Title:

Do the shuffle: Expanding the synthetic biology toolkit for shufflon-like recombination systems

Author List

Jan Katalinić[1], Morgan Richards[1], Alex Auyang[1], James H. Millett[1], Nikolai Windbichler[1,*]

*nikolai.windbichler@imperial.ac.uk

Affiliations

1. Department of Life Sciences, Imperial College London, UK

Abstract

Naturally occurring DNA inversion systems play an important role in the generation of genetic variation and adaptation in prokaryotes. Shufflon invertase (SI) *Rci* from plasmid R64, recognizing asymmetric *sfx* sites, has been adopted as a tool for synthetic biology. However, the availability of a single enzyme with moderate rates of recombination has hampered the more widespread use of SIs. We identified 14 previously untested SI genes and their *sfx* sites in public databases. We established an assay based on single-molecule sequencing that allows the quantification of the inversion rates of these enzymes and determined cross-recognition to identify orthogonal SI/*sfx* pairs. We describe SI enzymes with substantially improved shuffling rates when expressed in an inducible manner in *E. coli*. Our findings will facilitate the use of SIs in engineering biology where synthetic shufflons enable the generation of millions of sequence variants *in vivo* for applications such as barcoding or experimental selection.

Keywords

Rci, recombinase, invertase, DNA inversion, shufflon, barcoding

Introduction

Bacteria employ multiple mechanisms for DNA inversion to achieve phenotypic heterogeneity within the population[1], a process that is now believed to be more widespread than previously thought[2]. This strategy generates variation and allows the selection of lineages with a set of beneficial genetic traits, for instance under conditions of environmental stress. One mechanism to achieve this is conservative site-specific DNA recombination mediated by a recombinase which typically binds to a single pair of recognition sites and inverts the intervening DNA sequence. However, a more interesting case is presented by shufflons - multiple inversion systems that contain several recombination sites that flank and separate multiple invertible coding regions and as a result can produce a larger set of alternative proteins[3]. The presence of multiple recombination sites enables a single recombinase or shufflon invertase (SI) to 'shuffle' between DNA segments to produce alternative alleles. In *Bacteroides fragilis* the SI Tsr0667 shuffles a gene cluster of outer-membrane proteins of the SusC/SusD family to produce alternative alleles utilising different promoters[4]. This mechanism leads to varied combinations and expression levels of SusC/SusD proteins contributing to differential polysaccharide utilisation of *B. fragilis*. Other multiple inversion systems include those responsible for antigen variation by alternating expression of *vsa* gene variants in *Mycoplasma pulmonis*[5], shuffling outer membrane protein genes in *Dichelobacter nodosus*[3] or immune evasion in *Campylobacter fetus* by varying surface layer protein genes[6].

One of the best characterised systems is the shufflon of the pil operon on plasmid R64[7], a DNA inversion system that functions as a biological switch to select between alternative C-terminal segments of the pilV protein[8,9]. Conjugative thin pili which are required for liquid matings and the *pilV* products are tip-located adhesins of the type IV pilus that recognize lipopolysaccharides of recipient bacterial cells and thus determine recipient specificity in matings. The pil operon harbours the *rci* gene, coding for shufflon invertase *Rci*, and the adjacent shufflon made of four DNA segments separated and flanked by seven recognition sites, also known as shufflon crossover or *sfx* sites, specifically recognized by *Rci*[10]. *Rci's sfx* sites are composed of a 12bp left arm, a 7bp core or spacer sequence, and a 12bp right arm. Fully conserved nucleotides are found only in the core and right arm whereas the sequence of the left arm is highly variable[10]. It has been proposed that this asymmetry predetermines *Rci* binding and limits its activity to perform inversions only. The model suggests that *Rci* bound to the right arm recruits a second *Rci* monomer to bind to the left arm sequence nonspecifically and is supported by experiments with artificially symmetric *sfx* sites that yield both inversions and excisions events. This cooperative binding is thought to be mediated by an additional C-terminal domain not present in related Tyrosine recombinases like Cre[11] which perform both excisions and inversions.

It has been demonstrated that *Rci*-mediated shuffling in *E. coli* requires no co-factors making it an attractive synthetic biology tool for sequence diversification in different contexts[8]. A proof-of-principle study for the use of DNA inversion systems to enable *in vivo* genetic barcoding combined *Rci* with artificial 5-module and 11-module shufflons yielding a theoretical maximum of 384 and 176,947,200 barcodes respectively[12]. The authors used long-read sequencing to explore the unique barcodes generated by this system and demonstrated *Rci*'s potential for generating a large amount of sequence diversity within a synthetic construct. Whilst the literature thus suggests that *Rci* is capable of shuffling synthetic constructs in *E. coli*, its efficiency in doing so is inconsistent

and requires long incubation periods[8,12]. *Rci* has also been demonstrated to show low activity in eukaryotes[13,14], including the occurrence of deletions not previously observed in studies on wildtype *Rci*. Recently an approach that employed directed evolution managed to generate *Rci* mutants with a higher inversion frequency and a negligible rate of deletion in eukaryotes including human HEK293 cells[14]. The study did not test how these *Rci* mutants behave in the bacterial context (Y. Wu, personal communication).
Together this suggests that the availability of a greater range of SIs and variant *sfx* sites could provide more enzymes suitable for particular applications, contexts, and organisms and additional starting points for evolving enzymes for custom synthetic biology applications. It could also provide the starting point for the establishment of orthogonal shuffling systems or yield SI enzymes with higher rates of activity especially in the context of larger synthetic shufflons. We thus sought to identify variant SI enzymes homologous to *Rci* recognizing differing *sfx* sites and to establish a reporter system that allows us to compare their shuffling capabilities.

Results and Discussion

To identify a set of novel SI enzymes we relied on the fact that in natural DNA inversion systems SI genes and shufflons are typically co-located in close proximity. To that end, we took an iterative DNA sequence mining approach (Figure 1A) starting with the amino acid sequence of *Rci* (WP_001139955) in a query against predicted DNA translations of sequences in the NCBI core nucleotide database 'core_nt'. We extracted putative SI protein sequences as well as up to 5kb of the DNA regions flanking each genomic or plasmid hit. Flanking DNA sequences were then screened for 13-mers which occurred multiple times or occurred at least twice as inverted repeats. This is based on the longest conserved subsequence within *Rci's* 31-nt *sfx* site of 13 nucleotides[10]. Organisms with hits lacking identifiable repeats or hits with repeats identical to the conserved portion of the original *sfx* site, were excluded from the database in the next iterative searches. The protein sequences of SIs featuring *sfx* repeat sequences non-identical to *Rci*'s *sfx* site were then used as seeds to initiate additional search iterations. In total, out of 314 DNA loci which share translated sequence homology with *Rci,* 94 contained at least two 13-nt direct repeats, of which 56 also had at least one inverted repeat. Of those, 14 were non-identical to *Rci's sfx* site. Two homologs with palindromic repeats were not considered due to the lack of asymmetry in their cognate recognition sites. For simplicity, here we named *Rci* homologs according to their species of origin in the following format: SI-species name. Of the 14 homologs, one was discarded since its protein sequence was significantly shorter than the other SIs, and one because its sequence was identical to SI-E.E76. In addition to these 12, two further homologs had already been identified in a prior database screen (*SI-H.paralvei* and *SI-Y.pseudotuberculosis*). Furthermore, we included three *Rci* mutants from Han, P. et al. (2021). Mutant m8 was reported to have the highest inversion to deletion ratio in *Saccharomyces cerevisiae,* while mutants m18 and m34 were reported to have the highest inversion rates in the same organism[14]. Thus, a total of 17 SIs were evaluated alongside the original *Rci* in our experimental system (Table 1, Supplementary File 1). A single *sfx* site was defined for each enzyme picking, where possible, the most common sequence variant for the variable left arm or by randomly choosing the variable arm from a single *sfx* repeat.

Next, we established a standardised system for comparing Rci's shuffling capabilities with the other 17 SIs. It has been reported that in *E. coli* expressing *Rci* from low copy number plasmids prevents protein aggregation[12]. Our initial attempts also showed that the *rci* gene or its promoter acquire mutations when *Rci* is overexpressed from high-copy vectors suggesting some level of toxicity (data not shown). All synthesised SI's open reading frames were therefore cloned into a low copy vector (Figure 1B) with the p15a origin of replication under the transcriptional control of the pLacIQ promoter (Registry of Standard Biological Parts: BBa_K1695000). However, out of the 18 assembled SIs three (*SI-P.punjabense*, *SI-P.parmentieri*, and *SI-K.radicincitans*) were found to accumulate mutations even in the low-copy expression vector as evident during standard passaging and plasmid sequencing. They were therefore cloned into an L-arabinose-inducible[15] expression vector (Figure 1B) alongside *Rci*.

Next, we assembled, for each *sfx* variant a reporter plasmid harbouring a five-module shufflon of "dummy" sequences of increasing length (143 to 521bp) flanked by six recombination sites in alternating directions (Figure 1B). Since shuffling of 100 bp fragments with *Rci* had been successfully demonstrated before[12], we chose our minimum length to be close to that value. Each module can be inverted in place and it can also occupy any other position of the same parity: even modules can occupy any even position, odd modules can only occupy any odd positions for a total of 384 possible permutations.

Figure 1C illustrates the experimental workflow for our shuffling assay. First, each SI was co-transformed with its cognate reporter into Stbl3 *E. coli* cells followed by liquid culture both in the presence of ampicillin and chloramphenicol to select for both reporter and SI expression plasmids. For constitutively expressed SIs, the cultures were incubated for 24 hours (Figure 1C) following Peikon et al. who demonstrated Rci shuffling of multi-modular cassettes after overnight incubation of their *E. coli* liquid culture[12]. For inducibly expressed SIs, cultures were first incubated overnight and then used to inoculate fresh cultures at an optical density OD600 of 0.1 units. After roughly two doublings, L-arabinose was added for a final concentration of 0.2% upon which the cultures were incubated for a further 5 hours. The plasmid DNA library was then extracted from each culture and subjected to a restriction digest to isolate the shufflons. The DNA shufflon libraries were then barcoded for pooled single nucleotide sequencing using the Oxford Nanopore Technologies (ONT) MinION system. Following basecalling and deconvolution, each library was mapped against the 384 possible reference permutations that can be generated by each reporter shufflon of five modules. For further details describing the bioinformatic analysis, see Materials & Methods.

Figure 2A shows the mean shuffling rates, i.e. the fraction of shufflons not in the original (i.e. starting) configuration over two biological replicates for each constitutively-expressed SI tested. Five SIs produced at least one replicate with a shuffling rate above 50%: wild type *Rci*, homologs *SI-E. E76* and *SI-B.ubonensis*, and *Rci* mutants m18 and m34. Of those, wild type *Rci* displayed the highest mean shuffling rate of 97% followed by *SI-E.E76* and *SI-B.ubonensis* with 79% and 74% mean shuffling rates respectively. All other homologs exhibited either no shuffling or a minimal amount of up to about 7%. Of the three *Rci* mutants, mutant m8 yielded no shuffling whilst mutant m18 yielded the highest mean shuffling rate. Figure 2B shows the module occupancy after shuffling for those SIs with an appreciable shuffling rate as well as *SI-P.citronellolis* as an example of an unshuffled library. We observed a previously reported bias, namely that terminal modules are less frequently shuffled than central modules flanked by multiple recombination sites. The

shuffling rates of homologs *SI-E.E76* and *SI-B.ubonensis* did not exhibit statistically significant differences in pairwise comparisons to wildtype *Rci* (Dunnett's C test). However, a closer look at the reporter libraries generated shows that these SI's also generated fewer library members with more than 1 inversion (Supplementary Figure 1A) and the libraries covered a smaller share of the total diversity possible (384) in an analysis of 4000 randomly selected library members per SI (Supplementary Figure 1B). Interestingly, even though *SI-E. E76* yielded a higher mean shuffling rate than *SI-B. ubonensis*, the *SI-B.ubonensis* library displays more evenly distributed module occupancies in every position (Figure 2B). It also yielded 259 out of 384 permutations of its reporter shufflon while *SI-E.E76* yielded only 217 permutations (Supplementary Figure 1B). Similarly, despite *Rci* m18's lower mean shuffling rate compared *SI-E.E76* and *SI-B.ubonensis*, and obvious biases in module occupancy (Figure 2B), it generated 293 different permutations. These findings suggest that a high shuffling rate may not always reflect a library's diversity. Wild type *Rci* library has the most evenly distributed populations in its reporter library's shufflon positions (Figure 2B). Among the 4000 randomly sampled library members from both biological replicates, 373 out of 384 permutations were identified (Supplementary Figure 1B). Therefore, among the constitutively expressed SIs, it appears to be the most active SI and generated the most diverse library.

Next, we aimed to determine whether orthogonality between SIs existed using our set of active enzymes and reporters with different *sfx* variants. An orthogonal system of SIs would be a useful addition to the toolkit, as it could provide precise spatial and temporal control of SI-mediated inversions or could be used to establish dual index barcodes shuffled by different SIs at different times. For this purpose *Rci*, and the two most active homologs, *SI-E.E76* and *SI-B.ubonensis,* were tested for shuffling each other's reporters, as well as for shuffling their cognate reporters as controls. Additionally, they were tested for shuffling reporters with *sfx* sites most dissimilar to their own. The result of this experiment is shown in Figure 3. Overall, shuffling rates were lower for all 3 SI's tested in this experiment with *SI-B.ubonensis* showing the most substantial decrease of shuffling its own reporter (by 49% on average). *SI-B.ubonensis* was the only one out of the three SIs that did not shuffle any of the non-cognate reporters at an appreciable rate (Figure 3A). Surprisingly, *SI-E.E76* exhibited high shuffling activity with 3 different non-cognate reporters, in all three cases surpassing the rate for its cognate reporter. There is no statistical difference between mean shuffling rates of *SI-E.E76* with *E.cloacae* and *B.ubonensis* reporters compared to *Rci*'s shuffling rate for its cognate reporter (two-sample t-tests with Welch's correction) which is the combination that had previously yielded the most diverse reporter library. This is supported by the occupancy graphs across the reporter shufflons for both *E.cloacae* and *B.ubonensis* (Figure 3B). Whilst there is a statistically significant difference between the mean shuffling rates of *Rci* and *SI-E.E76* for various reporters, our analysis did not yield a fully orthogonal pair of SIs in combination with reporters harbouring different *sfx* variants.

The results of the experiments with inducibly expressed SIs are summarised in Figure 4. In accordance with expectations, *Rci* induced no significant shuffling of its reporter in the absence of L-arabinose in the culture. Following induction with L-arabinose, the mean shuffling rate observed after 5 hours was 42.1% for *Rci*. The other 3 SI variants also generated shuffling of their cognate reporters. 2 of them, *SI-P.punjabense* and *SI-K.radicincitans,* produced mean shuffling rates of 89.8% and 93.2% respectively for their cognate reporters, significantly outperforming *Rci* under these conditions (Dunnett C test for pairwise comparisons, Figure 4A). *SI-K.radincincitans*

in particular displayed more balanced module occupancy (Figure 4B) and generated the most diverse library (370 out of 384 possible shufflon permutations) within the 4000 randomly sampled library members (Supplementary Figure 1B) compared to *Rci* (95 out of 384).

It has previously been demonstrated that the 563 bp long segment A of the original R64 shufflon had the highest inversion frequency. Shortening it by about 100 bp and 250 bp had no significant impact on the inversion frequency, while increasing the sequence length by approximately 2.5 times as well as 3.5 times reduced its inversion frequency[8]. We therefore investigated the impact of module length on SI inversion frequency in our experiments. We considered the data of all five constitutively expressed SIs that we found to be active as well as all inducibly expressed SIs. From each library we exclusively considered reads which represented the original reference order of modules and quantified the inverted and uninverted configurations of each module. The reason for this selection was to consider library members which were likely the products of only a single module inversion. A multiple regression analysis was carried out to assess the effect of SI and module length on the module inversion frequency across SI's (Supplementary Figure 2). It suggests a decline in inversion frequency with increasing module length (module length being a significant coefficient $p<0.00361$) and confirms the expected significant effect of the SI employed ($p<1.484\times10^{-9}$) on inversion efficiency.

In this study we identified and tested SIs homologous to *Rci*. We established a standardised system for comparing the recombination or shuffling capabilities of these enzymes when paired with their respective recombination sites based on the ONT sequencing platform. Two such invertases, *SI-K.radicincitans* and *SI-P.punjabense*, were found to outperform *Rci* when expressed inducibly. Our findings thus expand the toolbox for shufflon-like systems in synthetic biology. *SI-K.radicincitans* is the most active shufflon invertase tested so far. It is able to randomise a 5 module shufflon during a 5-hour incubation and it recognizes an *sfx* site that differs from that of *Rci* at two nucleotide positions in both the core and right arm sequences. However, as our experiments on orthogonality demonstrate, *sfx* site variants can have a large effect on shuffling rates. Homolog *SI-E.E76* was found to exhibit a comparable shuffling activity to *Rci* with two different non-cognate *sfx* sites belonging to *SI-E.cloacae* and *SI-B.ubonensis*. Thus, while for each new SI we have based our choice of *sfx* site on the most common site found within the respective shufflon, this does not rule out that other variant recombination sites could lead to higher shuffling rates. Also, while we have tested all SI's at 37°C some homologs might have different optimal working conditions related to their host organism. For example, *P.parmentieri* does not survive temperatures above 33°C.[16] *H. paralvei* grows optimally at 30°C and pH 6.0[17]. To further increase the performance of the shuffling systems we describe, the parameter space around *sfx* site variants and conditions requires further mapping. The enzyme kinetics and the exact inversion mechanism of even the original *Rci* have not been fully elucidated. It is also still unclear how exactly the left arm of an sfx site impacts the inversion frequency. The standardised reporter assay, SIs and *sfx* variants we describe here will aid this effort. Further performance improvements may be required when one considers the shuffling of synthetic shufflons with dozens or even hundreds of modules where the ratio of *sfx* binding sites to active SI molecules may be shifted unfavourably. The performance of the enzymes we describe here should also be evaluated in eukaryotes, given that *Rci* mutants show a somewhat reduced performance in bacteria in our assays. In summary, *SI-SI-K. radicincitans*, *SI-P. punjabense*, *SI-E. E76*, and *SI-B.ubonensis* are the only SIs in the literature which can perform *in vivo* shuffling in *E. coli* under

standard growth conditions and without the need of any co-factors, performing like *Rci* or significantly better. Therefore, this study expands the invertase toolkit for lossless recombination employed to generate sequence diversity. It also paves the way for further discovery and testing of other such invertases which could eventually lead to establishing a set of fully orthogonal SI systems.

Materials & Methods

Plasmid construction

The sequence of all expression and reporter plasmids is provided in Supplementary File 2 and all primers used for cloning are provided in Supplementary Table 1. Briefly, *Rci* and its homologs were obtained by gene synthesis (ATUM, Newark, CA) in the low-copy Electra MOTHER™ pM279 backbone including an AatII restriction site and pLacIQ promoter upstream, and an XhoI restriction site and T1 terminator downstream. For inducible expression, the plasmid backbone was first obtained via a restriction digest of one of the constitutive expression plasmids using XhoI and AatII restriction enzymes which removed the pLacIq promoter and the SI. The L-arabinose operon was then PCR-amplified using the 'AraC-pBAD FWD' and 'AraC-pBAD REV' primers. *Rci*, and homologs *SI-K. radicincitans*, *SI-P. parmentieri* and *SI-P. punjabense* – which were delivered by ATUM as PCR products, were PCR-amplified using the forward primer 'Homolog FWD universal' and the reverse primers 'Inducible Rci EXP REV', 'K. rad. EXP REV', and 'P. parm. & P. punj. EXP REV' respectively. Tripart Gibson assembly was then carried out using the backbone, the L-arabinose operon, and the SI fragments. For the reporter plasmids, five dummy sequences were PCR-amplified from the CDS of the transcription factor p65 activation domain, the Epstein Barr virus R transactivator activation domain, and eGFP in an expression vector in the Windbichler lab. Each variant *sfx* site was introduced via overhangs of these primers. Modules were PCR-amplified with the following primer pairs: SI-specific FWD1 and SI-specific REV1 for module 1; 'Universal FWD2' and SI-specific REV2 for module 2; 'Universal FWD3' and SI-specific REV3 for module 3; 'Universal FWD4' and SI-specific REV4 for module 4; 'Universal FWD5' and SI-specific REV5 for module 5 (Supplementary Table 1). Each set of five modules was then Gibson-assembled into a cloning vector encoding an ampicillin resistance marker, ColE1 origin of replication and ROP gene.

Shuffling assay

One Shot™ Stbl3™ Chemically Competent E. coli (Invitrogen) cells were co-transformed with 3 ng DNA of both expression and reporter plasmids for an average molar ratio of 1:1.7, following the dedicated heat-shock protocol. After cell recovery, 6.8 mL of LB containing 100 µg/mL ampicillin and 34 µg/mL chloramphenicol were inoculated with the total volume of transformed cells, and incubated for 24 hours in a shaking incubator, at 37°C and 225 rpm. For inducible SIs, overnight cultures were used to seed a fresh culture at an OD600 of 0.1 which was incubated as before until an OD600 of 0.5. L-arabinose was then added to each culture for a final concentration of 0.2% followed by another 5 hour incubation. DNA was extracted using the QIAprep Spin Miniprep Kit (QIAGEN) with the following modifications. P1, P2, and N3 reagent volumes were doubled, and the cell lysis was thus carried out in 2 mL Eppendorf tubes. After the 10 min spin, the supernatant was loaded in roughly even parts on two columns. Finally, the DNA was eluted in 50 µL 53°C DNase/RNase-free Distilled water (ddH_2O) from each column, and the two eluates were subsequently pooled to yield roughly 100 µL of DNA per sample.

Library preparation and nanopore sequencing

Plasmid DNA was digested using NdeI (NEB) and following gel electrophoresis the 4.5kb band corresponding to the shufflon was extracted for each sample and gel-purified using Monarch®

DNA Gel Extraction Kit (NEB). 200 fmol of linearised DNA per condition were barcoded using the Native Barcoding Kit 24 (Q20+) (SQK-NBD112.24, Oxford Nanopore Technologies). Sequencing was performed using a Flongle Flow Cell R9.4.1 (FLO-FLG001), and the MinION Mk1B device. 20 fmol of the pooled library was sequenced until 150,000-200,000 reads were obtained. Basecalling was performed using Guppy[18] via the Imperial College Research Computing Service[20]. Using the super accurate model (dna_r9.4.1_450bps_sup.cfg), deconvoluting reads according to barcodes (barcode_kits SQK-NBD112-24) and with trim_barcodes and trim_adapters flags on.

Bioinformatic analysis

The analysis pipeline is found here https://gitfront.io/r/jkatalinic/xf8oSaYQpkmC/Rci-Variants/. Briefly, for each reporter the set of 384 references was generated including ~2 kb anchor sequences flanking the shufflon. Sequencing reads were size filtered using Seqkit[19] for DNA of the approximate size of the shufflon and anchor sequences and mapped against the reference set using Minimap2[21]. Samtools[22] was used to generate one compressed .bam file per reference which is filtered for alignment quality (AS score of >7500)[21]. From each filtered .bam file, a text file is generated that reflects the order of modules for each read. These files are then compiled and analysed in R.

Statistics

All tests were conducted in R. For comparing the shuffling rates of Rci and its homologs and mutants, Welch's ANOVA was performed for a one-way analysis of means of unequal variance[24] to determine if the shuffling rate means are different. Unequal variance was confirmed using Bartlett test of homogeneity of variances[25]. Dunnett's C test was performed for pairwise comparisons to a control, in this case Rci, in order to determine which SIs produced a mean shuffling rate of statistically insignificant difference to Rci. For assessing the correlation between module inversion frequency, module length and SI, a regression analysis was carried out in R using a linear model. For the test of orthogonality we used a two sample t-test with Welch's correction since unequal variance was indicated by the difference in standard deviations between groups.

Figure 1

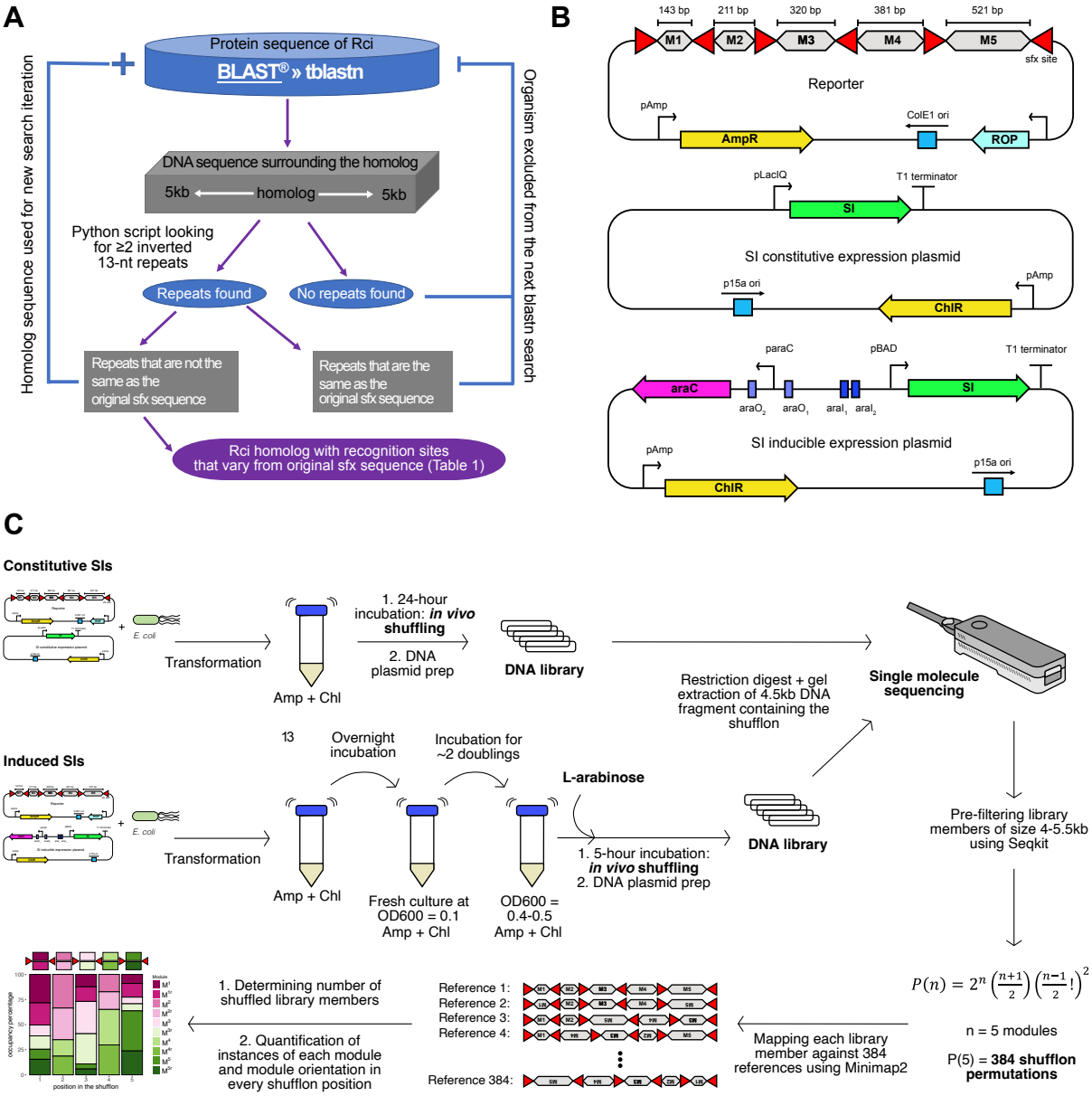
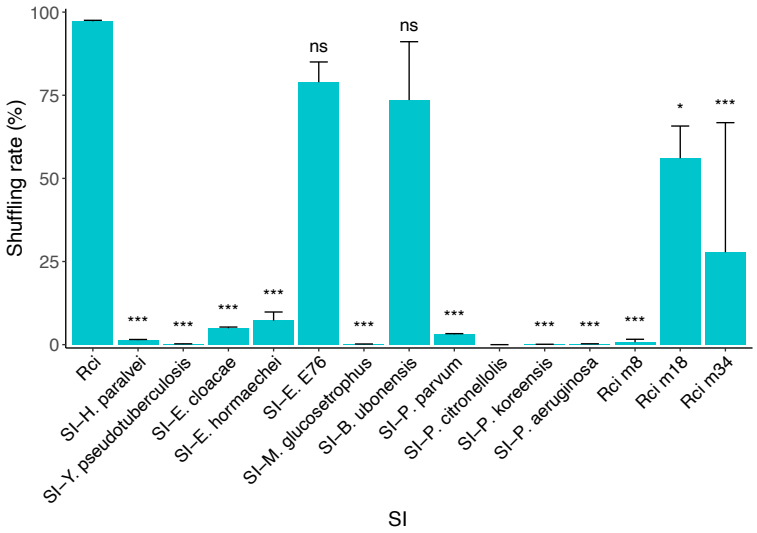

A
Protein sequence of Rci
BLAST® » tblastn
DNA sequence surrounding the homolog
5kb
homolog
5kb
Python script looking for ≥2 inverted 13-nt repeats
Repeats found
No repeats found
Repeats that are not the same as the original sfx sequence
Repeats that are the same as the original sfx sequence
Rci homolog with recognition sites that vary from original sfx sequence (Table 1)
Homolog sequence used for new search iteration
Organism excluded from the next blastn search
B
143 bp
211 bp
320 bp
381 bp
521 bp
M1
M2
M3
M4
M5
sfx site
Reporter
pAmp
ColE1 ori
AmpR
ROP
pLacIQ
T1 terminator
SI
SI constitutive expression plasmid
p15a ori
pAmp
ChlR
paraC
pBAD
T1 terminator
araC
SI
SI inducible expression plasmid
pAmp
p15a ori
ChlR
C
Constitutive SIs
E. coli
Transformation
Amp + Chl
1. 24-hour incubation: in vivo shuffling
2. DNA plasmid prep
DNA library
Restriction digest + gel extraction of 4.5kb DNA fragment containing the shufflon
Single molecule sequencing
Induced SIs
13
Overnight incubation
Incubation for ~2 doublings
L-arabinose
E. coli
Transformation
Amp + Chl
Fresh culture at OD600 = 0.1 Amp + Chl
OD600 = 0.4-0.5 Amp + Chl
1. 5-hour incubation: in vivo shuffling
2. DNA plasmid prep
DNA library
Pre-filtering library members of size 4-5.5kb using Seqkit
$P(n) = 2^n \left(\frac{n+1}{2}\right)\left(\frac{n-1}{2}!\right)^2$
n = 5 modules
P(5) = 384 shufflon permutations
Mapping each library member against 384 references using Minimap2
Reference 1:
Reference 2:
Reference 3:
Reference 4:
Reference 384:
1. Determining number of shuffled library members
2. Quantification of instances of each module and module orientation in every shufflon position
position in the shufflon

Figure 2

A


Shuffling rate (%)
100
75
50
25
0
ns
ns
*
***
***
Rci
SI–H. paralvei
SI–Y. pseudotuberculosis
SI–E. cloacae
SI–E. hormaechei
SI–E. E76
SI–M. glucosetrophus
SI–B. ubonensis
SI–P. parvum
SI–P. citronellolis
SI–P. koreensis
SI–P. aeruginosa
Rci m8
Rci m18
Rci m34
SI

B

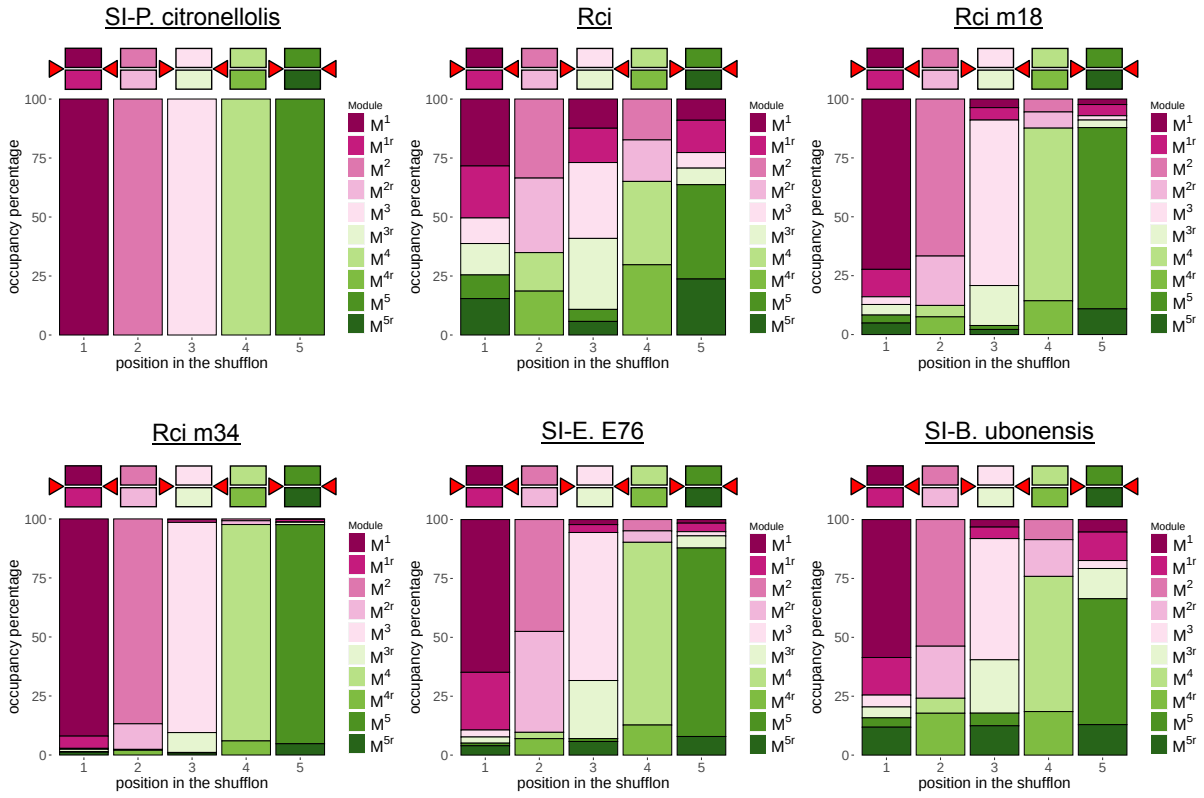
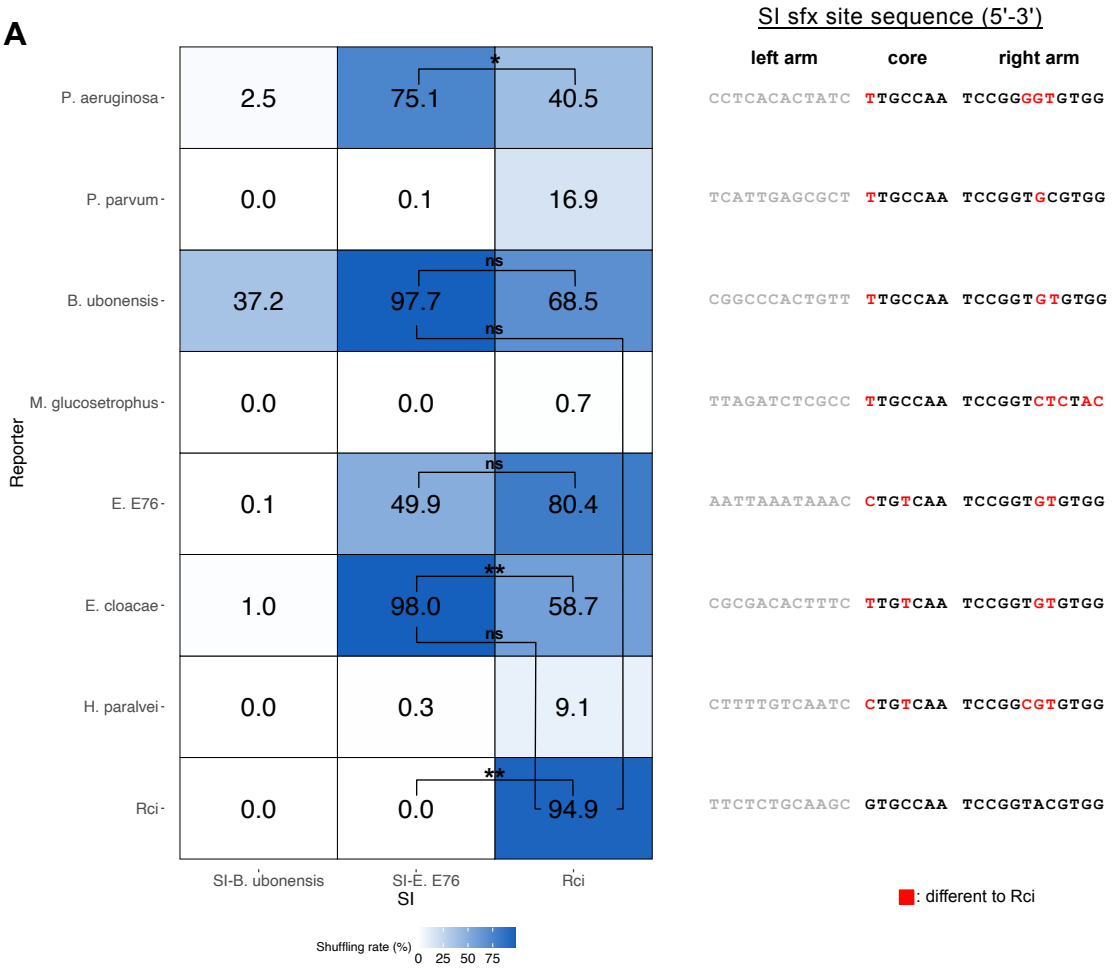

SI-P. citronellolis
Rci
Rci m18
Rci m34
SI-E. E76
SI-B. ubonensis
occupancy percentage
position in the shufflon
Module
M1
M1r
M2
M2r
M3
M3r
M4
M4r
M5
M5r

Figure 3

A


SI sfx site sequence (5'-3')
left arm
core
right arm
Reporter
P. aeruginosa
P. parvum
B. ubonensis
M. glucosetrophus
E. E76
E. cloacae
H. paralvei
Rci
2.5
75.1
40.5
0.0
0.1
16.9
37.2
97.7
68.5
0.0
0.0
0.7
0.1
49.9
80.4
1.0
98.0
58.7
0.0
0.3
9.1
0.0
0.0
94.9
*
ns
ns
ns
**
ns
**
CCTCACACTATC TTGCCAA TCCGGGGTGTGG
TCATTGAGCGCT TTGCCAA TCCGGTGCGTGG
CGGCCCACTGTT TTGCCAA TCCGGTGTGTGG
TTAGATCTCGCC TTGCCAA TCCGGTCTCTAC
AATTAAATAAAC CTGTCAA TCCGGTGTGTGG
CGCGACACTTTC TTGTCAA TCCGGTGTGTGG
CTTTTGTCAATC CTGTCAA TCCGGCGTGTGG
TTCTCTGCAAGC GTGCCAA TCCGGTACGTGG
SI-B. ubonensis
SI-E. E76
Rci
SI
: different to Rci
Shuffling rate (%)
0 25 50 75

B

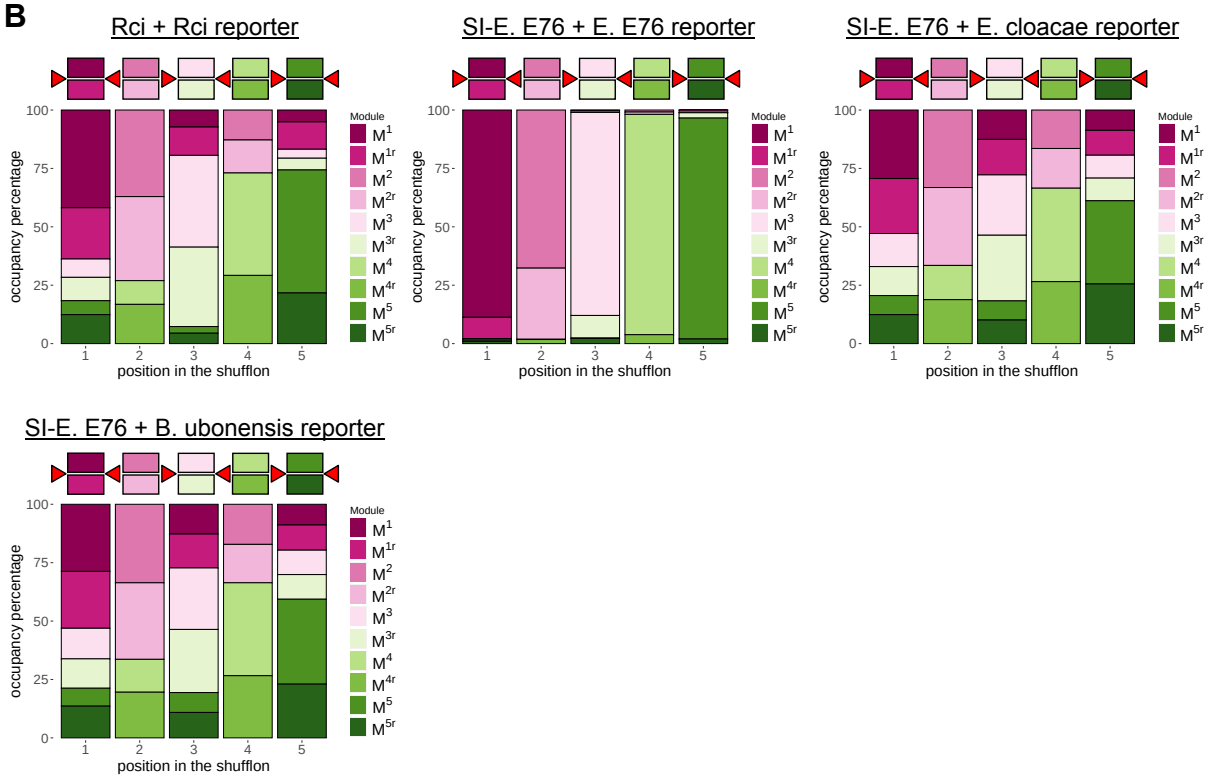
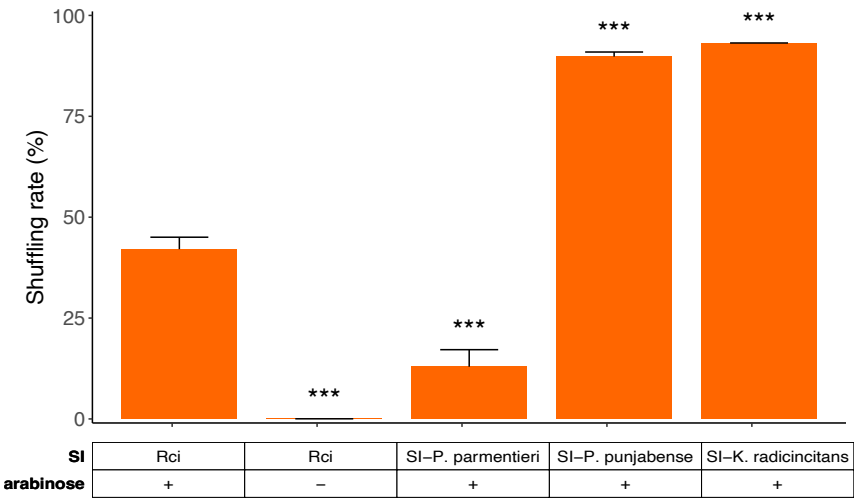

Rci + Rci reporter
SI-E. E76 + E. E76 reporter
SI-E. E76 + E. cloacae reporter
SI-E. E76 + B. ubonensis reporter
occupancy percentage
position in the shufflon
Module
M1
M1r
M2
M2r
M3
M3r
M4
M4r
M5
M5r

Figure 4

A


Shuffling rate (%)
100
75
50
25
0
***
***
***
***
SI
arabinose
Rci
Rci
SI–P. parmentieri
SI–P. punjabense
SI–K. radicincitans
+
–
+
+
+

B

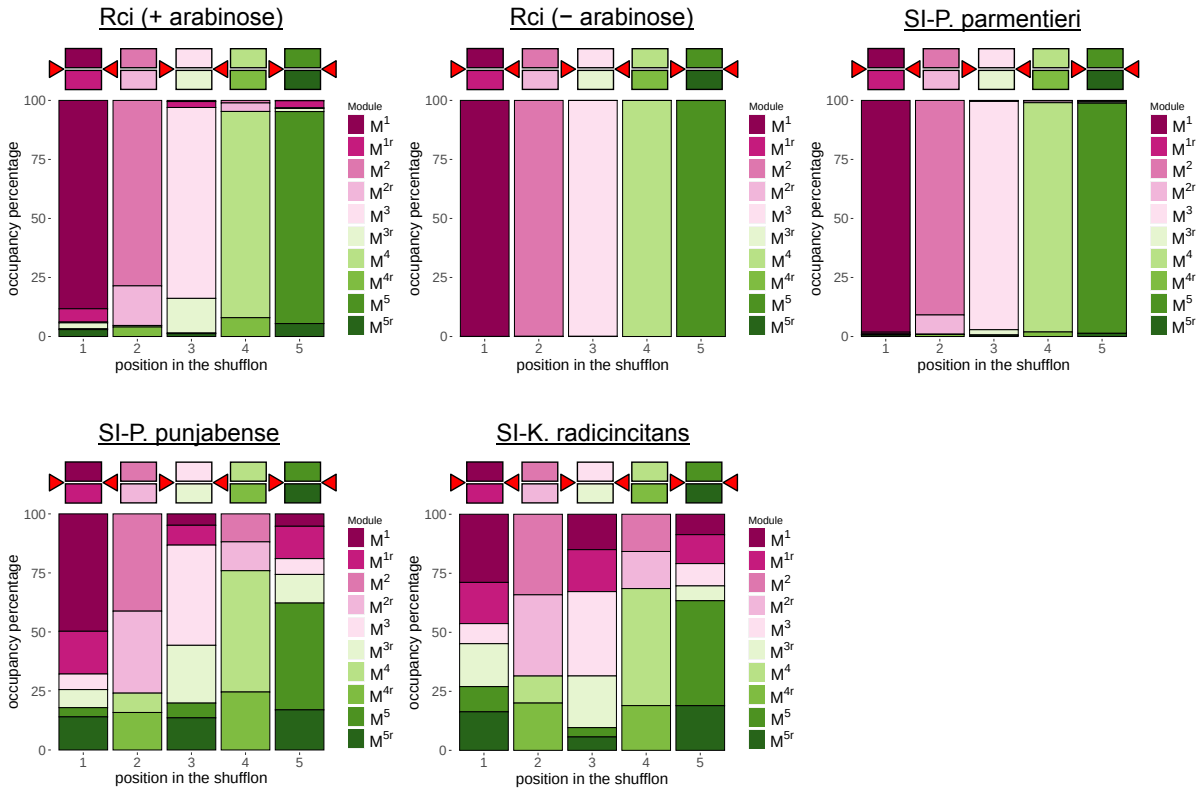
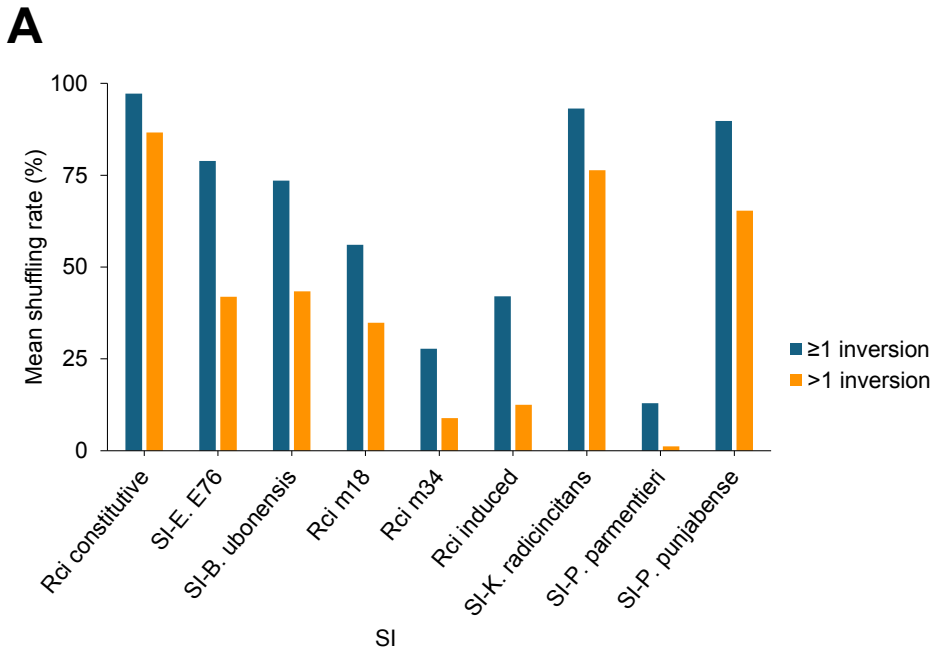

Rci (+ arabinose)
Rci (− arabinose)
SI-P. parmentieri
SI-P. punjabense
SI-K. radicincitans
occupancy percentage
position in the shufflon
Module
M1
M1r
M2
M2r
M3
M3r
M4
M4r
M5
M5r

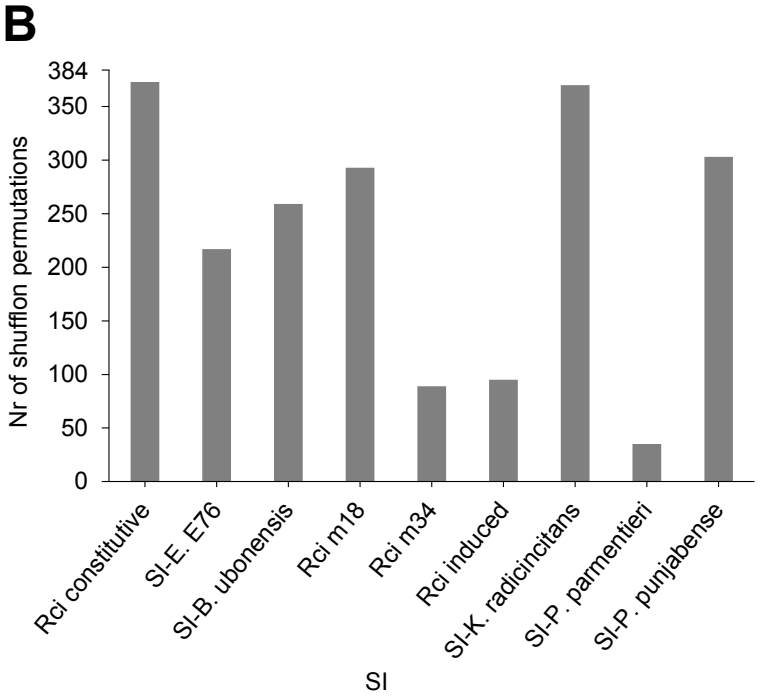
Table 1: *Tested SIs and their cognate sfx site sequences*

SI name	Origin of SI	Accession nr	Protein sequence homology to Rci (%)	Sfx site sequence (5'-3') Left arm	Core	Right arm	Nr of sfx sites at locus of origin	Gap between inverted repeats closest to SI gene
Rci*	*Salmonella enterica* subsp. *enterica* serovar Typhimurium	AP005147.1	NA	TTCTCTGCAAGC	GTGCCAA	TCCGGTACGTGG	7	499
Rci m8	*Salmonella enterica* subsp. *enterica* serovar Typhimurium	DOI: 10.1021/acssynbio.1c00132	99.2	TTCTCTGCAAGC	GTGCCAA	TCCGGTACGTGG	7	499
Rci m18	*Salmonella enterica* subsp. *enterica* serovar Typhimurium	DOI: 10.1021/acssynbio.1c00132	99.2	TTCTCTGCAAGC	GTGCCAA	TCCGGTACGTGG	7	499
Rci m34	*Salmonella enterica* subsp. *enterica* serovar Typhimurium	DOI: 10.1021/acssynbio.1c00132	99.2	TTCTCTGCAAGC	GTGCCAA	TCCGGTACGTGG	7	499
SI-P. parmentieri*	*Pectobacterium parmentieri* WPP163 chromosome	CP001790.1	63.0	CTTGTGAGCACC	TTGCCAA	TCCGGTGCGTGG	7	2140
SI-P. punjabense*	*Pectobacterium punjabense* strain SS95 chromosome	CP038498.1	62.8	TGCAATACTTTC	TTGCCAA	TCCGGTGCGTGG	5	635
SI-P. parvum	*Pectobacterium parvum* strain PZ1	CP046377.1	62.5	TCATTGAGCGCT	TTGCCAA	TCCGGTGCGTGG	5	976
SI-K. radicincitans*	*Kosakonia radicincitans* DSM 16656	CP018016.1	58.1	CGCAATACTTTC	CTGTCAA	TCCGGTGTGTGG	4	966
SI-E. E76	*Enterobacter* sp. *E76*	CP042499.1	57.9	AATTAAATAAAC	CTGTCAA	TCCGGTGTGTGG	6	1353
SI-H. paralvei	*Hafnia paralvei* strain AVS0177	CP083738.1	56.8	CTTTTGTCAATC	CTGTCAA	TCCGGCGTGTGG	2	775
SI-Y. pseudotuberculosis	*Yersinia pseudotuberculosis* strain 598	CP071946.1	51.3	CATTTGTCAATC	TTGTCAA	TCCGGTGTGTGG	3	814
SI-P. aeruginosa	*Pseudomonas aeruginosa strain* PA15W	MN961672.1	49.9	CCTCACACTATC	TTGCCAA	TCCGGGGTGTGG	7	2126
SI-P. citronellolis	*Pseudomonas citronellolis* strain SJTE-3	CP015879.1	48.9	CGTAATACTTTC	TTGCCAA	TCCGGTGTTTGG	6	955
SI-M. glucosetrophus	*Methylovorus glucosetrophus* SIP3-4	CP001675.1	47.4	TTAGATCTCGCC	TTGCCAA	TCCGGTCTCTAC	4	1699
SI-B. ubonensis	*Burkholderia ubonensis* subsp. *mesacidophila* strain ATCC 31433	CP020738.1	47.1	CGGCCCACTGTT	TTGCCAA	TCCGGTGTGTGG	3	1495
SI-P. koreensis	*Pseudomonas koreensis* strain P19E3	CP027479.1	46.2	AAAGCGCGTGCC	TTGCCAA	TCCGGTGTGTGG	16	2169
SI-E. cloacae	*Enterobacter cloacae* ECNIH3	CP008900.1	44.3	CGCGACACTTTC	TTGTCAA	TCCGGTGTGTGG	3	738
SI-E. hormaechei	*Enterobacter hormaechei* subsp. *xiangfangensis* strain WCHEX045001	CP043384.1	44.3	CGCGACACTTTC	TTGTCAA	TCCGGTGTGTGG	3	738

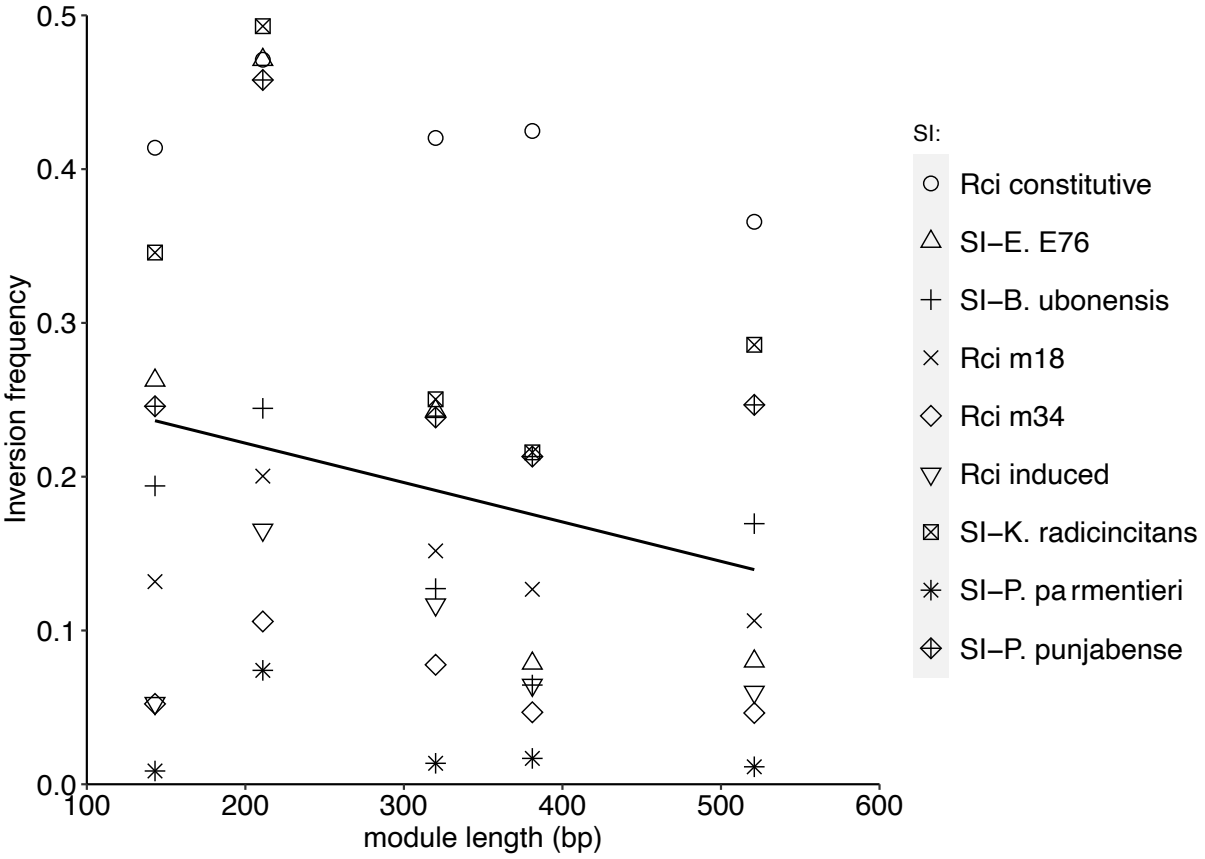
The sfx site sequences were used to assemble each SI's respective reporter shufflon. *: SIs which were inducibly expressed. Rci was tested for both constitutive and inducible expression.

Supplementary Figure 1





Supplementary Figure 2



Supplementary Table 1: Reporter and inducible expression vector assembly primers

Primer ID	Sequence (5' to 3')
Universal FWD2	CAGGAGCGCACCATCTTC
Universal FWD3	GAGATGGCCGATACTGTG
Universal FWD4	CGACGAGGACTTCAGCTC
Universal FWD5	CCTACGAGACATTCAAGAGC
Rci FWD1	CTAGTGGATCCAGTGTGGTGGAATTCTTCTCTGCAAGCGTGCCAATCCGGTACGTGGGGTACCGACGTAAACGGCCACAAG
Rci REV1	CTTGAAGAAGATGGTGCGCTCCTGGGTACCTTCTCTGCAAGCGTGCCAATCCGGTACGTGGTAGGTCAGGGTGGTCACG
Rci REV2	GGGAATCACAGTATCGGCCATCTCGGTACCCCACGTACCGGATTGGCACGCTTGCAGAGAATGATGCCGTTCTTCTGCTTG
Rci REV3	GCGATAGAGCTGAAGTCCTCGTCGGGTACCTTCTCTGCAAGCGTGCCAATCCGGTACGTGGCCAGGATTCTCCTCGACG
Rci REV4	TGATGCTCTTGAATGTCTCGTAGGGGTACCCCACGTACCGGATTGGCACGCTTGCAGAGAAGCTTTGACAGCCTGGCTC
Rci REV5	CCAGTGACTACATCTAGACTCGAGTTTCTCTGCAAGCGTGCCAATCCGGTACGTGGCTGCTGGAACTCGCTGTTG
H. paralvei FWD1	CTAGTGGATCCAGTGTGGTGGAATTCTTTTGTCAATCCTGTCAATCCGGCGTGTGGGGTACCGACGTAAACGGCCACAAG
H. paralvei REV1	CTTGAAGAAGATGGTGCGCTCCTGGGTACCTTTTGTCAATCCTGTCAATCCGGCGTGTGGTAGGTCAGGGTGGTCACG
H. paralvei REV2	GGGAATCACAGTATCGGCCATCTCGGTACCCCACACGCCGGATTGACAGGATTGACAAAATGATGCCGTTCTTCTGCTTG
H. paralvei REV3	GCGATAGAGCTGAAGTCCTCGTCGGGTACCTTTTGTCAATCCTGTCAATCCGGCGTGTGGCCAGGATTCTCCTCGACG
H. paralvei REV4	TGATGCTCTTGAATGTCTCGTAGGGGTACCCCACACGCCGGATTGACAGGATTGACAAAAGCTTTGACAGCCTGGCTC
H. paralvei REV5	CCAGTGACTACATCTAGACTCGAGTTTTTGTCAATCCTGTCAATCCGGCGTGTGGCTGCTGGAACTCGCTGTTG
Y. pseudo. FWD1	CTAGTGGATCCAGTGTGGTGGAATTCATTTGTCAATCTTGTCAATCCGGTGTGTGGGGTACCGACGTAAACGGCCACAAG
Y. pseudo. REV1	CTTGAAGAAGATGGTGCGCTCCTGGGTACCATTTGTCAATCTTGTCAATCCGGTGTGTGGTAGGTCAGGGTGGTCACG
Y. pseudo. REV2	GGGAATCACAGTATCGGCCATCTCGGTACCCCACACACCGGATTGACAAGATTGACAAATTGATGCCGTTCTTCTGCTTG
Y. pseudo. REV3	GCGATAGAGCTGAAGTCCTCGTCGGGTACCATTTGTCAATCTTGTCAATCCGGTGTGTGGCCAGGATTCTCCTCGACG
Y. pseudo. REV4	TGATGCTCTTGAATGTCTCGTAGGGGTACCCCACACACCGGATTGACAAGATTGACAAATGCTTTGACAGCCTGGCTC
Y. pseudo. REV5	CCAGTGACTACATCTAGACTCGAGTTTGTCAATCCTGTCAATCCGGTGTGTGGCTGCTGGAACTCGCTGTTG
E. clo/hor FWD1*	CTAGTGGATCCAGTGTGGTGGAATTCGACAAAGGCGCGACACTTTCTTGTCAATCCGGTGTGTGGGGTACCGACGTAAACGGCCACAAG
E. clo/hor REV1*	CTTGAAGAAGATGGTGCGCTCCTGGGTACCGACAAAGGCGCGACACTTTCTTGTCAATCCGGTGTGTGGTAGGTCAGGGTGGTCACG
E. clo/hor REV2*	GGGAATCACAGTATCGGCCATCTCGGTACCCCACACACCGGATTGACAAGAAAGTGTCGCGCCTTTGTCTGATGCCGTTCTTCTGCTTG
E. clo/hor REV3*	GCGATAGAGCTGAAGTCCTCGTCGGGTACCGACAAAGGCGCGACACTTTCTTGTCAATCCGGTGTGTGGCCAGGATTCTCCTCGACG
E. clo/hor REV4*	TGATGCTCTTGAATGTCTCGTAGGGGTACCCCACACACCGGATTGACAAGAAAGTGTCGCGCCTTTGTCGCTTTGACAGCCTGGCTC
E. clo/hor REV5*	CCAGTGACTACATCTAGACTCGAGTGACAAAGGCGCGACACTTTCTTGTCAATCCGGTGTGTGGCTGCTGGAACTCGCTGTTG
E. E76 FWD1	CTAGTGGATCCAGTGTGGTGGAATTCGCATTAGCAATTAAATAAACCTGTCAATCCGGTGTGTGGGGTACCGACGTAAACGGCCACAAG
E. E76 REV1	CTTGAAGAAGATGGTGCGCTCCTGGGTACCGCATTAGCAATTAAATAAACCTGTCAATCCGGTGTGTGGTAGGTCAGGGTGGTCACG
E. E76 REV2	GGGAATCACAGTATCGGCCATCTCGGTACCCCACACACCGGATTGACAGGTTTATTTAATTGCTAATGCTGATGCCGTTCTTCTGCTTG
E. E76 REV3	GCGATAGAGCTGAAGTCCTCGTCGGGTACCGCATTAGCAATTAAATAAACCTGTCAATCCGGTGTGTGGCCAGGATTCTCCTCGACG
E. E76 REV4	TGATGCTCTTGAATGTCTCGTAGGGGTACCCCACACACCGGATTGACAGGTTTATTTAATTGCTAATGCGCTTTGACAGCCTGGCTC
E. E76 REV5	CCAGTGACTACATCTAGACTCGAGTGCATTAGCAATTAAATAAACCTGTCAATCCGGTGTGTGGCTGCTGGAACTCGCTGTTG
M. glu. FWD1	CTAGTGGATCCAGTGTGGTGGAATTCTTTTGTCAATCCTGTCAATCCGGCGTGTGGGGTACCGACGTAAACGGCCACAAG
M. glu. REV1	CTTGAAGAAGATGGTGCGCTCCTGGGTACCTTTTGTCAATCCTGTCAATCCGGCGTGTGGTAGGTCAGGGTGGTCACG
M. glu. REV2	GGGAATCACAGTATCGGCCATCTCGGTACCCCACACGCCGGATTGACAGGATTGACAAAATGATGCCGTTCTTCTGCTTG
M. glu. REV3	GCGATAGAGCTGAAGTCCTCGTCGGGTACCTTTTGTCAATCCTGTCAATCCGGCGTGTGGCCAGGATTCTCCTCGACG
M. glu. REV4	TGATGCTCTTGAATGTCTCGTAGGGGTACCCCACACGCCGGATTGACAGGATTGACAAAAGCTTTGACAGCCTGGCTC
M. glu. REV5	CCAGTGACTACATCTAGACTCGAGTTTTTGTCAATCCTGTCAATCCGGCGTGTGGCTGCTGGAACTCGCTGTTG
B. ubo. FWD1	CTAGTGGATCCAGTGTGGTGGAATTCTTTTGTCAATCCTGTCAATCCGGCGTGTGGGGTACCGACGTAAACGGCCACAAG
B. ubo. REV1	CTTGAAGAAGATGGTGCGCTCCTGGGTACCTTTTGTCAATCCTGTCAATCCGGCGTGTGGTAGGTCAGGGTGGTCACG
B. ubo. REV2	GGGAATCACAGTATCGGCCATCTCGGTACCCCACACGCCGGATTGACAGGATTGACAAAATGATGCCGTTCTTCTGCTTG
B. ubo. REV3	GCGATAGAGCTGAAGTCCTCGTCGGGTACCTTTTGTCAATCCTGTCAATCCGGCGTGTGGCCAGGATTCTCCTCGACG
B. ubo. REV4	TGATGCTCTTGAATGTCTCGTAGGGGTACCCCACACGCCGGATTGACAGGATTGACAAAAGCTTTGACAGCCTGGCTC
B. ubo. REV5	CCAGTGACTACATCTAGACTCGAGTTTTTGTCAATCCTGTCAATCCGGCGTGTGGCTGCTGGAACTCGCTGTTG
P. parvum FWD1	CTAGTGGATCCAGTGTGGTGGAATTCTTTTGTCAATCCTGTCAATCCGGCGTGTGGGGTACCGACGTAAACGGCCACAAG
P. parvum REV1	CTTGAAGAAGATGGTGCGCTCCTGGGTACCTTTTGTCAATCCTGTCAATCCGGCGTGTGGTAGGTCAGGGTGGTCACG
P. parvum REV2	GGGAATCACAGTATCGGCCATCTCGGTACCCCACACGCCGGATTGACAGGATTGACAAAATGATGCCGTTCTTCTGCTTG
P. parvum REV3	GCGATAGAGCTGAAGTCCTCGTCGGGTACCTTTTGTCAATCCTGTCAATCCGGCGTGTGGCCAGGATTCTCCTCGACG
P. parvum REV4	TGATGCTCTTGAATGTCTCGTAGGGGTACCCCACACGCCGGATTGACAGGATTGACAAAAGCTTTGACAGCCTGGCTC
P. parvum REV5	CCAGTGACTACATCTAGACTCGAGTTTTTGTCAATCCTGTCAATCCGGCGTGTGGCTGCTGGAACTCGCTGTTG
P. cit. FWD1	CTAGTGGATCCAGTGTGGTGGAATTCTTTTGTCAATCCTGTCAATCCGGCGTGTGGGGTACCGACGTAAACGGCCACAAG
P. cit. REV1	CTTGAAGAAGATGGTGCGCTCCTGGGTACCTTTTGTCAATCCTGTCAATCCGGCGTGTGGTAGGTCAGGGTGGTCACG
P. cit. REV2	GGGAATCACAGTATCGGCCATCTCGGTACCCCACACGCCGGATTGACAGGATTGACAAAATGATGCCGTTCTTCTGCTTG
P. cit. REV3	GCGATAGAGCTGAAGTCCTCGTCGGGTACCTTTTGTCAATCCTGTCAATCCGGCGTGTGGCCAGGATTCTCCTCGACG
P. cit. REV4	TGATGCTCTTGAATGTCTCGTAGGGGTACCCCACACGCCGGATTGACAGGATTGACAAAAGCTTTGACAGCCTGGCTC
P. cit. REV5	CCAGTGACTACATCTAGACTCGAGTTTTTGTCAATCCTGTCAATCCGGCGTGTGGCTGCTGGAACTCGCTGTTG
P. kor. FWD1	CTAGTGGATCCAGTGTGGTGGAATTCTTTTGTCAATCCTGTCAATCCGGCGTGTGGGGTACCGACGTAAACGGCCACAAG
P. kor. REV1	CTTGAAGAAGATGGTGCGCTCCTGGGTACCTTTTGTCAATCCTGTCAATCCGGCGTGTGGTAGGTCAGGGTGGTCACG
P. kor. REV2	GGGAATCACAGTATCGGCCATCTCGGTACCCCACACGCCGGATTGACAGGATTGACAAAATGATGCCGTTCTTCTGCTTG
P. kor. REV3	GCGATAGAGCTGAAGTCCTCGTCGGGTACCTTTTGTCAATCCTGTCAATCCGGCGTGTGGCCAGGATTCTCCTCGACG
P. kor. REV4	TGATGCTCTTGAATGTCTCGTAGGGGTACCCCACACGCCGGATTGACAGGATTGACAAAAGCTTTGACAGCCTGGCTC
P. kor. REV5	CCAGTGACTACATCTAGACTCGAGTTTTTGTCAATCCTGTCAATCCGGCGTGTGGCTGCTGGAACTCGCTGTTG
P. aer. FWD1	CTAGTGGATCCAGTGTGGTGGAATTCTTTTGTCAATCCTGTCAATCCGGCGTGTGGGGTACCGACGTAAACGGCCACAAG
P. aer. REV1	CTTGAAGAAGATGGTGCGCTCCTGGGTACCTTTTGTCAATCCTGTCAATCCGGCGTGTGGTAGGTCAGGGTGGTCACG
P. aer. REV2	GGGAATCACAGTATCGGCCATCTCGGTACCCCACACGCCGGATTGACAGGATTGACAAAATGATGCCGTTCTTCTGCTTG
P. aer. REV3	GCGATAGAGCTGAAGTCCTCGTCGGGTACCTTTTGTCAATCCTGTCAATCCGGCGTGTGGCCAGGATTCTCCTCGACG
P. aer. REV4	TGATGCTCTTGAATGTCTCGTAGGGGTACCCCACACGCCGGATTGACAGGATTGACAAAAGCTTTGACAGCCTGGCTC
P. aer. REV5	CCAGTGACTACATCTAGACTCGAGTTTTTGTCAATCCTGTCAATCCGGCGTGTGGCTGCTGGAACTCGCTGTTG
P. parm. FWD1	CTAGTGGATCCAGTGTGGTGGAATTCTTGAACCGCTTGTGAGCACCTTGCCAATCCGGTGCGTGGGGTACCGACGTAAACGGCCACAAGTT
P. parm. REV1	CTTGAAGAAGATGGTGCGCTCCTGGGTACCTTGAACCGCTTGTGAGCACCTTGCCAATCCGGTGCGTGGTAGGTCAGGGTGGTCACGAG
P. parm. REV2	GGGAATCACAGTATCGGCCATCTCGGTACCCCACGCACCGGATTGGCAAGGTGCTCACAAGCGGTTCAATGATGCCGTTCTTCTGCTTG
P. parm. REV3	GCGATAGAGCTGAAGTCCTCGTCGGGTACCTTGAACCGCTTGTGAGCACCTTGCCAATCCGGTGCGTGGCCAGGATTCTCCTCGACGTC
P. parm. REV4	TGATGCTCTTGAATGTCTCGTAGGGGTACCCCACGCACCGGATTGGCAAGGTGCTCACAAGCGGTTCAAGCTTTGACAGCCTGGCTCG
P. parm. REV5	CCAGTGACTACATCTAGACTCGAGTTTGAACCGCTTGTGAGCACCTTGCCAATCCGGTGCGTGGCTGCTGGAACTCGCTGTTGT
P. punj. FWD1	CTAGTGGATCCAGTGTGGTGGAATTCAATGTTGGTGCAATACTTTCTTGCCAATCCGGTGCGTGGGGTACCGACGTAAACGGCCACAAGTT
P. punj. REV1	CTTGAAGAAGATGGTGCGCTCCTGGGTACCAATGTTGGTGCAATACTTTCTTGCCAATCCGGTGCGTGGTAGGTCAGGGTGGTCACGAG
P. punj. REV2	GGGAATCACAGTATCGGCCATCTCGGTACCCCACGCACCGGATTGGCAAGAAAGTATTGCACCAACATTTGATGCCGTTCTTCTGCTTG
P. punj. REV3	GCGATAGAGCTGAAGTCCTCGTCGGGTACCAATGTTGGTGCAATACTTTCTTGCCAATCCGGTGCGTGGCCAGGATTCTCCTCGACGTC
P. punj. REV4	TGATGCTCTTGAATGTCTCGTAGGGGTACCCCACGCACCGGATTGGCAAGAAAGTATTGCACCAACATTGCTTTGACAGCCTGGCTCG
P. punj. REV5	CCAGTGACTACATCTAGACTCGAGTAATGTTGGTGCAATACTTTCTTGCCAATCCGGTGCGTGGCTGCTGGAACTCGCTGTTGT
K. rad. FWD1	CTAGTGGATCCAGTGTGGTGGAATTCGCGTCTGGCGCAATACTTTCCTGTCAATCCGGTGTGTGGGGTACCGACGTAAACGGCCACAAGTT
K. rad. REV1	CTTGAAGAAGATGGTGCGCTCCTGGGTACCGCGTCTGGCGCAATACTTTCCTGTCAATCCGGTGTGTGGTAGGTCAGGGTGGTCACGAG
K. rad. REV2	GGGAATCACAGTATCGGCCATCTCGGTACCCCACACACCGGATTGACAGGAAAGTATTGCGCCAGACGCTGATGCCGTTCTTCTGCTTG
K. rad. REV3	GCGATAGAGCTGAAGTCCTCGTCGGGTACCGCGTCTGGCGCAATACTTTCCTGTCAATCCGGTGTGGCCAGGATTCTCCTCGACGTC
K. rad. REV4	TGATGCTCTTGAATGTCTCGTAGGGGTACCCCACACACCGGATTGACAGGAAAGTATTGCGCCAGACGCGCTTTGACAGCCTGGCTCG
K. rad. REV5	CCAGTGACTACATCTAGACTCGAGTGCGTCTGGCGCAATACTTTCCTGTCAATCCGGTGTGTGGCTGCTGGAACTCGCTGTTGT
AraC-pBAD FWD	GCCCTTCGCTGAAGACGTTTATGACAACTTGACGGCTACA
AraC-pBAD REV	CTCTCTTCCGGTATGGAGAAACAGTAGAGAGT
Homolog FWD universal	CTCCATACCGGAAGAGAGACAATTCAGGGTG
K. rad. EXP REV	GTTTTATTTGATGCCTCTACTTAGGTACCGTTGTCGGTACC
P. parm. & P. punj. EXP REV	GTTTTATTTGATGCCTCTACTTACAGCGGAGAGATCAGCAG
Inducible Rci EXP REV	GTTTTATTTGATGCCTCTACTTACAGCGGTGTGCTG

*: SI-E. cloacae and SI-E. hormaechei have identical sfx sites.

Figure Legends

Figure 1. Experimental overview. A. Bioinformatic strategy for the isolation of shufflon invertases (SIs) and their recognition sites based on homology to *Rci*. **B.** Structure of the reporter (top) as well as the constitutive (middle) and inducible (bottom) expression plasmids used in this study. AmpR: ampicillin resistance gene. ROP: repressor of primer. ChlR: chloramphenicol resistance gene. M1 to M5 represent the invertible dummy modules. **C.** Experimental workflow for both inducible and constitutively expressed SIs.

Figure 2. Shuffling of reporters by constitutively expressed SIs. A. Bar graphs showing the mean shuffling rate between two biological replicates. Each estimate is based on a minimum of 1500 analysed reads. Error bars represent standard deviation (SD). Following ANOVA, Dunnett's C test was performed for pairwise comparisons to *Rci*. p-values are represented by *, $0.05 \geq p > 0.01$; **, $0.01 \geq p > 0.001$; ***, $0.001 \geq p$; 'ns', non-significant. **B.** Indicative module occupancy in the shuffled library of selected SIs. The diagram above each plot represents the shufflon in its reference (unshuffled) form. Coloured squares represent the dummy modules when present on sense or antisense strands and red triangles indicate the *sfx* sites. For each panel, 6000 library members were randomly sampled and pooled from two biological replicates.

Figure 3. Testing constitutively expressed SIs for orthogonality and cross-recognition. A. Each panel displays the mean shuffling rate from two biological replicates for each corresponding SI + reporter pair. An average of 6093 reads were analysed. Two sample t-tests with Welch's correction were performed for pairwise comparisons of mean shuffling rates. p-values are represented by *, $0.05 \geq p > 0.01$; **, $0.01 \geq p > 0.001$; ***, $0.001 \geq p > 0.0001$; ****, $p \leq 0.0001$; 'ns', non-significant. **B.** Indicative module occupancy in the shuffled library of selected SIs. The diagram above each plot represents the shufflon in its reference (unshuffled) form. Coloured squares represent the dummy modules when present on sense or antisense strands and red triangles indicate the *sfx* sites. For each panel, 6000 library members were randomly sampled and pooled from two biological replicates.

Figure 4. Shuffling of reporters by constitutively expressed SIs. A. Bar graphs showing the mean shuffling rate between two biological replicates. Each estimate is based on a minimum of 1300 analysed reads. Error bars represent standard deviation (SD). Following ANOVA, Dunnett's C test was performed for pairwise comparisons to *Rci*. p-values are represented by *, $0.05 \geq p > 0.01$; **, $0.01 \geq p > 0.001$; ***, $0.001 \geq p$; 'ns', non-significant. **B.** Indicative module occupancy in the shuffled library of selected SIs. The diagram above each plot represents the shufflon in its reference (unshuffled) form. Coloured squares represent the dummy modules when present on sense or antisense strands and red triangles indicate the *sfx* sites. For each panel, 4000 library members were randomly sampled and pooled from two biological replicates.

Supplementary Figure 1. Analysis of shuffled libraries. A. For each SI, the mean shuffling rate (orange) which measures the proportion of analysed library members which contain shufflons that are not in the original (unshuffled) configuration as well as the multiple-shuffling rate (blue) which measures the proportion library members that are the product of more than 1 inversion are shown for the pooled data of both biological replicates. **B.** Permutations were quantified from 4000 randomly sampled library members for each SI from the pooled data of both biological replicates. There are a maximum of 384 possible permutations for a five-module shufflon.

Supplementary Figure 2. Analysis of single module inversions. For each SI, those library members from both pooled biological replicates that contained the reference module configuration (in either inverted or non-inverted form) were extracted and then randomly downsampled to the smallest library of 2201 members. A multiple regression analysis indicated that module length ($p<0.00361$) and SI ($p<1.484\times10^{-9}$) were regression coefficients significantly explaining the module inversion frequencies.

Author Contributions

JK: Conceptualization, Investigation, Software, Formal Analysis, Visualization, Writing - Original Draft Preparation, Supervision. **MR**: Investigation, Software. **AA**: Methodology, Investigation. **JHM**: Data Curation. **NW**: Funding Acquisition, Conceptualization, Supervision, Writing - Review & Editing.

Supporting information

Supporting information: protein sequences of Rci homologs (.fasta), expression plasmid and reporter plasmid maps as GenBank files (.gb).

Acknowledgments

The authors thank Manjunatha Kogenaru and Alex Ivanov. The coding sequences of *Rci* mutants were kindly provided by Yi Wu from School of Chemical Engineering and Technology, Tianjin University, China.

Funding

Funded by the BBSRC via PhD studentship 2131394.

References

(1) Trzilova, D., and Tamayo, R. (2021) Site-Specific Recombination – How Simple DNA Inversions Produce Complex Phenotypic Heterogeneity in Bacterial Populations. *Microb. Genet. Stress Manag. 37*, 59–72.
(2) Chanin, R. B., West, P. T., Wirbel, J., Gill, M. O., Green, G. Z. M., Park, R. M., Enright, N., Miklos, A. M., Hickey, A. S., Brooks, E. F., Lum, K. K., Cristea, I. M., and Bhatt, A. S. (2024) Intragenic DNA inversions expand bacterial coding capacity. *Nature 634*, 234–242.
(3) Komano, T. (1999) Shufflons: multiple inversion systems and integrons. *Annu. Rev. Genet. 33*, 171–191.
(4) Nakayama-Imaohji, H., Hirakawa, H., Ichimura, M., Wakimoto, S., Kuhara, S., Hayashi, T., and Kuwahara, T. (2009) Identification of the site-specific DNA invertase responsible for the phase variation of SusC/SusD family outer membrane proteins in Bacteroides fragilis. *J. Bacteriol. 191*, 6003–6011.
(5) Sitaraman, R., Denison, A. M., and Dybvig, K. (2002) A unique, bifunctional site-specific DNA recombinase from Mycoplasma pulmonis. *Mol. Microbiol. 46*, 1033–1040.
(6) Dworkin, J., and Blaser, M. J. (1997) Molecular mechanisms of Campylobacter fetus surface layer protein expression. *Mol. Microbiol. 26*, 433–440.
(7) Sampei, G.-I., Furuya, N., Tachibana, K., Saitou, Y., Suzuki, T., Mizobuchi, K., and Komano, T. (2010) Complete genome sequence of the incompatibility group I1 plasmid R64. *Plasmid 64*, 92–103.
(8) Gyohda, A., Funayama, N., and Komano, T. (1997) Analysis of DNA inversions in the shufflon of plasmid R64. *J. Bacteriol. 179*, 1867–1871.
(9) Gyohda, A., Furuya, N., Ishiwa, A., Zhu, S., and Komano, T. (2004) Structure and function of the shufflon in plasmid r64. *Adv. Biophys. 38*, 183–213.
(10) Gyohda, A., Furuya, N., Kogure, N., and Komano, T. (2002) Sequence-specific and non-specific binding of the Rci protein to the asymmetric recombination sites of the R64 shufflon. *J. Mol. Biol. 318*, 975–983.
(11) Gyohda, A., Zhu, S., Furuya, N., and Komano, T. (2006) Asymmetry of shufflon-specific

recombination sites in plasmid R64 inhibits recombination between direct sfx sequences. *J. Biol. Chem. 281*, 20772–20779.
(12) Peikon, I. D., Gizatullina, D. I., and Zador, A. M. (2014) In vivo generation of DNA sequence diversity for cellular barcoding. *Nucleic Acids Res. 42*, e127–e127.
(13) Peikon, I.D. (2014, November 19) Encoding of neural circuit information into DNA. Cold Spring Harbor Laboratory.
(14) Han, P., Ma, Y., Fu, Z., Guo, Z., Xie, J., Wu, Y., and Yuan, Y.-J. (2021) A DNA Inversion System in Eukaryotes Established via Laboratory Evolution. *ACS Synth. Biol. 10*, 2222–2230.
(15) Naseri, G., Raasch, H., Charpentier, E., and Erhardt, M. (2023) A versatile regulatory toolkit of arabinose-inducible artificial transcription factors for Enterobacteriaceae. *Commun. Biol. 6*, 1005.
(16) Bastas, K. K. (2023) Chapter 10 - Bacterial diseases of potato and their control, in *Potato Production Worldwide* (Çalişkan, M. E., Bakhsh, A., and Jabran, K., Eds.), pp 179–197. Academic Press.
(17) Merchán, A. V., Ruiz-Moyano, S., Hernández, M. V., Martín, A., Lorenzo, M. J., and Benito, M. J. (2022) Characterization of autochthonal Hafnia spp. strains isolated from Spanish soft raw ewe's milk PDO cheeses to be used as adjunct culture. *Int. J. Food Microbiol. 373*, 109703.
(18) Wick, R. R., Judd, L. M., and Holt, K. E. (2019) Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol. 20*, 129.
(19) Shen, W., Le, S., Li, Y., and Hu, F. (2016) SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLOS ONE 11*, e0163962.
(20) Harvey, M. (2017) Imperial College Research Computing Service. Imperial College London.
(21) Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics 34*, 3094–3100.
(22) Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl. 25*, 2078–2079.
(23) Sankoff, D. (2000) The early introduction of dynamic programming into computational biology. *Bioinformatics 16*, 41–47.
(24) Liu, H. (2015) Comparing Welch ANOVA, a Kruskal-Wallis test, and traditional ANOVA in case of heterogeneity of variance.
(25) Zhou, Y., Zhu, Y., and Wong, W. K. (2023) Statistical tests for homogeneity of variance for clinical trials and recommendations. *Contemp. Clin. Trials Commun. 33*, 101119.
(26) Dunnett, C. W. (1955) A Multiple Comparison Procedure for Comparing Several Treatments with a Control. *J. Am. Stat. Assoc. 50*, 1096–1121.
(27) Lee, S., and Lee, D. K. (2018) What is the proper way to apply the multiple comparison test? *Korean J. Anesthesiol. 71*, 353–360.
(28) Zhang, Z., and Wang, L. (2017) Advanced Statistics Using R, p 138. ISDSA PRESS GRANGER, Indiana, USA.
(29) Delacre, M., Lakens, D., and Leys, C. (2017) Why Psychologists Should by Default Use Welch's t-test Instead of Student's t-test. *Int. Rev. Soc. Psychol.*