided. The sequencing chromatograms were aligned with the hypothetical constructs in GenomeCompiler using the Clustal Omega algorithm.

2.10. Preparation of spacers and introduction into new pDonor V2

New pDonor V2 (pTU395; contains transposon and CRISPR array with interchangeable spacers) was digested as described above with BsaI, run in an agarose gel and purified with a Zymoclean Gel DNA Recovery Kit (Zymo Research). Spacers were prepared by phosphorylation of the corresponding two oligos (Supplementary Table 2 in section 8.2.), using 1 µL of each oligo (100 µM), 5 µL of 10x T4 ligase buffer (NEB), 1 µL of T4 PNK (NEB) and 42 µL of miliQ water. The oligos were annealed in a thermocycler at 95 ºC for 5 min, followed by 5 ºC decreases in temperature at intervals of 30 s until a temperature of 70 ºC. Samples were then let to cool down at room temperature and diluted 10 times in MiliQ water. The annealed phosphorylated oligos form a spacer sequence with overhangs matching those of the digested new pDonor V2 vector. For the ligation, 1 µL (~10 ng/µL) of gel-purified digested new pDonor V2 was mixed with 2 µL of diluted annealed oligos, 2 µL of 10x T4 DNA ligase and 14 µL of miliQ water. The samples were incubated at 16 ºC overnight and transformed into chemically competent cells.

Table 6 provides details of the spacers used, as well as the results of their ligation into new pDonor V2.

Table 6. List of spacers designed to be ligated into new pDonor V2, including their name, corresponding PAM, PAM-containing strand, the sequences of the annealed oligos (with the spacer sequence highlighted in bold), which locus they target, and whether the annealed oligos were successfully ligated.

Name	PAM	Strand	Oligo sequences	Locus	Ligated
PSL1	GGTT	+	5'-AAAG**TCCGGCACCAGAAGCGGTGCCGG**-3' 3'-**AGGCCGTGGTCTTCGCCACGGCCA**CCG-5'	*lacZ*	Yes
PSL2	GGTT	+	5'-AAAG**TTCTCCGGCGCGTAAAAATGCGC**-3' 3'-**AAGAGGCCGCGCATTTTTACGCG**ACCG-5'	*lacZ*	No
PSL3	GGTT	-	5'-AAAG**TCAATATTGGCTTCATCCACCAC**-3' 3'-**AGTTATAACCGAAGTAGGTGGTG**ACCG-5'	*lacZ*	Yes
PSL4	GGTT	-	5'-AAAG**TTCCGCCAGACGCCACTGCTGCC**-3' 3'-**AAGGCGGTCTGCGGTGACGACGG**ACCG-5'	*lacZ*	No
PSL5	GGTT	-	5'-AAAG**TTCCGCCAGACGCCACTGCTGCCAGGCGCTG**-3' 3'-**AAGGCGGTCTGCGGTGACGACGGTCCGCGAC**ACCG-5'	*lacZ*	Yes
PSP49	AGTT	-	5'-AAG**ATAGCGATCCCTTGCTGAAAATA**-3' 3'-**TATCGCTAGGGAACGACTTTTAT**ACCG-5'	Non-coding	Yes

2.11. Integration assays

Electrocompetent *E. coli* K-12 MG1655 cells were transformed with pHelper w/o array (pTU393; contains tracrRNA and protein-coding sequences) and new pDonor (contains transposon and CRISPR array) or new pDonor V2, and plated overnight in LB agar containing AMP and CHL at 37 ºC. Individual colonies where inoculated in 5 mL of LB medium. IPTG (final concentration of 1 mM throughout) was added after 2.5 h of growth at 37 ºC under agitation. After an additional 2 h of incubation, dilutions of the cells were prepared, plated in LB agar containing AMP, CHL, X-Gal (final concentration of 0.04 mg/mL throughout) and IPTG, and incubated overnight at 37 ºC (figure 5A).

Some assays involved the pFree plasmid. For those, electrocompetent cells were transformed with pFree and plated overnight in LB agar containing KAN, after which they were made electrocompetent. Those cells were then transformed with pHelper w/o array and new pDonor as described above and plated overnight in LB agar containing AMP, CHL and KAN at 37 ºC. Again, individual colonies where inoculated in 5 mL of LB medium. Anhydrotetracycline (final concentration of 200 ng/mL throughout) and rhamnose (final 0.2% throughout) were added after 2.5 h of growth and IPTG after 4.5 h at 37 ºC under agitation. After an additional 2 h of incubation, dilutions of the cells were prepared, plated in LB agar containing AMP, CHL, KAN, X-Gal and IPTG, and incubated overnight at 37 ºC (figure 5B).

A blue/white colony screening was performed on the plates for spacers targeting *lacZ* (assays involving new pDonor and new pDonor V2 containing spacers PSL1-PSL5; figures 5A and 5B).

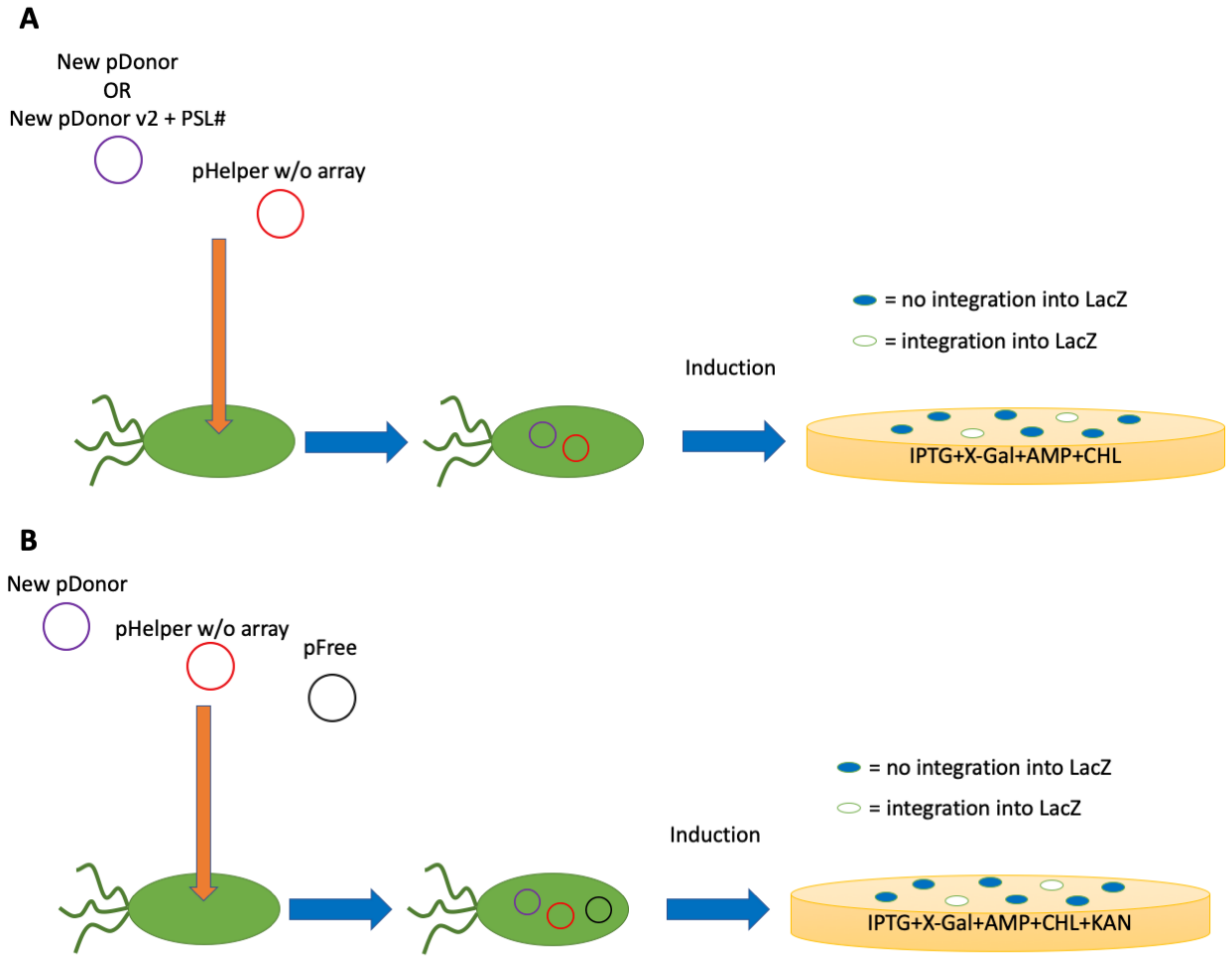


Figure 5. Integration assays and blue/white screening
A. *E. coli* was transformed with pHelper w/o array and new pDonor or new pDonor V2. Following induction of pHelper w/o array, cells were plated and blue/white screening was performed to detect integration events **B.** *E. coli* was transformed with pHelper w/o array, new pDonor and pFree. Following induction of pFree and later pHelper w/o array, cells were plated and blue/white screening was performed to detect integration events.

For assays involving new pDonor V2 containing spacers targeting *lacZ* (PSL1-PSL5), PCR was performed to detect integration events and their directionality on white colonies and cells growing in liquid media (figure 6A), using: I) external forward (BN2170) and external reverse (BN2171), II) external forward (BN2170) and internal (BN2305), and III) internal (BN2305) and external reverse (BN2171).

Similar PCRs were also performed on cells growing in liquid media for a spacer targeting a non-coding region (PSP49; figure 6B), using: I) external forward (BN2346) and external reverse (BN2347), II) external forward (BN2346) and internal (BN2305), and III) internal (BN2305) and external reverse (BN2357).

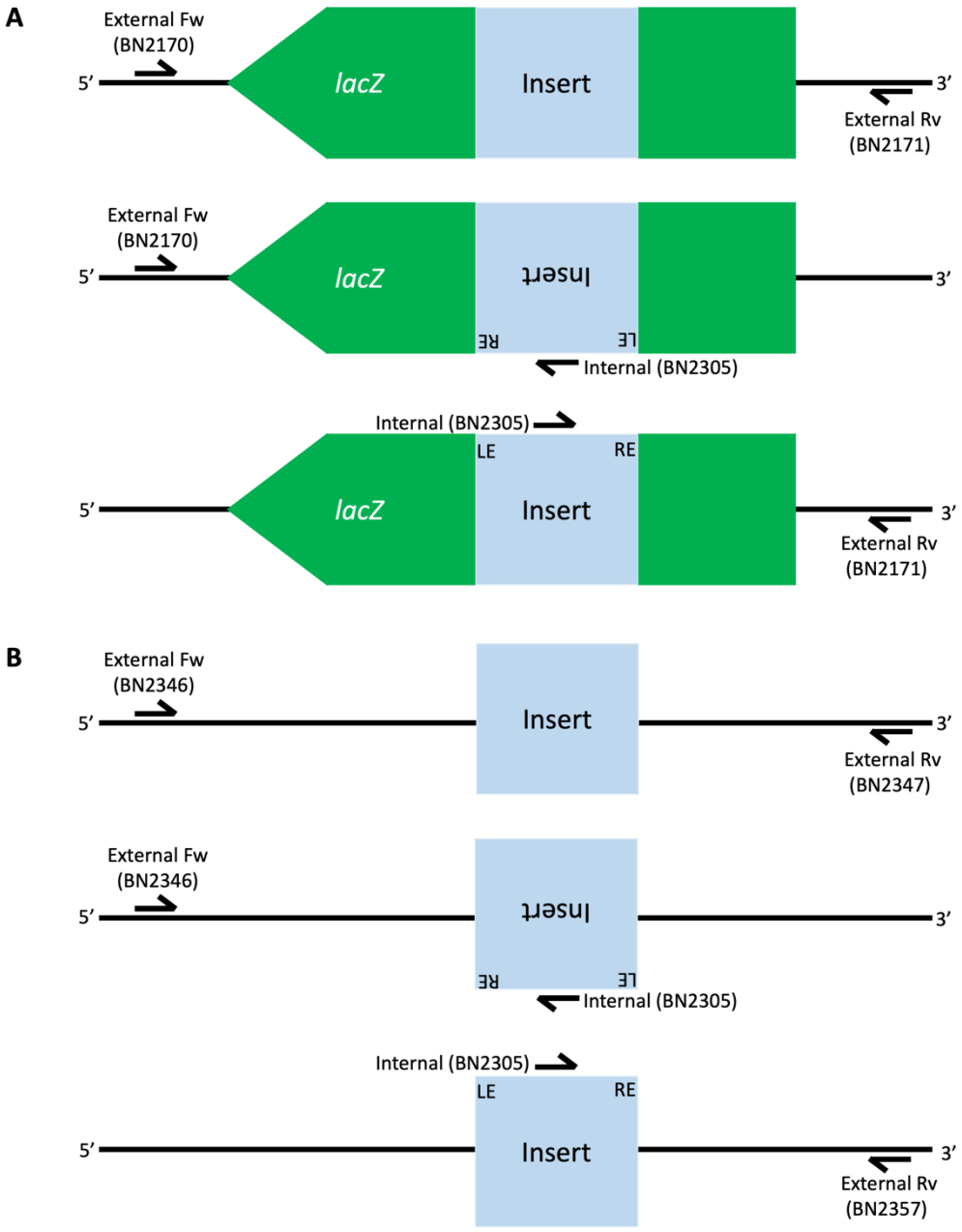


Figure 6. Detection of integrations and its directionality by PCR
Schematics of the designed PCRs aimed at detecting integration events and its directionality, performed by a CAST complex guided by a crRNA targeting **A.** a region inside *lacZ* (PSL1-PSL5) and **B.** a non-coding region (PSP49).

2.12. pFree self-plasmid loss assay

To corroborate that pFree was working correctly, a self-plasmid loss assay was performed. *E. coli* K-12 MG1655 cells were transformed with pFree, plated in LB agar containing KAN, and grown overnight at 37 ºC. An individual colony was inoculated in 5 mL of LB medium. Anhydrotetracycline and rhamnose were added after 2.5 h of growth at 37 ºC under agitation.

The cells were plated in LB agar and incubated overnight at 37 ºC. 50 individual colonies were streaked on both an LB agar plate containing KAN and an LB agar plate without antibiotic (supplementary figure 1 in section 8.3.).

3. Results

3.1. The integration efficiency of CAST guided by a spacer targeting *lacZ* is negligible or null

Blue/white screening relies on the disruption of *lacZ* via an integration event. One of such integration events in the *lacZ* locus would lead to a truncated and non-functional version of β-galactosidase (the product of *lacZ*). β-galactosidase metabolizes X-Gal (added to the plates) creating a blue compound. Therefore, colonies with an undisrupted version of *lacZ* (where no integration event occurred) appear blue, while colonies with a truncated version of *lacZ* (where an integration event happened) appear white.

In this light, an integration assay to assess if CAST is able to integrate a cargo into the *lacZ* locus of a bacterial genome was performed. *E. coli* K-12 MG1655 cells were transformed with pHelper w/o array and new pDonor. The latter plasmid contains a CRISPR array with a single spacer (5'-GCGCAGCCTGAATGGCGAATGGCGCTTTGCC-3') targeting the plus DNA strand of *lacZ* with an AGTT PAM. After induction and plating, blue/white screenings were performed to detect transformants. Most colonies obtained were blue, indicating that no integration events occurred. The few white colonies observed also did not contain the integrated transposon, as confirmed by PCR (supplementary figure 2 in section 8.3.). After several attempts, it was not possible to detect integration events using this spacer targeting *lacZ*.

3.2. CAST is unable to integrate a cargo using a spacer targeting *lacZ*

To better assess if the integration efficiency was null or too small to be detected, the same assay was performed, now including the pFree plasmid. pFree codes for Cas9 and harbors several spacers targeting origins of replication of various plasmids, including those of pHelper w/o array, new pDonor and pFree itself. Upon addition of anhydrotetracycline (inductor of Cas9) and rhamnose (inductor of the CRISPR array), Cas9 produces a DSB in the origins of replication, which aborts plasmid replication. It has been shown that pFree is capable of completely curing around 80% of cells transformed with pFree and other three plasmids (Lauritsen, Porse, Sommer, & Nørholm, 2017).

A cell could be resistant to chloramphenicol either because: I) it bears at least one copy of new pDonor and no integration occurs; II) the CAST complex correctly integrated the CHL^R carrying transposon into the genome and new pDonor has not been curated; or III) the integration was successful and new pDonor was curated. The addition of pFree will drastically reduce the number of cells that are resistant to chloramphenicol because they retain at least one copy of new pDonor (cases I and II). In the presence of chloramphenicol and pFree, cells pertaining to case I that are

successfully curated from new pDonor will die. Cells belonging to case II will now become part of case III.

The objective of this assay was to drastically reduce the number of colonies where no integration occurred (by an estimate of more than 90%; Lauritsen et al., 2017). This is critical for the detection of rare integration events in *lacZ* using this particular spacer, assuming the integration efficiency is not equal to zero. However, results using pFree were identical to those of the prior section (3.1.) with an integration efficiency of zero.

3.3. CAST is unable to integrate a cargo in *lacZ* independently of the spacer used

To investigate whether the observed null efficiency was a consequence of the spacer used, three additional spacers were tested (see table 6 in section 2.10.), targeting different regions of *lacZ* both in the plus and minus strand. Assays analogous to those previously described in section 3.1. (i.e., without pFree) were performed for each new spacer. The integration efficiencies were once again zero for all three spacers. In a final attempt to detect integration events, population PCRs (see section 2.11.) were performed using the liquid cell cultures and the products were run in an agarose gel (figure 7).

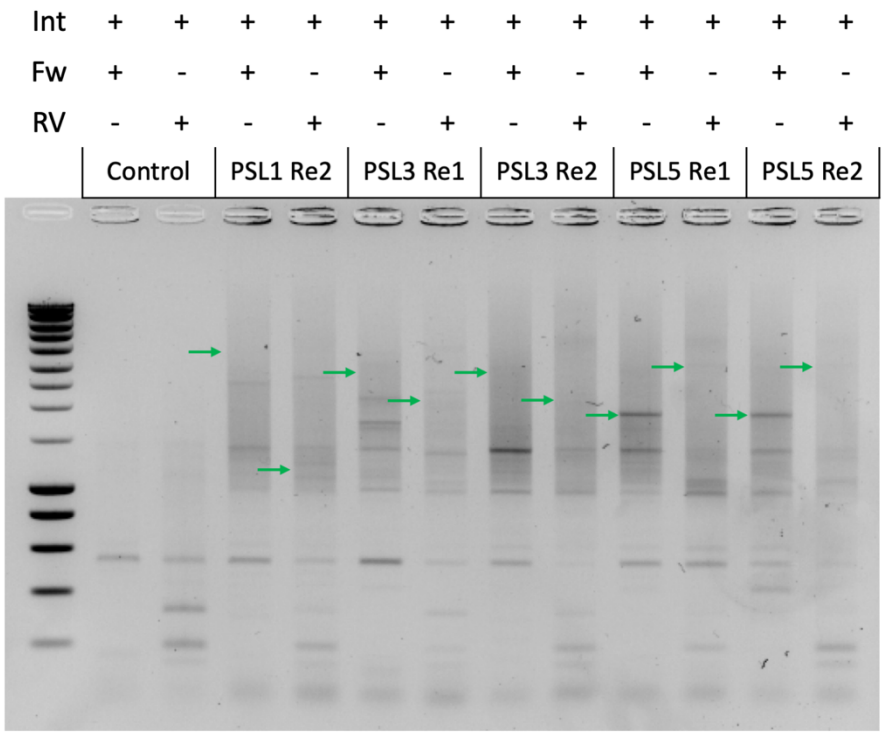


Figure 7. PCR products aimed at detecting integration events into *lacZ* and their directionality.
Integration assays using spacers PSL1, PSL3 and PSL5 were performed in duplicate (two replicas; Re1 and Re2). PCRs directly from liquid media with either an internal (Int) and an external forward (Fw) primers or an internal (Int) and an external reverse (Rv) primers were performed and ran in an agarose gel. The green arrows indicate the expected product size for amplicons derived from a *lacZ* locus interrupted by a transposon. PSL1 Re1 was ran in a separate gel and is not shown here.

Faint bands of the expected size were observed for an integration event occurring in the 5'-RE→LE-3' orientation in both PSL5 spacer replicas, using the internal and forward external primers. However, after repeating the PCR using Q5 polymerase instead of Taq, the bands could not be seen, suggesting these to be unspecific PCR products. Overall, these results strongly suggest that integration events are inexistent at the *lacZ* locus with the current experimental setup.

3.4. CAST is able to integrate a cargo in a non-coding region

To confirm that the setup was functional, and the lack of integration events was not related to the specific experimental setup, a spacer (PSP49) previously shown to be functional in the setup of Strecker, *et al.* (Strecker et al., 2019) was tested. This spacer targets a non-coding region in the genome of *E. coli* and it showed the highest integration efficiencies (~70%) of all spacers tested in the study (Strecker et al., 2019).

Possible integration events were evaluated by population PCRs using combinations of external (complementary to regions adjacent to the non-coding region targeted by PSP49) and internal (complementary to a region inside the CHL^R gene contained in the transposon) primers (figure 8).

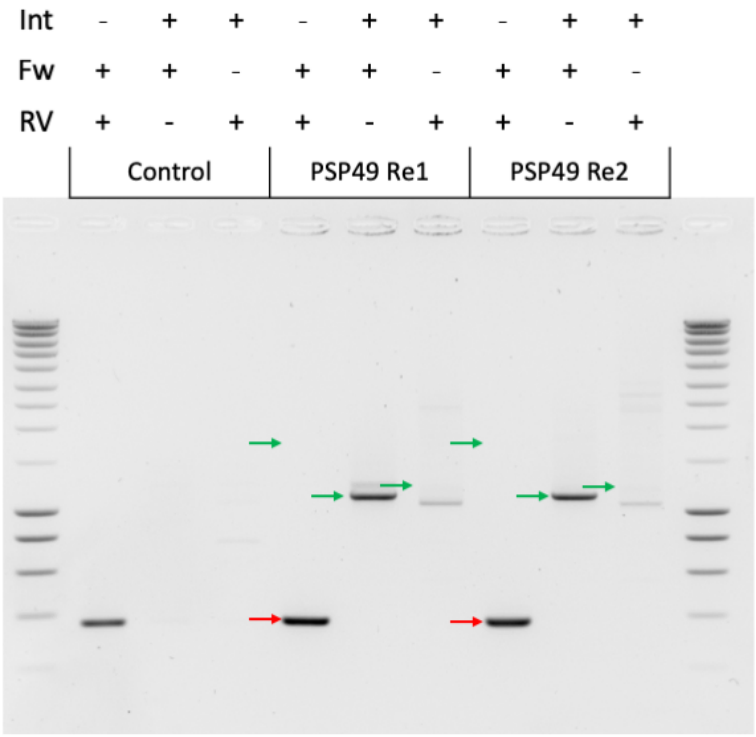


Figure 8. Population PCRs aimed at detecting integration events into a non-coding region and their directionality.
Integration assays using spacer PSP49 were performed in duplicate (Re1 and Re2). PCRs directly from liquid media with either an internal (Int) and an external forward (Fw) primers or an internal (Int) and an external reverse (Rv) primers were performed and ran in an agarose gel. The green arrows indicate the expected product size for amplicons derived from the aforementioned non-coding region interrupted by a transposon and the red arrows indicate the expected product size for amplicons derived from an undisturbed non-coding region.

A sharp band of a size corresponding to an integration event is visible for both replicates when using I) internal and forward external primers, but not when using II) internal and reverse external primers or III) forward external and reverse external (figure 8). Observation I suggests that the transposon was integrated by the CAST complex in a 5'-LE→RE-3' orientation. This was confirmed by sequencing the band. Moreover, the sequencing resulted in a chromatogram with defined and sharp peaks, which when aligned to the non-coding region, revealed the transposon was consistently integrated 61 bp downstream of the PAM (supplementary figure 3 in section 8.3.). If the integration position was not consistent, the sequencing would have yielded a chromatogram with overlapping and broad peaks due the presence of frameshifted sequences. From observation II, it is clear that integrations did not occur in any detectable amount in the opposite (5'-RE→LE-3') orientation. Finally, a relatively low overall integration efficiency is evidenced by observation III, as the undisrupted locus was preferentially amplified over the transposon-interrupted one, due to an overwhelming relative abundance of the undisrupted version in the population.

4. Discussion

Our results demonstrate that a transposon target cannot be integrated by CAST into the *lacZ* locus, but can be integrated in a non-coding region. When the CAST complex is induced with IPTG, the lac operon is also induced, causing *lacZ*, *cas12K* and the genes that form the transposon all to be expressed at high levels. When a gene is highly expressed there is high abundance of RNA polymerase (RNAP) at the corresponding locus, which does not occur in non-coding regions. We therefore hypothesize that RNAP molecules inhibit the integration of transposons by the CAST complex in coding regions.

In vitro, it has been shown that Cas9 (as well as dCas9) can remain stably bound to the DSB site for 5.5 h (Richardson, Ray, Dewitt, Curie, & Corn, 2016). *In* vivo, this would prevent the DNA repairing machinery to reach the DSB site. Clarke *et al.* found that RNAP present in transcriptionally active sites can increase the efficiency of editing by dislodging the Cas9 from the DSB site. However, this only happens when the single guide RNA (sgRNA; a fusion of a crRNA and a tracrRNA) base pairs with the same strand the RNAP uses as a template for transcription (template orientation), but not when it base pairs the opposite strand (non-template orientation; Clarke et al., 2018). Once a Cas9 has found its target, it promptly performs a DSB, after which the RNAP can dislodge it, ultimately leading to an increase in editing efficiency. If, however, Cas9 was to take a considerable amount of time to perform a DSB after finding the target, RNAPs would likely dislodge it before it can cleave the DNA. This last scenario would lead to a decrease in editing efficiency rather than an increase.

It is reasonable to assume that Cas12 behaves similarly. In this study, the integration rate of the CAST complex ($k_{i\text{-}CAST}$), rather than the DSB rate is of concern. If the CAST complex binds in the template orientation, and if $k_{i\text{-}CAST}$ is small when compared to the rate at which RNAPs bind the promoter to start transcription for a typical intracellular RNAP concentration ($k_{on\text{-}RNAP}$), this will result in the ubiquitous dislodgement of the CAST complex before it is able to perform the integration (figure 9A). This would happen when CAST targets a highly expressed gene, as $k_{on\text{-}RNAP}$ increases with the concentration of the inducer. On the other hand, when targeting a non-coding region, $k_{on\text{-}RNAP}$ effectively equals zero. In this scenario, CAST would not be dislodged, resulting in

an integration event. When the CAST complex binds the DNA in the non-template orientation, the coalition with the RNAP would not result in a dislodgment event, but in the physical impediment of the transposition proteins to reach the target DNA (figure 9B).

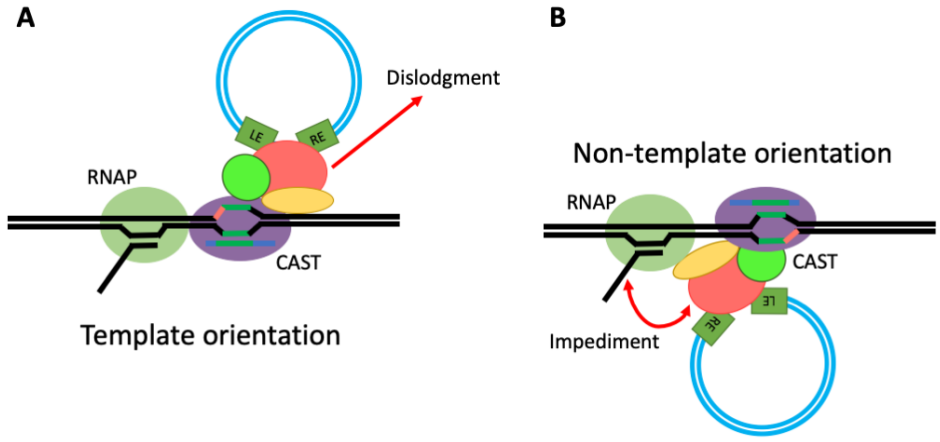


Figure 9. Interaction model between RNAP and the CAST complex.
A. An incoming RNAP is able to dislodge the CAST complex when bound to the target DNA in the template orientation, thereby preventing an integration event. **B.** After an RNAP collides with the CAST complex when bound to the target DNA in the non-template orientation, it blocks the access of the transposition proteins to the target DNA, and in doing so prevents an integration event.

This model could be further supported by additional experiments. For example, an RNAP inhibitor (like rifampin; Wehrli, 1983) could be added some hours after the CAST system is induced with IPTG. If the model holds, integration events would now be detected when targeting *lacZ*. Alternatively, a washing step, where the liquid culture is centrifuged and the supernatant is replaced by fresh media free of IPTG, can be performed instead of using an RNAP inhibitor. A third option would be to replace the promoter in new pHelper w/o array, for a promoter induced by a chemical other than IPTG. In that case, new pHelper w/o array would be induced first when on liquid media, and IPTG could be added during the plating stage for *lacZ* screening purposes. The independent control on the number of CAST complexes and the activity of RNAP in the *lacZ* locus also poses the possibility of carrying out assays to determine the integration efficiency as a function of those parameters (or at the very least as a function of the concentration of the inductors). Alternative screening methods could be used as well, providing data of integration efficiencies for different genes. For example the *tonA* gene in *E. coli* codes for an outer membrane protein involved in ferrichrome-iron uptake, which is used by the T5 phage as a receptor (Menichi, Buu, Microbiologie, & Paris-xi, 1983). This gene should be expressed at significantly lower levels than *lacZ* in presence of IPTG for the concentrations used in this study. The assay would consist in inducing the integration of the transposon amidst *tonA* in liquid media and then plating in a petri dish. Next, single colonies would be isolated and each would be suspended in individual wells of a microplate together with the bacteriophage T5. A time series of the optical density (which should depend on the concentration of living cells in a monotonically increasing fashion) would then be recorded by using a plate reader. If the optical density decreases over time, this would mean the bacterial cells are being lysed by the phage due to a lack of disruption of *tonA*

that would be caused by an integration event. On the other hand, an integration event would result in truncated TonA receptors, making infection impossible. The optical density would then increase or remain roughly the same mimicking a non-infected control.

The fact that the I-F system is capable of integrating targets into the *lacZ* locus even when the Cascade and transposition proteins are induced by IPTG, suggests this system is robust when confronted with RNAPs (Klompe et al., 2019). The model described above for CAST therefore does not apply to the I-F system. This higher robustness is also evidenced by the results obtained while developing MUCICAT (mentioned in section 1.5), where the authors settled for the I-F system over the V-K, due in part to cargo integration problems while targeting a gene (*insB1*) forming part of the *E. coli* multicopy insertion sequence 1 (IS1; Zhang et al., 2020). However, the I-F system has, as previously mentioned, the downside of integration efficiencies falling dramatically as the size of the cargo increases (Klompe et al., 2019; Strecker et al., 2019).

The finding that CAST is unable to integrate transposons in an active gene, coupled with previously mentioned observations about the occurrence of high number of co-integration (Strecker et al., 2020) and off-target integration events (Strecker et al., 2019), suggest that this system might not be suitable for certain applications in its current form. Although promising, the discovery and characterization of "better behaved" CRISPR-associated transposases or the engineering of the CAST complex would be required for the wide adaption of this technology by the scientific community and the biotechnological and medical industries. Because it is speculated that the transposition proteins can work on their own, independently of the effector complex, the creation of a fusion protein composed of Cas12K and one of the transposition proteins could greatly reduce the number of off-target integration events by imposing the condition that all the transposition proteins can form a complex only if Cas12K is also a part of it. Additionally, this might increase $k_{i\text{-}CAST}$, as one of the main factors limiting the rapid integration of a transposon might be that most CAST complexes fully form upon Cas12K finding its target rather than before, which leads to an RNAP-mediated dislodgment or impediment to access the target DNA prior to the formation of full CAST complexes. A fusion as such could provide a head start for the full complex formation. It is critical, however, to obtain a structural reconstruction of the CAST complex to better understand the precise way in which all the proteins interact with each other. The problem of co-integrations could be solved without nicking the 5' ends of the transposon by adding an appropriate *tnsA* homologue to the system. Most likely, however, some rational engineering or directed evolution approaches would be necessary for TnsA to correctly work with the other components of the CAST complex.

Another important aspect of this study is the integration efficiency when targeting the non-coding region specified by the PSP49 spacer. In the assays performed by Strecker *et al.*, this efficiency was on average around 70% (Strecker et al., 2019). Although the integration efficiency in genomic loci other than *lacZ* was not quantified in this study, as explained in section 3.4, the integration efficiency was found to be low, certainly well below 70%. This discrepancy could be due to the different backbones used in the donor plasmids. The pDonor used by Strecker *et al.* has an R6Kγ origin of replication (R6Kγori). The replication of plasmids containing this origin rely on the π protein encoded by the *pir* gene. Cells harboring the wild type *pir* gene maintain R6Kγori^{+} plasmids at a copy number of about 15. However, some mutated versions of the *pir* gene maintain the plasmid at around 250 copies (Kvitko et al., 2012). For their integration assays, Strecker *et al*.

cloned pDonor and pHelper into One Shot Pir1 Chemically Competent *E. coli* (Invitrogen; Strecker et al., 2019), which contain this mutant allele. Because *pir*$^{+}$ cells are engineered, most *E. coli* strains do not contain this gene. As the creation of a system compatible with most *E. coli* strains was desired for this study, the backbone of pDonor was here replaced for that of pACYCDuet-1, which maintains the number of copies per cell at 10-12. The integration efficiency using this backbone was apparently reduced compared to that observed with the original backbone; this suggests that transposition efficiencies are highly dependent on the number of transposons present in the cell for the range of 10-250 copies.

The choice of backbone in this study aimed at maintaining a low copy number for the integration assays involving targeting of *lacZ*. The curation process by pFree can be more efficient when targeting such a low copy number plasmid. Therefore, the transformation efficiency calculated for assays involving pFree is "apparent" rather than "real", as a background-reduction process was at play. Alternative background-reducing strategies involving the utilization of a thermosensitive new pDonor V2 with a pKD46 backbone (for more information see Datsenko & Wanner, 2000), and a version of new pDonor V2 containing a gene (*sacB*) coding for a toxin-producing enzyme in its backbone (to learn more see X. Li, Thomason, Sawitzke, Costantino, & Court, 2013), were considered. Because the integration efficiency seems to be null for the case discussed above, these strategies will probably be unfruitful as well. However, they could prove useful for cases in which the efficiency is small but not null; perhaps when targeting less active genes.

The discovery of CRISPR-associated transposases represents a major milestone in the CRISPR field. It has the potential of opening a myriad of possibilities for genetic engineers and synthetic biologists, allowing for the efficient genomic integration of genes or even bigger elements like operons by bypassing the main limiting factor of Cas9-based technologies, the homology directed repair. However, several issues related to high levels of off-target integrations (Strecker et al., 2019), considerable amount of co-integration events (Strecker et al., 2020), and extremely low levels of integration efficiencies in transcriptionally highly active loci, have to be dealt with before CRISPR-associated transposases are incorporated into the mainstream CRISPR toolbox. Recent discoveries and developments like CRISPR-associated transposases or prime editing make the CRISPR field as exciting today as it was when Mojica realized he stumbled upon a prokaryotic adaptive immune system (Mojica et al., 2005), or when Cas proteins were first exploited to edit the genomes of mammalian cells (Cong et al., 2013).

5. Acknowledgments

I would like to thank Stan Brouns for hosting me in his lab and giving me the opportunity to further develop my scientific skills. I would also like to thank Rita Costa and Cristóbal Almendros for their day to day supervision. Cristóbal was my supervisor for about the first half of the project. He was involved in the setup process and taught me the necessary techniques to carry out the project. Rita supervised me for about the second half and was involved in the further development of the project and in the writing process of this thesis. I learned a lot from both.

Furthermore, I would like to thank everyone at the lab; including Stan, all the staff, and the students. All of you helped me at some point by indicating ways I could improve my experiments, pointing me towards a different direction than I originally envisioned, providing materials, or simply making the lab an enjoyable place to work at. I would like to highlight the contribution of Sebastian Kieper, with whom I wrote a review paper on Cas4, the introduction of which was included in an adapted form as section 1.1. of this thesis. I am also grateful to Franklin Nobrega for his support and for kick-starting interesting conversations.

For forming part of the assessment committee I would like to thank Christophe Danelon and Chirlmin Joo, who introduced me to the world of CRISPR and scientific research.

Lastly, I would like to thank my parents for the support I have received from them all these years.

6. Notes

Due to the COVID-19 outbreak in the Netherlands, all experimental work came to a halt as of the beginning of March. Because of this, experiments planned to strengthen the main conclusion of the thesis, to reduce the background for blue/white screening, and to develop alternative screening methods, could not be performed but are mentioned in the discussion section.

7. References

Adli, M. (2018). The CRISPR tool kit for genome editing and beyond. *Nat Commun*, *9*.

Anzalone, A. V, Randolph, P. B., Davis, J. R., Sousa, A. A., Koblan, L. W., Levy, J. M., … Liu, D. R. (2019). Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature*, *576*, 149–157.

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

Bolotin, A., Quinquis, B., Sorokin, A., & Ehrlich, S. D. (2005). Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology*, *151*(8), 2551–2561.

Clarke, R., Heler, R., Macdougall, M. S., Church, G. M., Marraffini, L. A., Merrill, B. J., … Merrill, B. J. (2018). Enhanced Bacterial Immunity and Mammalian Genome Editing via RNA-Polymerase-Mediated Dislodging of Cas9 from Double-Strand DNA Breaks. *Mol Cell*, *71*, 42–55.

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., … Zhang, F. (2013). Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science*, *339*, 819–823.

Datsenko, K. A., & Wanner, B. L. (2000). One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *PNAS*, *97*, 6640–6645.

Deltcheva, E., Chylinski, K., Sharma, C. M., Gonzales, K., Chao, Y., Pirzada, Z. A., … Charpentier, E. (2011). CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature*, *471*, 602–607.

Durai, S., Mani, M., Kandavelou, K., Wu, J., Porteus, M. H., & Chandrasegaran, S. (2005). Zinc finger nucleases: custom-designed molecular scissors for genome engineering of plant and mammalian cells. *Nucleic Acids Res*, *33*, 5978–5990.

East-seletsky, A., Connell, M. R. O., Knight, S. C., Burstein, D., Cate, J. H. D., Tjian, R., & Doudna, J. A. (2016). Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection. *Nature*, *538*, 270–273.

Faure, G., Scott, D. A., & Peters, J. E. (2019). CRISPR–Cas in mobile genetic elements: counter-defence and beyond. *Nat Rev Microbio*, *17*, 513–525.

Fonfara, I., Richter, H., Bratovič, M., Le Rhun, A., & Charpentier, E. (2016). The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. *Nature*, *532*, 517–532.

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

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

Jackson, S. A., Mckenzie, R. E., Fagerlund, R. D., Kieper, S. N., Fineran, P. C., & Brouns, S. J. J. (2017). CRISPR-Cas : Adapting to change. *Science*, *356*.

Jiang, F., & Doudna, J. A. (2017). CRISPR–Cas9 Structures and Mechanisms. *Annu Rev Biophys*, *46*, 505–529.

Joung, J. K., & Sander, J. D. (2013). TALENs: a widely applicable technology for targeted genome editing. *Nat Rev Mol Cell Biol*, *14*, 49–55.

Kazazian Jr., H. H. (2004). Mobile Elements: Drivers of Genome Evolution. *Science*, *303*, 1626–1633.

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

Koonin, E. V., Makarova, K. S., & Zhang, F. (2017). Diversity, classification and evolution of CRISPR-Cas systems. *Curr Opin Microbiol*, *37*, 67–78.

Kvitko, B. H., Bruckbauer, S., Prucha, J., Mcmillan, I., Breland, E. J., Lehman, S., … Schweizer, H. P. (2012). A simple method for construction of pir + Enterobacterial hosts for maintenance of R6K replicon plasmids. *BMC Res Notes*, *5*.

Lauritsen, I., Porse, A., Sommer, M. O. A., & Nørholm, M. H. H. (2017). A versatile one-step CRISPR-Cas9 based approach to plasmid-curing. *Microb Cell Fact*, *16*.

Li, K., Wang, G., Andersen, T., Zhou, P., & Pu, W. T. (2014). Optimization of Genome Engineering Approaches with the CRISPR/Cas9 System. *PLoS ONE*, *9*.

Li, X., Thomason, L. C., Sawitzke, J. A., Costantino, N., & Court, D. L. (2013). Positive and negative selection using the tetA-sacB cassette: recombineering and P1 transduction in Escherichia coli. *Nucleic Acids Res*, *41*.

Makarova, K. S., Wolf, Y. I., Alkhnbashi, O. S., Costa, F., Shah, S. A., Saunders, S. J., … Koonin, E. V. (2015). An updated evolutionary classification of CRISPR–Cas systems. *Nat Rev Microbiol*, *13*, 722–736.

Makarova, K. S., Wolf, Y. I., Iranzo, J., Shmakov, S. A., Alkhnbashi, O. S., Brouns, S. J. J., … Koonin, E. V. (2020). Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nat Rev Microbiol*, *18*, 67–83.

Marraffini, L. A., & Sontheimer, E. J. (2008). CRISPR Interference Limits Horizontal Gene Transfer in Staphylococci by Targeting DNA. *Science*, *322*, 1843–1845.

May, E. W., & Craig, N. L. (1996). Switching from Cut-and-Paste to Replicative Tn7 Transposition. *Science*, *272*, 401–404.

Menichi, B., Buu, A., Microbiologie, I. De, & Paris-xi, U. (1983). Integration of the Overproduced Bacteriophage T5 Receptor Protein in the Outer Membrane of Escherichia coli. *J Bacteriol*, *154*, 130–138.

Mohanraju, P., Makarova, K. S., Zetsche, B., Zhang, F., Koonin, E. V., & Oost, J. van der. (2016). Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. *Science*, *353*, aad5147-1-add5147-12.

Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J., & Soria, E. (2005). Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements. *J Mol Evol*, *60*, 174–182.

Mojica, F. J. M., Juez, G., & Rodríguez-Valera, F. (1993). Transcription at different salinities of Haloferax mediterranei sequences adjacent to partially modified Pstl sites. *Mol Microbiol*, *9*, 613–621.

Peters, J. E., & Craig, N. L. (2001). Tn7: Smarter Than We Thought. *Nat Rev Mol Cell Biol*, *2*, 806–814.

Rice, P. A., Craig, N. L., & Dyda, F. (2020). Comment on "RNA-guided DNA insertion with CRISPR-associated transposases." *Science*.

Richardson, C. D., Ray, G. J., Dewitt, M. A., Curie, G. L., & Corn, J. E. (2016). Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. *Nat Biotech*, *34*, 339–344.

Rozov, S. M., Permyakova, N. V, & Deineko, E. V. (2019). The Problem of the Low Rates of

CRISPR/Cas9-Mediated Knock-ins in Plants: Approaches and Solutions. *Int J Mol Sci*, *20*.
Strecker, J., Ladha, A., Gardner, Z., Schmid-burgk, J. L., & Kira, S. (2019). RNA-guided DNA insertion with CRISPR-associated transposases. *Science*, *365*, 1–12.
Strecker, J., Ladha, A., Makarova, K. S., Koonin, E. V, & Zhang, F. (2020). Response to Comment on "RNA-guided DNA insertion with CRISPR-associated transposases." *Science*.
Sundstrom, L., & Sköld, O. L. A. (1990). The dhfrI Trimethoprim Resistance Gene of Tn7 Can Be Found at Specific Sites in Other Genetic Surroundings. *Antimicrob Agents Ch*, *34*, 642–650.
Van Der Oost, J., Westra, E. R., Jackson, R. N., & Wiedenheft, B. (2014). Unravelling the structural and mechanistic basis of CRISPR–Cas systems. *Nat Rev Microbio*, *12*, 479–492.
Wehrli, W. (1983). Rifampin: Mechanisms of Action and Resistance. *Rev Infect Dis*, *5*, S407–S411.
Westra, E. R., Dowling, A. J., Broniewski, J. M., & Houte, S. Van. (2016). Evolution and Ecology of CRISPR. *Annu Rev Ecol Evol Syst*, *47*, 307–331.
Zhang, Y., Sun, X., Wang, Q., Xu, J., Dong, F., Yang, S., … Zhang, Z. (2020). Multicopy chromosomal integration using CRISPR-associated transposases. *ACS Syn Bio*.

8. Appendices

8.1. Plasmids

Supplementary Table 1. List of plasmids used in this study. "-" in lab name indicates that the plasmid is commercially available.

Lab name	Common name	Description	Antibiotic resistance
-	pHelper	Contains a CRISPR array, tracrRNA and genes coding for transposition proteins and Cas12K	AMP
pTU393	pHelper w/o array	pHelper without array. Contains tracrRNA and genes coding for transposition proteins and Cas12K. Induced by IPTG	AMP
-	pDonor	Contains a Tn7-like transposon	KAN
-	pACYCDuet-1	Low copy number expression plasmid	CHL
pTU394	New pDonor	pACYCDuet-1 backbone with a Tn7-like transposon and a CRISPR array	CHL
-	pFree	Expression of Cas9 and crRNAs targeting common origins of replication for plasmid curing.	KAN
pTU395	New pDonor V2	New pDonor with interchangeable spacer	CHL
-	pKD46	Low copy number thermosensitive expression plasmid. Curated by growing at ≥37 ºC	AMP
pTU396	Thermosensitive new pDonor V2	New pDonor V2 with pKD46 backbone. Curated by growing at ≥37 ºC	CHL

8.2. Oligonucleotides

Supplementary Table 2. List of oligonucleotides used in this study, including primers and oligonucleotides used for the creation of spacers.

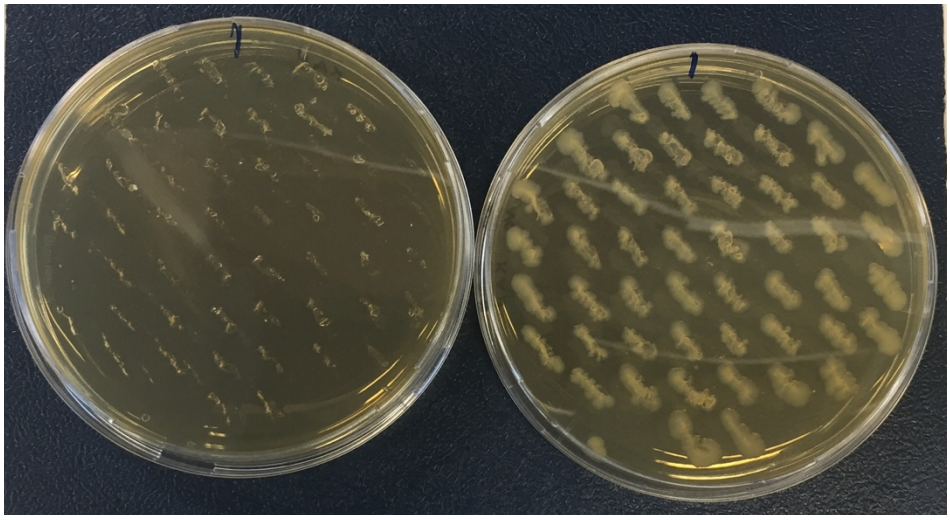
Name	Sequence (5'→3')	Description
T7 ter Rv	GCTAGTTATTGCTCAGCGG	Rv. Confirm cloning of the synthetic transposon into pACYCDuet-1
BN514	GATGGTGTCCGGGATCTC	Fw. Confirm cloning of the synthetic transposon into pACYCDuet-1
BN1915	TTTCCTAATGCAGGAGTCGCATAAGGGAGA	Rv (BN2054 as Fw). Eliminate multiple cloning site from pACYCDuet-1
BN2040	CTCAGTACAATCTGCTCTGATGC	Fw. Eliminate array from pHelper
BN2041	CTACCCAGTTACAAAATGTTGTG	Rv. Eliminate array from pHelper
BN2042	AACCAAGGCATGATTCTTGAGAC	Fw (BN2041 as Rv). Confirm elimination of array from pHelper
BN2054	TTTAAGAAGGAGATATAGATATCCCAACTCCATAACAGCTTAATTAACCTAGGCTGCTG	Fw (BN1915 as Rv). Introduce LIC site pACYCDuet-1 and eliminate multiple cloning site
BN2166	TTTAAGAAGGAGATATAGATTTAGACATCTCCACAAAAGGCG	Fw. Amplify synthetic transposon for new pDonor
BN2167	TTATGGAGTTGGGATCTTATTAGGATCCCTTTCAAC	Rv. Amplify synthetic transposon for new pDonor
BN2170	TTACGCGAAATACGGGCAGA	Fw. Confirm integration in LacZ locus (*E. coli* K12 MG1655 genome)
BN2171	TGTGTGGAATTGTGAGCGGA	Rv. Confirm integration in LacZ locus (*E. coli* K12 MG1655 genome)
BN2172	CACTATAGGGCGAATTGGCGG	Fw. Amplify synthetic CRISPR array with BsaI sites for new pDonor V2

BN2173	ACTGGAAAGCGGGCAGTGA	Fw. Amplify synthetic CRISPR array with BsaI sites for new pDonor V2
BN2187	AGGTAGTCACGCAACTCGCC	Fw. Confirm integration in LacZ locus (*E. coli* K12 MG1655 genome; length w/o integration=719 bp)
BN2188	GACTGGGAAAACCCTGGCGT	Rv. Confirm integration in LacZ locus (*E. coli* K12 MG1655 genome; length w/o integration=719 bp)
BN2197	GTGAGACCAGTCTCGGAA	Fw (T7 ter Rv as Rv). Confirm integration of CRISPR array into new pDonor
BN2239	AACCCTTTAGAATTCGTGGC	Fw. Amplify new pDonor without including array
BN2240	TAGGATCCCTTTCAACCCAC	Rv. Amplify new pDonor without including array
BN2251	AAAGTCCGGCACCAGAAGCGGTGCCGG	Spacer PSL1 for new pDonor v2 + strand
BN2252	GCCACCGGCACCGCTTCTGGTGCCGGA	Spacer PSL1 for new pDonor v2 - strand
BN2253	AAAGTTCTCCGGCGCGTAAAAATGCGC	Spacer PSL2 for new pDonor v2 + strand
BN2254	GCCAGCGCATTTTTACGCGCCGGAGAA	Spacer PSL2 for new pDonor v2 - strand
BN2255	AAAGTCAATATTGGCTTCATCCACCAC	Spacer PSL3 for new pDonor v2 + strand
BN2256	GCCAGTGGTGGATGAAGCCAATATTGA	Spacer PSL3 for new pDonor v2 - strand
BN2257	AAAGTTCCGCCAGACGCCACTGCTGCC	Spacer PSL4 for new pDonor v2 + strand
BN2258	GCCAGGCAGCAGTGGCGTCTGGCGGAA	Spacer PSL4 for new pDonor v2 - strand
BN2259	AAAGTTCCGCCAGACGCCACTGCTGCCAG GCGCTG	Spacer PSL5 for new pDonor v2 + strand
BN2260	GCCACAGCGCCTGGCAGCAGTGGCGTCTG GCGGAA	Spacer PSL5 for new pDonor v2 - strand

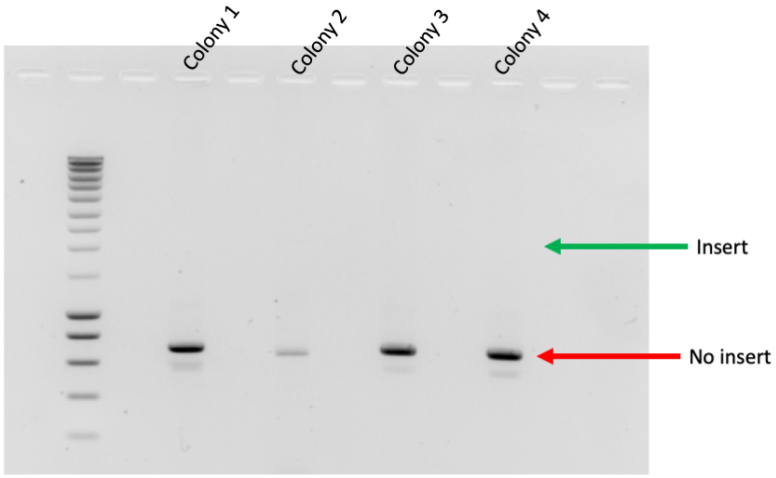
BN2261	AAAGATAGCGATCCCTTGCTGAAAATA	Spacer PSP49 for new pDonor v2 + strand
BN2262	GCCATATTTTCAGCAAGGGATCGCTAT	Spacer PSP49 for new pDonor v2 - strand
BN2263	GGATGGAGGCGGATAAAGTTGCA	Sequence pHelper w/o array (seq2)
BN2264	GAGCCTATGGAAAAACGCCAGCA	Sequence pHelper w/o array (seq3)
BN2265	CACTTGAGTCAGATTGGGGCG	Sequence pHelper w/o array (seq4)
BN2266	GTCTCAAGGGTAAGAGCATTGTG	Sequence pHelper w/o array (seq5)
BN2267	CCAGATGTTACCGCCTGCT	Sequence pHelper w/o array (seq6)
BN2268	GGTGACACTTTAGAGGCTATTCG	Sequence pHelper w/o array (seq7)
BN2269	GGTTGGAGTAAGCCTGGGACTA	Sequence pHelper w/o array (seq8)
BN2270	GAGTATCGAGATGGCACATAGC	Fw. PCR out pKD46 ori
BN2271	GGTAACTGTCAGACCAAGTTTACTC	PCR out pKD46 ori Rv
BN2272	CGCCTTATCCGGTAACTATCG	PCR new pDonor w/o ori Fw
BN2273	CAGTAGCTGAACAGGAGGG	PCR new pDonor w/o ori Rv
BN2305	CCACCGTTGATATATCCCAATGGCATCG	Fw (for 5'-L→R-3' orientation). Internal primer for transposon (CHL^R gene)
BN2335	GGTTGGACTCAAGACGATAGTTACCGGAT AAGGGGGTTATTGTCTCATGAGCGG	Fw (BN2271 as Rv). Amplify pKD46 without AMP^R and with spacer and PAM targeted by pFree
BN2346	GTCAGGTAGCCAGAACACCC	Fw. Confirm integration in non-coding region targeted by PSP49 (*E. coli* K12 MG1655 genome)
BN2347	GCCGGGATACGTTCCTTCTT	Rv. Confirm integration in non-coding region targeted by PSP49 (*E. coli* K12 MG1655 genome)
BN2362	AAAGGCAGGGCAACTTCTATAACGATG	Spacer PST1 for new pDonor v2 + strand

BN2363	GCCACATCGTTATAGAAGTTGCCCTGC	Spacer PST1 for new pDonor v2 - strand

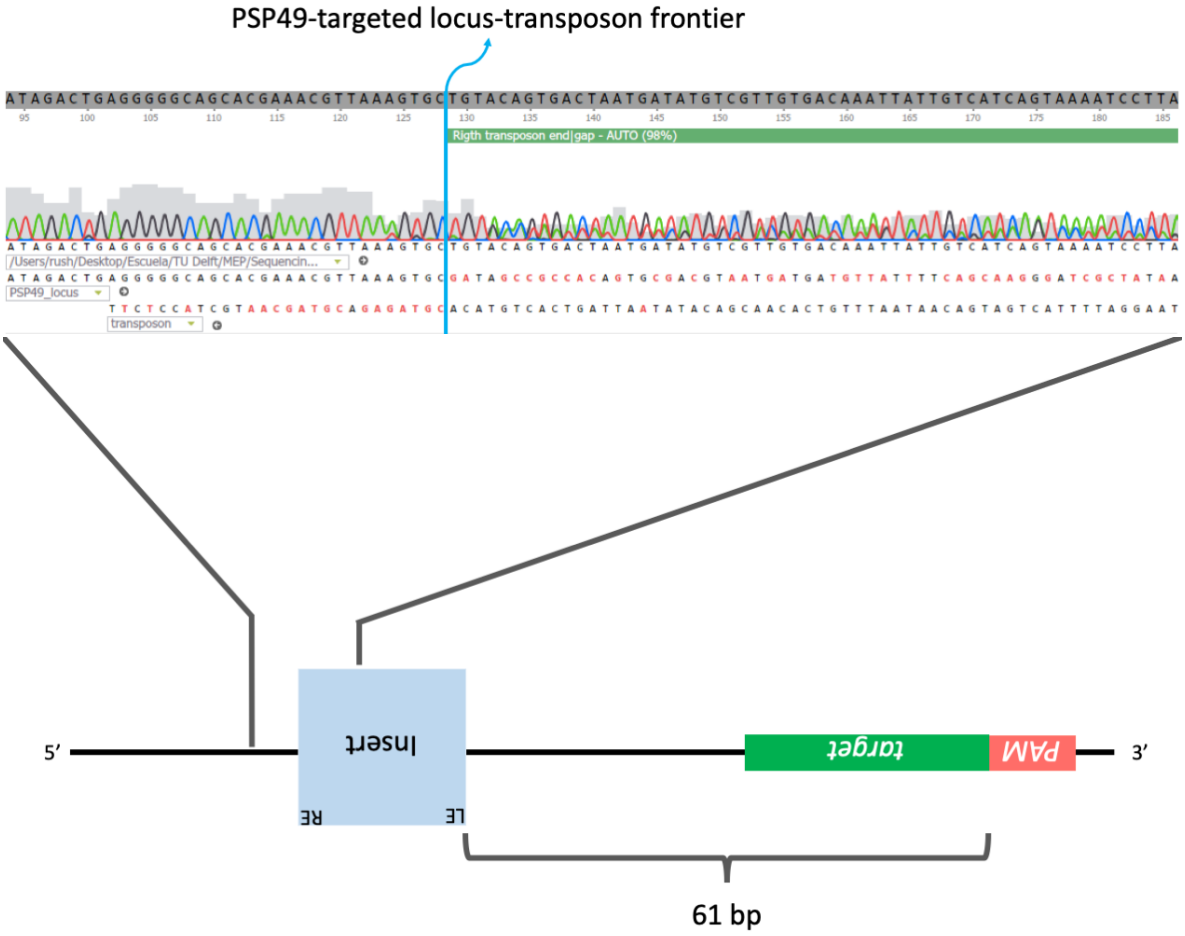
8.3. Supplementary figures



Supplementary figure 1. pFree self-plasmid loss assay results.
Cells in which pFree was not curated would grow in the plate containing KAN (left) and the one containing no antibiotic (right). Cells in which pFree was successfully curated would exclusively grow in the plate containing no antibiotic. All 50 cells tested successfully for loss of pFree.



Supplementary figure 2. PCR products aimed at detecting integration events into *lacZ* in white colonies after blue/white screening.
PCRs on white colonies after blue/white screening were performed and ran in an agarose gel. The green arrows indicate the expected product size for amplicons derived from *lacZ* interrupted by a transposon and the red arrows indicate the expected product size for amplicons derived from an undisturbed *lacZ* locus.



Supplementary figure 3. Transposon insertion site, sequencing results and alignments in the PSP49-targeted locus.
Sequencing results (from DNA contained in the band indicated by a green arrow in the int +, fw +, rv -, lane in figure 8) aligned with the non-coding PSP49-targeted locus and the transposon. The frontier between the locus and the inserted transposon (where one side matches the PSP49-targeted locus and the other matches the transposon) is indicated with a blue line. The relative position of the sequencing results, as well as the distance from the PAM to the inserted transposon are indicated.