

Experimental demonstration of operon formation catalyzed by insertion sequence

Yuki Kanai¹, Saburo Tsuru^{2,*} and Chikara Furusawa^{2,3,*}

¹Department of Biological Sciences, Graduate School of Science, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan, ²Universal Biology Institute, Graduate School of Science, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan and ³Center for Biosystem Dynamics Research, RIKEN, 6-2-3 Furuedai, Suita, Osaka 565-0874, Japan

Received November 02, 2021; Revised December 23, 2021; Editorial Decision December 24, 2021; Accepted January 11, 2022

ABSTRACT

Operons are a hallmark of the genomic and regulatory architecture of prokaryotes. However, the mechanism by which two genes placed far apart gradually come close and form operons remains to be elucidated. Here, we propose a new model of the origin of operons: Mobile genetic elements called insertion sequences can facilitate the formation of operons by consecutive insertion–deletion–excision reactions. This mechanism barely leaves traces of insertion sequences and thus difficult to detect in nature. In this study, as a proof-of-concept, we reproducibly demonstrated operon formation in the laboratory. The insertion sequence IS3 and the insertion sequence excision enhancer are genes found in a broad range of bacterial species. We introduced these genes into insertion sequence-less *Escherichia coli* and found that, supporting our hypothesis, the activity of the two genes altered the expression of genes surrounding IS3, closed a 2.7 kb gap between a pair of genes, and formed new operons. This study shows how insertion sequences can facilitate the rapid formation of operons through locally increasing the structural mutation rates and highlights how coevolution with mobile elements may shape the organization of prokaryotic genomes and gene regulation.

INTRODUCTION

Operons are clusters of genes under the control of the same promoter sequences and are a hallmark of prokaryotic genome structure (1,2) and gene regulation (3). A significant proportion of genes of the prokaryotic genome are organized into operons (4), and it is widely accepted that oper-

ons are beneficial for multiple reasons, including for better coregulation of genes (5–8).

On the other hand, the mechanism by which operons form is poorly understood (9). As understanding operon formation is fundamental to our understanding of the evolution of prokaryotes, many models for operon formation have been proposed in the last 60 years (7,9–12). However, none of the proposed mechanisms of operon formation have been corroborated by experimental evidence, limiting our understanding of the types of mutations that drive them.

In principle, operons form by rearrangements, including insertions, deletions (9) (Figure 1A), and duplications (10), but these mechanisms alone seem to be insufficient to explain the prevalence of operons in prokaryotic genomes. For instance, duplication may explain the evolution of some operons (10), but genes in operons do not necessarily have similar structures (7). Moreover, while pairs of genes can culminate into operons by random rearrangements, they are more likely to end up apart (7).

Thus, selection for gene coregulation is thought to be crucial after operon formation by the aforementioned mechanisms to account for the prevalence of operons (5,9,13,14). However, while selection for coregulation can facilitate the maintenance and sophistication of operons (9), it does not facilitate their formation (7). This is because bringing two genes closer becomes adaptive only when the genes were initially clustered close enough (7). The ‘selfish operon model’ proposed that instead of relying on rare random rearrangements to cluster genes, operons form gradually through small intermediate steps (7). Unfortunately, this model has not found general acceptance as it fails to explain how essential genes likely form operons (5,15). Operon formation likely proceeds through a mechanism that gradually ties two far apart genes (7); however, a plausible mechanism remains to be elucidated.

As genes in operons are clustered together, mechanisms that cluster functionally related genes may also facilitate the formation of operons (13). Essential genes in prokaryotic genomes seem to cluster because the genomes tend to lose

*To whom correspondence should be addressed. Tel: +81 3 5841 4229; Fax: +81 3 5841 4229; Email: saburotsuru@gmail.com
Correspondence may also be addressed to Chikara Furusawa. Email: chikara.furusawa@riken.jp

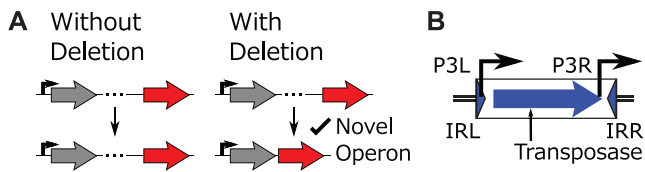


Figure 1. (A) Deletion facilitates the formation of operons. (B) IS3 has two promoters, P3L and P3R, in the same direction.

tandem sequences ('persistence model' (12)); Clustering of essential genes increases the robustness of genomes against tandem deletions by making large deletions less detrimental (16). Recently, a laboratory evolution of *Escherichia coli* revealed that the major cause of tandem deletions may be the activity of insertion sequences (IS) (17). This could mean that the tandem deletions by IS cluster important genes in prokaryotic genomes and may also make those genomes operon-rich.

ISs are a major cause of genetic rearrangements (17,18) and therefore may as well facilitate operons to form. A recent study found several examples of IS insertions between genes in widely conserved operons, suggesting that IS may facilitate rearrangements to form operons (19). However, no study seems to have associated the tandem deletion of surrounding genes by IS with the formation of operons.

Although ISs are known to delete neighboring sequences since at least the 1970s (20–22), deletions by IS as a mechanism that promotes the formation of operons have been undervalued for the following three main reasons: (i) IS-rich genomes seem to have more operons disrupted compared to other genomes (19,23). Therefore, ISs are often associated with the disruption rather than the formation of operons. (ii) As explained by our model described in the Results section, for an IS to form an operon, it must be excised after deleting adjacent sequences. However, IS excision would result in a double-strand break in the genome, which may lead to genomic degradation unless fixed by homologous recombination (24). (iii) Some of the major types of ISs rarely excise themselves (25).

The following findings have made these arguments less critical: (i) IS-rich genomes can be rich in operons (26), suggesting that IS-mediated rearrangements may not only destroy but also form operons (9). (ii) Many bacterial species have end-joining mechanisms (27). (iii) Many ISs are known to excise themselves (28,29). Moreover, the excision of one of the major types of IS, IS3, that rarely excise themselves, is significantly promoted by an enzyme called IS excision enhancer (IEE) found in various bacterial species (25).

We hypothesized that, considering that IS3 is known to rapidly create deletions of various lengths adjacent to it (21,30), the interaction of IS3 and IEE may promote the rapid formation of operons by IS3 deleting sequences intervening two genes and IEE excising IS3.

In this study, we assess if deletion by IS can promote the formation of operons. First, we propose a model of operon formation based on deletions by ISs, that does not rely on the selection for coregulation. We also suggest a mechanism by which an IS can facilitate the formation of operons through intermediate steps, gradually bringing closer

two genes placed far apart. According to this model, ISs do not leave apparent traces after forming operons. This makes it difficult to demonstrate the model using comparative genomics of extant genomes. Therefore, as a proof-of-concept, we engineered *E. coli* in a state that mimics an intermediate step of the model by inserting a copy of IS3 between two conditionally essential genes. To detect the diversity of structural mutations during the formation of operons, fluorescent reporters were placed around the copy of IS3. Selecting the cells cultured overnight based on their fluorescence, we examined how the concerted activity of IS3 and IEE affects the expression of surrounding genes, leading to the formation of novel operons. We believe that our study is the first attempt to demonstrate a plausible mechanism of prokaryotic operon formation in the laboratory.

MATERIALS AND METHODS

Reagents

For cell culture, we used: LB medium (Difco LB Broth, Miller, BD, USA, 244620), 50 $\mu\text{g}/\text{ml}$ kanamycin (Kanamycin Sulfate, Wako, Japan, 115-00342), 100 $\mu\text{g}/\text{ml}$ streptomycin (Streptomycin Sulfate, Wako, Japan, 190-14342), 50 $\mu\text{g}/\text{ml}$ ampicillin (Ampicillin Sodium, Wako, Japan, 012-23303), 50 nM anhydrotetracycline hydrochloride (aTc) (anhydrotetracycline hydrochloride, Sigma-Aldrich, USA, 37919), 0.1% (w/v) arabinose (L(+)-arabinose, Wako, Japan, 010-04582), 30 μM IPTG (isopropyl- β -D(-)-thiogalactopyranoside [IPTG], Wako, Japan, 190-14342), 1.5% agar (Agar, Powder, Wako, Japan, 010-08725).

For handling nucleic acids and DNA sequencing, we used: KOD One PCR Master Mix Blue (Toyobo, Japan, KMM-201), Rapid Barcoding Kit (Oxford Nanopore Technologies, UK, SQK-RBK004), DpnI (Takara Bio, Japan, 1235-A), PureLink RNA Mini Kit (Thermo Fisher Scientific, USA, 12183018A), PureLink DNase Set (Thermo Fisher Scientific, USA, 12185010), PrimeScript High Fidelity RT-PCR Kit (Takara Bio, Japan, R022A), In-Fusion Snap Assembly Master Mix (Takara Bio, Japan, 638948), FastGene Plasmid Mini Kit (NIPPON Genetics, Japan, FG-90502).

We used the following instruments: Infinite F200 multimode plate reader (Tecan, Switzerland), Flongle Flow Cell R9.4.1 (Oxford Nanopore Technologies, UK, FLO-FLG001), MinION Mk1B (Oxford Nanopore Technologies, UK, MIN-101B), FACSARIA III (BD, USA), MiniAmp Plus (Thermo Fisher Scientific, USA), blue/green LED transilluminator (NIPPON Genetics, Japan, LB-16BG).

The primers we used are listed in Supplementary Table S1.

Biological resources

The strains and plasmids used in this study are listed in Table 1.

To identify the effect of IS3 on the formation of operons, *E. coli* MDS42, a derivative of the wild-type K-12 MG1655, was used as it is absent of all mobile elements, including IS3

Table 1. Strains and plasmids used in this study

Strains and plasmids	Description	Source or reference
<i>E. coli</i> strains		
MDS42	Strain used for laboratory evolution	(31)
MG1655	Origin of IS3 transposase	Lab collection
HST08	Strain used to test <i>kanR-rfp</i> coregulation and for cloning	Takara Bio, Japan, 9128
Plasmids		
pCDSSara-L18	Backbone of pYK-1S3, <i>-iee</i> control	(32)
pRC2117	Origin of Tn5 transposase gene of pYK-1Q2	(33)
pYK-1N5	Fluorescence reporter plasmid (+ <i>tpn</i>)	This study
pYK-1Q2	pYK-1N5 with Tn5 transposase gene instead of IS3 <i>tpn</i> , <i>-tpn</i> control	This study
pYK-1S0	Origin of P _{lac} cassette	This study
pYK-1S3	IEE expression plasmid (+ <i>iee</i>)	This study

(31). IS3 and IEE were supplied via plasmids pYK-1N5 and pYK-1S3, respectively. All plasmids were constructed using the In-Fusion cloning kit.

The IS3 supplying vector (pYK-1N5, Figure 3A) was designed to mimic a state after the ‘insertion’ step of the model that we propose in this study (Figure 2A). pYK-1N5 is a low copy-number plasmid (pSC101 origin) with a kanamycin resistance gene (*kanR*), an engineered IS3 sequence, and red (mScarlet-I (34)) and yellow (Venus YFP (35)) fluorescent protein genes (*rfp*, *yfp*). To demonstrate our proposed model, IS3 (Figure 1B) was engineered to have between its two inverted repeats (IR) the IS3 transposase gene (*tpn*) downstream of an inducible promoter (based on P_{LtetO-1} (36,37)) and a synthetic promoter P_{J23105} (a Biobrick promoter). The P_{LtetO-1} derivative is repressed by the product of *tetR* on the same plasmid, and therefore, transcription is inducible by the addition of aTc to the growth medium. P_{J23105} was inserted to mimic the outward-facing promoter P3R of wild-type IS3 because, according to our model, it facilitates the formation of operons. We chose P_{J23105} because its strength was ideal for demonstrating using flow cytometry (FCM) the influence of IS3 activity on the expression of surrounding genes. The two promoters of IS3 were arranged in a divergent orientation and a synthetic terminator was placed upstream of P_{LtetO-1}. This design was to avoid induction by aTc to increase the transcription of the fluorescent protein, to avoid potential promoter activity from IRL to interfere with P_{LtetO-1}, and to prevent the plasmid from being degraded before evolution by inhibiting the activity of the transposase gene by antisense transcription from the *kanR* gene. In addition, following a previous study (21), the two open reading frames constituting the IS3 transposase gene were fused by a single base insertion in the A₄G motif to increase the activity of the transposase.

Excluding the origin of replication, the longest repetitive sequences in pYK-1N5 were the ribosomal binding site sequence of SD8 (20 bp) upstream of *yfp* and *rfp*, the tet operator sequence of P_{LtetO-1} (17 bp), and the sequences of two synthetic terminators (14 bp, (37)) located upstream of P_{LtetO-1} and downstream of *tpn*, respectively.

The IEE expression vector (pYK-1S3, Figure 3B) was constructed based on pCDSSara-L18 (32) by replacing the arabinose-inducible ribosomal protein gene with *iee* of *E. coli* O157:H7 strain Sakai (NCBI Gene ID: 912859, ECs_1305). This plasmid had a streptomycin resistance gene (*smr*) as a selection marker.

Derivatives of the two plasmid vectors without the IS3 and IEE genes were used as negative controls. A hyperactive mutant of Tn5 transposase gene (33) and ribosomal subunit L18 (32) were placed instead of IS3 transposase and IEE genes, respectively.

To introduce IS3 and IEE, the ancestor strains for evolution were prepared by two rounds of electrotransformation. First, pYK-1S3 (or its *-iee* derivative) was introduced by transformation, and the transformed cells were selected on LB agar plates containing streptomycin. Mixtures of electrocompetent cells were then prepared from the selected colonies. Finally, pYK-1N5 (or its *-tpn* derivative) was introduced by transformation into the cells and the transformed cells were grown on LB agar plates containing streptomycin and kanamycin.

Culture conditions

Cells were cultured aerobically in 200 μ l volume of LB medium in 96-well plates (Greiner Bio-One, Cellstar, 655180). First, the medium was filtered through a 0.2 μ m membrane filter (ThermoFisher Scientific, 567-0020) to remove debris that may interfere with FCM measurements. Next, the culture medium was supplemented with kanamycin and streptomycin (unless mentioned otherwise) to select for only those cells that harbor both pairs of plasmids. To induce transcription of transposase and IEE genes, aTc and arabinose, respectively, were added to the media where indicated.

Growth rates were measured in Infinite F200 multimode plate reader at 37°C overnight. The plate was shaken linearly and then orbitally for three minutes each, and OD₆₀₀ was measured every ten minutes to determine the growth rate.

To obtain colonies, cells were plated on agar plates supplemented with LB and appropriate antibiotics.

Cells were temporarily stored at 4°C as colonies grown on LB agar plates or as mixtures in filtered phosphate-buffered saline (PBS).

Flow cytometry and selection of cells by fluorescence-activated cell sorting (FACS)

For the evolution of the cells, four colonies of each of the four genotypes (\pm *tpn* and \pm *iee*) were picked with toothpicks from LB agar plates, diluted 100-fold, and cultured

overnight in 96-well plates with 200 μ l medium supplemented with aTc and arabinose. After overnight culture at 37°C, the cells were diluted in ice-cold PBS and stored at 4°C until FCM measurements. The single-cell fluorescence of these cells was measured by FCM using FACSAria III and FACS Diva software v.6.1.3. The analysis and visualization were based on single-cell fluorescence of cells with forward scatter measurements within a two-fold range, including the most frequent values to compare fluorescence among cells in similar physiological states (38).

For FACS, four populations of cells with both *tpn* and *iee* and one population of cells with *tpn* but without *iee* were prepared as described above. After a brief measurement of the distribution of fluorescence by FCM using 10 000 cells, gates were manually set by first subsetting the measurements with the same forward scattered values as above and then manually drawing the gates according to the measurement, as shown in Figure 5A (P1-3). The same gates were used for all five rounds of FACS. The cells were collected in 500 μ l of LB or PBS.

Genotyping cells

To confirm that the cells were sorted as expected, 200–500 μ l of the collected cells were cultured overnight in 5 ml volume, and 10 μ l of cultured cells were plated on agar plates with the antibiotics. The plates were incubated at 37°C for two days and imaged under a transilluminator. The brightness and contrast were adjusted using ImageJ (National Institutes of Health, USA).

Sanger sequencing was used to genotype the cells in each FACS gate. Colonies on agar plates showing uniform fluorescence were picked, and colony PCR was performed using KOD One PCR Master Mix Blue. Template DNA was prepared by picking the colonies by toothpicks into Tris-HCl (pH 8.5) with 0.1% Triton X-100, boiling at 98°C for 5 min, and spinning down briefly at 10 000 \times g. The supernatants were used as templates. Agarose gel electrophoresis was run for the amplified DNA, and fragments of typical lengths were chosen and sequenced by Sanger sequencing (GeneWiz). The same procedure, but with a different primer set, was followed to check that sequences were not deleted in plasmids of cells in the major dark fraction of FCM measurements in *+iee +tpn* cells.

To identify the diversity among the cells sorted by FACS, nanopore sequencing was performed. 400 μ l out of 500 μ l of PBS with 600 cells collected by a single round of FACS were transferred into 5 ml of LB medium containing only kanamycin and cultured overnight at 37°C. Streptomycin was not added to reduce the read mapping onto pYK-1S3 and increase read mapping onto pYK-1N5. Plasmid DNA was extracted with FastGene Plasmid Mini Kit (with optional washing), and the sample for sequencing was prepared by adding barcodes and adapters using Rapid Barcoding Kit following the manufacturer's instructions. The samples were applied to Flongle Flow Cell R9.4.1 on MinION Mk1B and sequenced using MinKNOW (v21.06.0). The FAST5 data was basecalled using ONT Guppy (v5.0.11) with the configuration set to the 'Super-accurate' model. The reads were filtered according to the default quality threshold, and the barcode sequences

were trimmed. We obtained 3522 reads with 14 705 564 base pairs (bp) corresponding to a sequencing depth of \sim 2000. The median read length and quality were 4359 bp and 13.03, respectively. Reads longer than one kb were filtered with Filtrong (v0.2.0, <https://github.com/rswick/Filtrong>) and mapped onto a FASTA file with two copies of the pYK-1N5 sequence placed in tandem using minimap2 (v2.17-r941 (39)) to avoid reads that were mapped across the origin being split. Deletions longer than 50 bp were detected and extracted from the CIGAR strings of BAM using Rsamtools (v2.6.0) and analyzed using a custom script. The 514 reads are plotted in Figure 5F.

Reverse transcription polymerase chain reaction (RT-PCR)

To assess if *kanR* and *rfp* formed an operon, RT-PCR was performed. Two colonies genotyped by Sanger sequencing were picked and cultured overnight at 30°C. The cultured cells were diluted 50-fold into a fresh growth medium and cultured at 30°C for 2 h. The total RNA was extracted using the PureLink RNA Mini Kit. DNA was removed using DNase during the extraction. Total RNA was amplified using PrimeScript High Fidelity RT-PCR Kit following the manufacturer's two-step protocol. First, reverse transcription was performed. RNA with a final concentration of \sim 50 ng/ μ l was denatured at 65°C for 5 min with the reverse primer and, as negative controls of reverse transcription and genome DNA contamination, with no primers or with the forward primer; reverse-transcriptase and buffers were added on ice; reverse transcription was performed at 42°C for 30 min; the reverse transcriptase was denatured at 95°C for 5 min; the RNA mixtures were stored at 4°C. Then, cDNA was amplified by PCR. The obtained cDNA was diluted 10-fold with a premix containing the pair of primers and the DNA polymerase; DNA was amplified by 20 cycles of PCR (MiniAmp Plus), with denaturation at 98°C for 10 s, followed by annealing at 57°C for 5 s, and extension at 72°C for 3 min. The PCR products were subject to gel electrophoresis, and the image was recorded under a transilluminator. The brightness and contrast of the image were adjusted using ImageJ.

Testing the coregulation of *kanR* and *rfp* in newly-formed operons

To further confirm operon formation after evolution, the *kanR* promoter of pYK-1N5 and its evolved derivatives were replaced with a lactose inducible promoter P_{lac} . DNA sequences, excluding the kanamycin promoter sequence, of the plasmids were amplified by PCR (KOD One PCR Master Mix Blue). Another sequence containing the ampicillin resistance gene, P_{lac} and *lacI* (Supplementary Figure S6) was amplified by PCR from plasmid pYK-1S0. After digestion of remnant plasmids with DpnI, the amplified sequences were fused with In-Fusion cloning following the manufacturer's manual, and the DNA mixture was introduced to *E. coli* HST08 Premium Competent Cells. The sequence of the insert is provided in the DNA Sequences section of Supplementary Data.

Cells with the re-engineered plasmids were spread on LB agar plates with ampicillin. After checking the genotypes

of the cells, their colonies were picked and transferred to 96-well plates with 200 μ l of LB supplemented with ampicillin and IPTG. The cells were cultured overnight at 32°C to prevent edge effects. After the cells reached the stationary phase, the cells were diluted in ice-cold PBS, and single-cell fluorescence was measured using FCM.

Statistical analyses

Type III ANOVA was performed using FCM measurements of 16 independent cell cultures consisting of four colonies for each of the four genotypes. Because there were three treatments, there were 12 degrees of freedom within groups.

For calculating the number of cells in the three gates used for FACS, the mean and 95% confidence interval (CI) of \log_{10} probability are presented ($n = 4$), assuming the normality of the distributions.

To show that IPTG induction increased red fluorescence, three independent cell cultures were prepared for each condition and genotype, and the intensity of single-cell red fluorescence was measured with FCM. First, the FCM data was subsetted based on their forward scatter values as above. Using the values of median \log_{10} red fluorescence of over 10 000 measurements per culture, we performed two-sided t -tests adjusted with the Bonferroni method.

Data availability/sequence data resources

Data related to nanopore sequencing and its analysis, including the FAST5 data, reference FASTA, and BAM files, were deposited in NCBI Sequence Read Archive (BioProject PRJNA768397). The raw flow cytometry data are available in FlowRepository (FR-FCM-Z4LV). DNA sequences of pYK-1N5, pYK-1S3 and the P_{tac} cassette are provided in the DNA sequences section of Supplementary Data.

Data availability/novel programs, software, algorithms

Not applicable.

Web sites/data base referencing

As references, we used NCBI Gene for the DNA sequence of *iee*, and the Registry of Standard Biological Parts to find various Biobrick promoter sequences, including P_{J23105} .

RESULTS

The IDE model of operon formation based on the activity of insertion sequences

To explain the formation of operons in prokaryotes, we propose a model that relies on the activity of ISs and experimentally demonstrate the formation of an operon. ISs may be able to cluster two genes placed far apart into an operon by a sequential insertion–deletion–excision reaction as follows (IDE model, Figure 2A).

Ancestor: Initially, two genes beneficial in the same environment are placed apart, as is the case for the genes in gray and red (I, III) in Figure 2A.

Insertion: Many ISs have promoters that facilitate the transcription of downstream genes (40,41). For example,

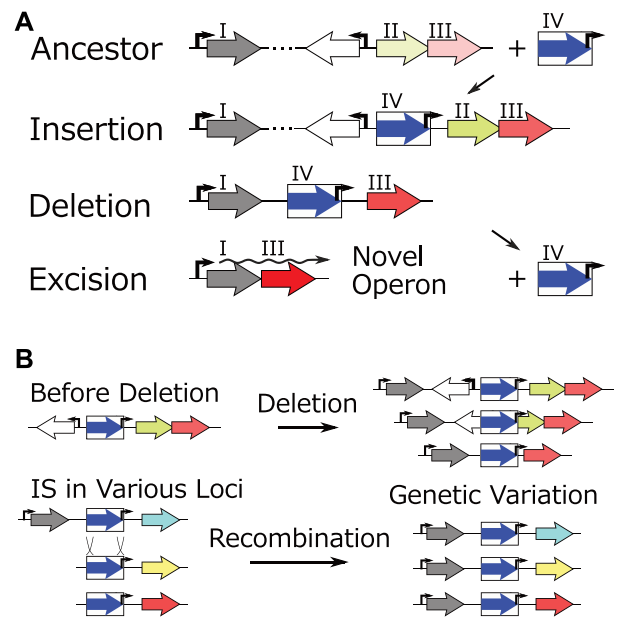


Figure 2. Insertion sequences (ISs) can facilitate the formation of novel operons through insertion–deletion–excision reactions (IDE model). (A) An example of how ISs (IV) can facilitate the formation of operons based on the IDE model. The block arrows indicate coding sequences, and the intensity of green (II) and red (III) colors indicates the gene expression level. In an environment where the genes indicated in red (III) and gray (I) are essential and that in green (II) is nonessential, the reactions from the top to bottom lead to the genes in gray and red to form a novel operon. (B) During the ‘deletion’ step of the IDE model, the inserted IS can locally accelerate structural mutation rates facilitating operon formation regardless of the intensity of selection. Firstly, IS can cause tandem deletions of various lengths at a high rate. Secondly, recombination among multiple copies of IS can bring various genes together. Some of the combinations of genes can be beneficial to form operons and get fixed after further deletion and excision.

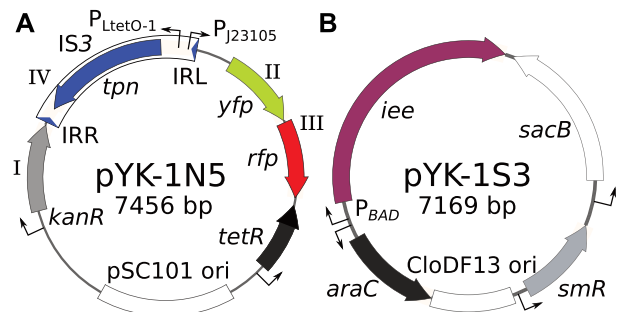


Figure 3. (A) Design of the fluorescent reporter plasmid used to detect operon formation. The plasmid was designed according to the gene arrangement shown in the ‘insertion’ step of Figure 2A. See *Biological resources* for details. (B) Design of the insertion sequence excision enhancer (IEE) expression plasmid. Tetracycline repressor gene, *tetR*; origin of replication, *ori*; arabinose operon regulator gene, *araC*; streptomycin resistance gene, *smR*. The genes of pYK-1N5 with roman numerals correspond to the genes in Figure 2A with the same roman numerals.

in addition to the promoter upstream of the transposase gene, IS3 has a strong promoter facing outward (P_{3R} , Figure 1B (42)). These promoters help an IS (IV) to activate initially dormant genes (II, III) by transposing upstream of the genes (41).

Deletion: Many types of ISs are known to delete sequences adjacent to them (20–22). Deletion can continue until nonessential genes between the IS and beneficial genes are deleted. This results in bringing together two beneficial genes on either side of the IS closer.

Excision: Many ISs can excise themselves, especially in the presence of IEE (25). This excision of an IS between two beneficial genes can lead to them forming a novel operon.

Thus, tandem deletions caused by IS can lead to the formation of operons without relying on selection for better coregulation before operons have formed.

IS can facilitate operon formation by the IDE model by increasing structural mutation rates

A characteristic of operon formation by the IDE reaction is the formation of an “enzyme-substrate complex,” where the IS is the “enzyme,” and the genome is the “substrate.” This facilitates operon formation regardless of the selection pressure to form operons. During the ‘deletion’ step, the copy of the inserted IS accelerates the local rate of structural mutations around its location by following two mechanisms (Figure 2B). Firstly, IS can “creep” towards genes by deleting adjacent sequences. Secondly, recombination among copies of the same IS allows genomes to test various pairs of genes, some of which are beneficial to form operons, to come together. This could be important as functionally related genes can be randomly dispersed within a genome.

Note that in our experimental demonstration of the IDE model to be described in this paper, for simplicity, we have ignored the role of recombination described here.

IDE reactions can also form novel operons through intermediate adaptive steps

While adaptive evolution is not essential, it can lead to more efficient operon formation via the IDE reactions. In theory, mutations that gradually bring two genes closer can be more effective than rare rearrangements that directly cluster two genes to form an operon (7). Our model involves such a mechanism and can be divided into small intermediate steps, all of which are adaptive (Figure 2A). For example, let us assume that the stronger expression of genes indicated in gray (I) and red (III) is beneficial and that initially, the gene in gray is upregulated, but the gene in red lacks an effective promoter (ancestor). The insertion of an IS (IV) upstream of the gene in red would increase the fitness, as the gene would be transcribed from the promoters within the IS. Further activity of the IS can delete sequences around the IS, increasing the expression of the gene in red by bringing it closer to the inner promoters of the IS and gradually closing the space between the IS and the gene. Finally, the excision of IS would be most beneficial when cells with both gray and red genes are adjacent to the IS as it would bring the promoter of the gene in gray closer to the gene in red. The premature excision of IS is less likely to be fixed because the gene in red would lose active promoters encoded within the IS. Also, the excision step can be adaptive if it stabilizes a new beneficial operon structure or eliminates the potential for deleterious mutations mediated by the IS element, such as adjacent deletions that remove beneficial genes.

A fluorescence reporter system to detect deletions related to the formation of *kanR-rfp* operon

To experimentally demonstrate our model of operon formation, we designed a plasmid as a model of the prokaryotic genome after the ‘insertion’ step (pYK-1N5, Figure 3A), which is common in both nature and laboratory settings (30,41,43). Specifically, the fluorescent reporter plasmid pYK-1N5 was designed such that the kanamycin resistance gene (*kanR*, I), a copy of IS3 (IV), and the red (*rfp*, III) and yellow (*yfp*, II) fluorescent protein genes correspond to the genes in gray, the copy of IS, and the genes in green and red in Figure 2A, respectively. The copy of IS3 was inserted within a 2.7 kb gap between *kanR* and *rfp*. For ease of the experiment, the wild-type IS3 was modified such that its transposase gene (*tpn*) was inducible by aTc, and the promoter P_{J23105} activated *rfp* instead of the native promoter of IS3. We expected that the IS would delete surrounding sequences or excise itself, changing the pattern of fluorescence measured by FCM. Using FACS, we selected cells with bright red fluorescence while adding kanamycin to the growth medium, expecting *kanR-rfp* operons to form by positive selection.

With the two fluorescent protein genes, various phenotypes were expected to be found after evolution. Deletions between IS3 and *yfp* would bring P_{J23105} closer to *yfp*, causing both yellow and red fluorescence to increase. Excision of IS3 would lead to the transcription of the two fluorescent protein genes from the *kanR* promoter instead of the weaker P_{J23105}, also causing both yellow and red fluorescence to increase. In this case, we expected *kanR-yfp-rfp* operon to “re-generate,” reverting to its pre-IS insertion state. When *yfp* was also deleted, we expected *kanR-rfp* operons to form.

The pattern of fluorescence observed under the presence of both IS3 and IEE

As a potential genetic background that facilitates operon formation, we examined whether the combined activity of IS3 and IEE in *E. coli* can accelerate operon formation. IS3 rarely excises itself on its own but is frequently excised under the presence of IEE (25). The presence of IEE thus is expected to promote the ‘excision’ step of operon formation. To test this, an arabinose-inducible IEE expression vector (pYK-1S3, Figure 3B), in which IEE was expressed from the promoter of *araBAD* operon (P_{BAD}), was constructed. As negative controls, derivatives of pYK-1N5 and pYK-1S3 without *tpn* and *iee*, respectively, were also constructed. We transformed an IS-less strain of *E. coli* MDS42 (31) with pairs of plasmids with and without *tpn* and *iee*. We expected that with both *tpn* and *iee*, a large proportion of cells would show an increased level of yellow and red fluorescence by losing the copy of IS3. We also expected some cells to show an increase in only red fluorescence as IS3 can disrupt *yfp*.

To test if the activity of IS and IEE creates the expected fluorescence patterns, cells transformed with the pairs of plasmids were cultured overnight with both *iee* and *tpn* fully induced (Figure 4, Supplementary Figure S1). We found that when both *tpn* and *iee* were present, a fraction of cells showed intense yellow and red fluorescence (Figure 4, white arrowhead). While these cells were also found in other conditions, they were most apparent when both *tpn* and *iee* were

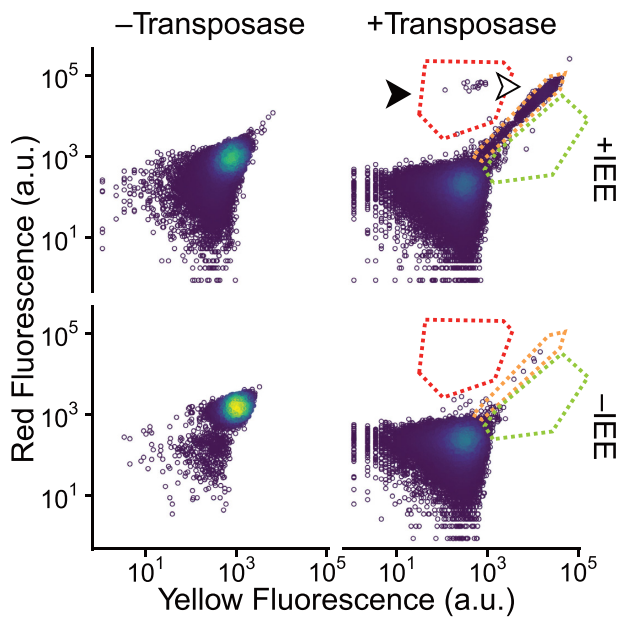


Figure 4. Single-cell fluorescence of cells with and without *tpn* and *iee* measured by flow cytometry ($n = 50\,000$). The colors of the points indicate the density of points in the fraction; yellow indicates dense, and blue indicates sparse. The red, orange, and green dashed lines indicate the gates P1-3 set for FACS as in Figure 5A, B.

present and fully induced. To validate the importance of *tpn* and *iee* co-expression, fluorescence data collected from four biological replicates of four genotypes ($n = 16$) were subjected to Type III ANOVA. We used the proportion of measurements with red fluorescence ten times brighter than the median of each measurement as the response variable and the presence of *tpn*, *iee* and the interaction of the two as the treatments. We found that the synergy of *tpn* and *iee* significantly facilitated the appearance of cells with intense fluorescence ($F_{1,12} = 4.2 \times 10^2$, $P = 1.0 \times 10^{-10}$, coefficient = 3.8×10^{-2}). Because the combined expression of *tpn* and *iee* has been associated with the excision of IS3 (25), this suggests that, in the bright cells, IS3 was excised, and *yfp* and *rfp* were transcribed from the stronger promoter of *kanR*.

Besides cells showing both stronger yellow and red fluorescence, some cells with *tpn* and *iee* showed stronger red fluorescence but modest yellow fluorescence (Figure 4, black arrowhead). This implies that, as intended, not just the excision of IS3 but also the deletion of adjacent *yfp* may have occurred.

We also noticed that the intensity of fluorescence decreased with the expression of either or both of *tpn* and *iee* (Figure 4, Supplementary Figure S2A). ANOVA with the median log fluorescence values as the response variable and the expression of *tpn*, *iee*, and the interaction of the two as treatments showed that the influence of *tpn* expression was the largest and most significant ($F_{1,12} = 7.9 \times 10^2$, $P = 2.6 \times 10^{-12}$). The influence of *iee* expression was also found to be significant ($F_{1,12} = 4.3 \times 10$, $P = 2.6 \times 10^{-5}$). The fold-changes were analyzed using the linear regression models used for the ANOVA. This analysis showed that the expression of *tpn* and *iee* reduced the fluorescence intensity by 85% and 36%, respectively. Given that we have used a

low copy-number plasmid, mutations that further lower the plasmid copy number will lead to cells without any plasmid with *kanR*, causing observable growth defects. However, growth defect was not apparent even with the presence of kanamycin in the growth medium (Supplementary Figure S2B). Moreover, cells collected from the major fraction of FCM distribution had both *yfp* and *rfp* intact (Supplementary Figure S3), implying that IS-mediated deletions of *yfp* and *rfp* were also not the cause of the reduced fluorescence intensity. The unexpected decrease in fluorescence seems to indicate that, during operon formation, the activity of IS can alter the expression pattern of neighboring genes.

IS generated a variety of genotypes, including those that seem to have formed operons

To determine whether the altered fluorescence of the cells reflects the formation of novel operons, some cells with *tpn* and with or without *iee* were sorted based on their fluorescence (gates P1-3) by FACS (Figure 5A). Gate P1 was set to collect cells with strong red fluorescence, which corresponds to cells shown with a black arrowhead in Figure 4. Gate P2 was set to collect cells with bright red and yellow fluorescence, which corresponds to cells shown with a white arrowhead in Figure 4. We noticed some cells had weak red fluorescence but strong yellow fluorescence; We set gate P3 to collect these cells. Because cells in P1 and P3 were less in number compared to those in P2, the gates were set broadly to improve the efficiency of FACS.

First, cells with both *tpn* and *iee* that evolved as four independent populations were analyzed. Consistent with Figure 4, a notable proportion of cells were found in P2 ($P = 10^{-1.3 \pm 0.1}$, $10^{\text{mean} \pm \text{CI}}$), and a few cells in P1 and P3 were also found with probabilities of $10^{-4.1 \pm 0.4}$ and $10^{-3.3 \pm 0.4}$, respectively. To investigate the diverse genotypes that seemed to have led to the formation of *kanR-rfp* operons, cells in P1 were collected from all four independent cultures. Cells in P2 and P3 were also collected from one population of cells. Next, a population of cells without the *iee* was also analyzed. Although the relative proportion of these cells was lower than the measurement limit ($P < 10^{-5.5}$), cells in gate P1 were also collected for further analysis (Figure 5B).

To assess if fluorescence was heritable, the cells were spread on LB agar plates with streptomycin and kanamycin and were regrown overnight. The fluorescence of colonies under blue/green LED was largely consistent with the fluorescence measured by FCM (Figure 5C, Supplementary Figure S4). The cells from P1 showed red fluorescence as expected; cells from P2 and P3 generally showed green to orange fluorescence. However, some cells collected from P3 did not show bright fluorescence, suggesting that gate P3 contains populations overlapping with those in the major dark fraction of cells.

Next, sequences around the IS3 of the colonies were amplified by colony PCR to genotype the sorted cells. Fragments with deletions of typical lengths were chosen and sequenced by Sanger sequencing (Figure 5D, Supplementary Table S2). The red colonies grown from gate P1 +*iee* cells had sequences deleted from the right inverted repeat (IRR) to *yfp*, including parts of *yfp*. In such cases, the

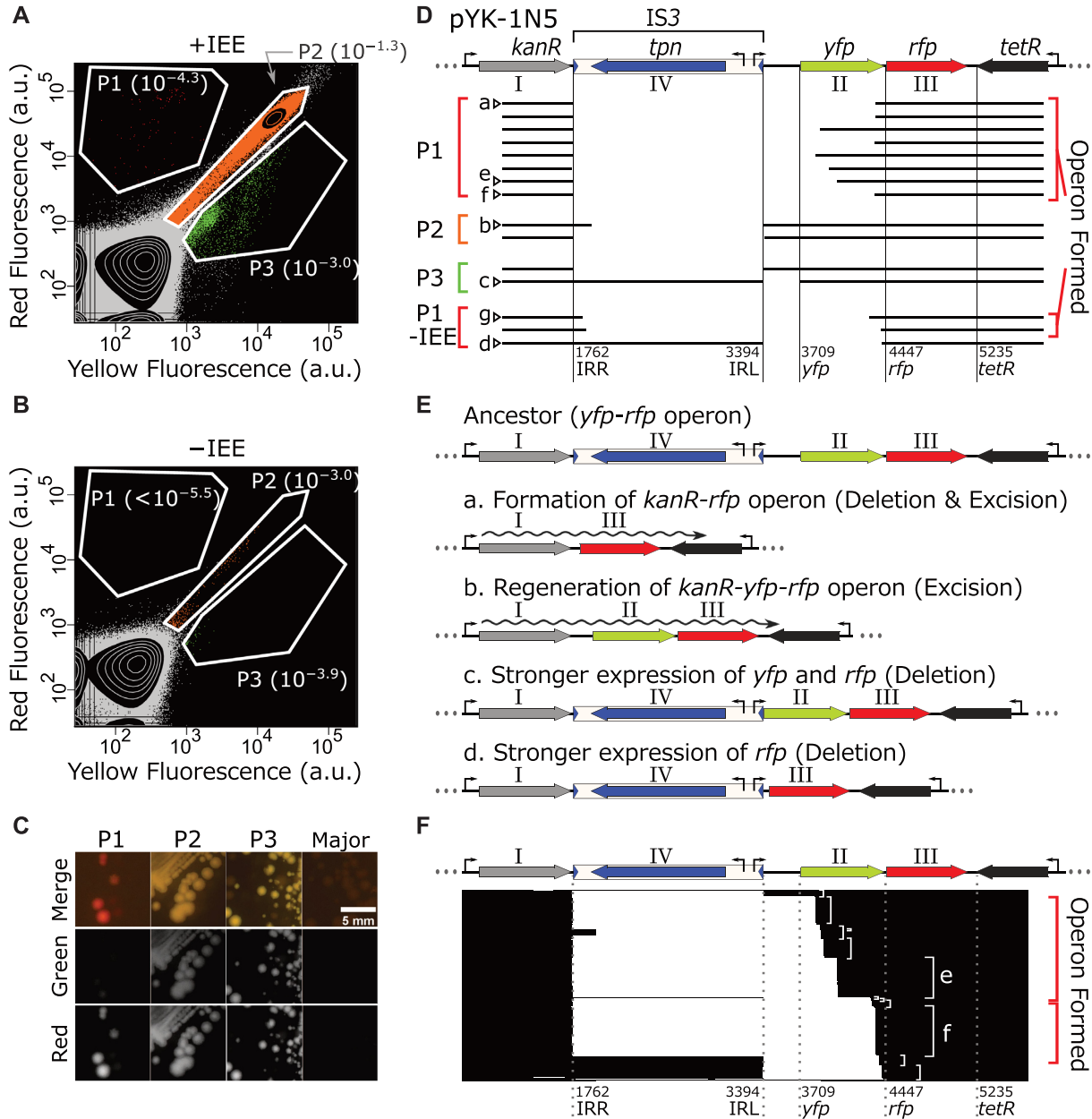


Figure 5. (A, B) Red and yellow fluorescence of cells with the IS3 transposase and with (A) or without (B) insertion sequence excision enhancer (IEE) expression vectors after evolution overnight (A: $n = 1\,986\,386$; B: $n = 300\,644$). The transcription of both transposase and IEE genes (when present) were fully induced. Each polygon (P) indicates the gates used to sort cells with fluorescence-activated cell sorting (FACS) for further analysis. The numbers within the parentheses show the proportion of cells found within each gate. The relative proportion of cells in P1 of subfigure B was lower than the measurement limit. The cells were prepared under the same conditions as the measurements of Figure 4 (A: +Transposase +IEE, B: +Transposase -IEE) but were from an independent population of cells. (C) Colonies of cells collected by FACS under a transilluminator. (D) Typical DNA sequences of pYK-1N5 that remained after evolution were determined by Sanger sequencing and are shown as black horizontal lines. The colors of the left square brackets correspond to the colors of points in subfigures A and B. The vertical lines indicate the boundaries of genetic elements. The details of each deletion are provided in Supplementary Table S2. (E) The order of genes after the deletions shown in subfigure D. Each map corresponds to the sequence with the identical characters (a–d) in subfigure D. The genotypes underwent steps of the IDE model as indicated in parenthesis. (F) Analysis of the variety of deletions using nanopore sequencing. Plasmids extracted from a population of cells sorted from P1 in subfigure A by FACS were sequenced. The loci of deletions found in each read are shown as white horizontal lines. The white square brackets on the right indicate reads showing similar deletions.

plasmids tended to have the IS3 completely deleted. This result is in line with a study (25) that showed that IEE promotes the complete excision of IS3. The sequenced cells collected from gate P2 had the IS3 deleted up to the IRL. While many colonies had the IS3 completely deleted with excision starting adjacent to the IRR, some colonies had part of the IS3 remaining as in read *b* of Figure 5D. Some cells collected from P3 had a genotype similar to those in P2, and others had only the sequence between IRL and *yfp* deleted. The latter cells had the maximum possible deletion preserving the yellow fluorescence (read *c*, Supplementary Figure S5), implying that diverse types of deletions were generated by the activity of the IS.

Rare cells in P1 without *iee* (Figure 5B) also had *yfp* deleted (Figure 5D, P1 –IEE) (25). Among these cells, some cells also had the copy of IS3 excised as in read *g* of Figure 5D, but with parts of IS3 remained. This shows that although rarely, IS3 may also be able to form operons without IEE.

We found at most 6 bp of homology at deletion endpoints (Supplementary Table S2); therefore, it is unlikely that the deletions and excisions were mediated by sequence repeats.

The deletions found by Sanger sequencing can be categorized into four types, as exemplified by reads *a–d* (Figure 5E). When sequences between IS3 and *rfp* were deleted together with IS3, novel *kanR-rfp* operons formed (P1, read *a*). The excision of only IS3 led to the strong promoter of *kanR* to come closer to the *yfp-rfp* operon, regenerating *kanR-yfp-rfp* operons (P2, read *b*). Some cells did not form operons but had sequences adjacent to the IS3 deleted, increasing red fluorescence. Deletion of sequences between IS3 and *yfp* led to stronger yellow fluorescence (P3, read *c*), and *yfp* deletion led to the loss of yellow fluorescence (P1 –IEE, read *d*). The type of deletion found in read *a* was the type of deletion we expected to find when new operons form by the IDE model in our experimental setup. As these cells were collected after a day of evolution, we cannot determine whether the ‘deletion’ and ‘excision’ steps occurred simultaneously or as two consecutive steps. Nevertheless, we also obtained genotypes that revealed cells that had been through only the ‘deletion’ (reads *c*, *d*) or ‘excision’ (read *b*) steps. This supports the IDE mechanism we have proposed in this study.

To demonstrate the diversity of operon-forming deletions IS can generate in the presence of *iee*, plasmids were extracted from cells sorted by FACS in P1. The deleted loci of the plasmids were analyzed by nanopore sequencing. Consistent with the results of Sanger sequencing, deletions of various lengths were observed (Figure 5F). The major types of deletions identified were those also found by Sanger sequencing of the colonies derived from the same population of cells collected by FACS (reads *e* and *f*). Although few, some reads had the IS3 sequence completely preserved, similar to read *d*. This implies that in some cells, the transposase may have deleted the sequences surrounding IS3 without IEE involvement. In addition, similar to read *b*, some cells seemed to have IS3 partially left undeleted. These results are consistent with a previous study where some cells had intact or partially excised IS3, even when both IS3 and IEE were activated (25). Although sequencing by nanopore sequencing is generally error-prone and the estimated position of

the junctions may have a few base pairs of errors, the deletions found were largely consistent with those detected by Sanger sequencing, reinforcing that plasmids collected from the population of sorted cells contained a variety of deletions, many of which have led to operon formation.

Overall, these results confirm our predictions of the genotypes for the distinct populations of cells in Figure 4. For example, the population of cells highlighted with the white arrowhead (gate P2) indeed had the IS3 excised. The proportion is negligible without IS3*tpn* and is significantly increased with *iee* as in a previous study (25). The population of cells highlighted with the black arrowhead (gate P1) indeed had sequences up to *yfp*, including parts of *yfp*, deleted. These cells had either the *kanR* promoter closer to *rfp* or, when the IS was intact, the inner promoter of IS (P_{J23105}) adjacent to *rfp* (Figure 5D, F). This may be due to the strong selection imposed in this study for strong red fluorescence. In +*tpn* conditions of Figure 4, there were some measurements with red fluorescence stronger than the majority of cells but still lower than that of gate P1 cells. This group of cells may have had shorter deletions of *yfp*. In –*tpn* conditions (Figure 4), no such cells, nor cells in P1 were observed.

In line with studies of IS-mediated deletions (20–22,25), IS3 transposase seems to have caused deletions that promote operon formation by the IDE model.

Validation of the formation of novel operons

To validate the formation of new operons, two experiments were performed.

First, the presence of mRNA transcribed from *kanR* to *rfp* was detected by RT-PCR (Figure 6A) using RNA extracted from cells corresponding to read *e* of Figure 5D. The RNA was reverse transcribed into cDNA using the forward or reverse primer, and subsequently, the DNA was amplified using the pair of primers by PCR. We detected RNA with the length expected from the genotype (835 bp) only in the direction of *kanR* (Figure 6B, e). This indicates that the evolved cells had both *kanR* and *rfp* on the same mRNA transcribed from the promoter of *kanR*, supporting the formation of a *kanR-rfp* operon. Using RNA extracted from cells corresponding to read *b* of Figure 5D instead, we detected 1951 bp RNA, supporting the regeneration of a *kanR-yfp-rfp* operon (Figure 6B, b).

Next, we tested whether the two genes could be controlled together. Specifically, we tested the coregulation of the two genes by replacing the promoter upstream of *kanR* with an IPTG inducible promoter P_{*tac*} (Figure 6A). We replaced the promoters of pYK-1N5 and its evolved derivatives and transformed the new plasmids into *E. coli* HST08 cells. The increase in the expression of *rfp* with the addition of IPTG to the growth media was checked (Figure 6C). We found that cells with the promoter replaced in the original pYK-1N5 plasmid did not show increased red fluorescence, indicating *rfp* in these cells is transcribed from the inner promoter of IS3. In contrast, cells with the promoter replaced in the plasmid of evolved cells showed significantly stronger fluorescence with IPTG induction. The baseline fluorescence of these cells without induction was significantly darker compared to that in the cells with plasmid

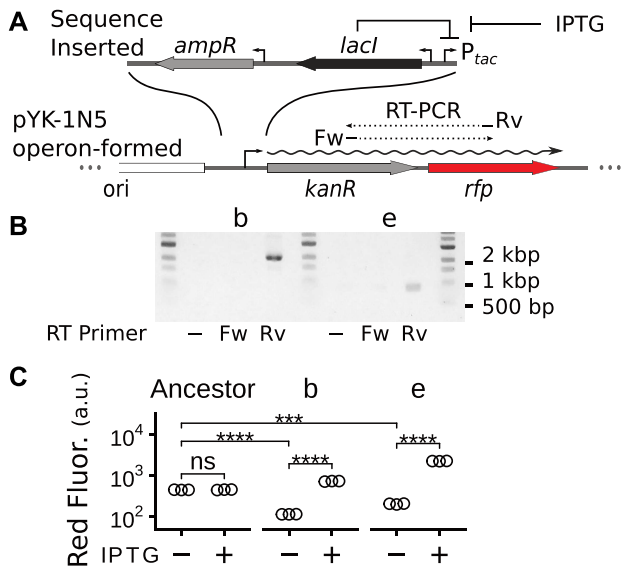


Figure 6. (A) The two experiments performed to confirm the formation of the *kanR-rfp* operon. (B) Agarose gel electrophoresis of RT-PCR products. The lengths were confirmed by Sanger sequencing. (C) Median red fluorescence measured by flow cytometry of cells with re-engineered plasmids ($***P < 10^{-3}$, $****P < 10^{-4}$). The constitutive promoter of *kanR* in pYK-1N5 was replaced with a DNA cassette containing an IPTG-inducible promoter (P_{tac}). The lower-case letters indicate cells corresponding to reads in Figure 5D with the same letter. Abbreviations: ampicillin resistance gene (*ampR*), lactose operon repressor gene (*lacI*), not significant (ns).

reconstructed from the ancestor plasmid. These results indicate that *rfp* in these cells was primarily transcribed from P_{tac} , and LacI tightly regulated their transcription.

Overall, these experiments strongly support that cells evolved to have *kanR* and *rfp* to be under the control of the same promoters by the IDE model.

DISCUSSION

Summary of the results

It is widely accepted that operon structures can change dramatically (1,9,10,23,44). However, since the mechanism of operon formation has generally been studied by comparing genomes at the level above species, the very mutations that form new operons have been missed.

To explain the formation of operons, we proposed a new model whereby ISs facilitate the formation of novel operons through consecutive insertion–deletion–excision reactions (IDE model, Figure 2A). According to this model, operons can form through IS locally accelerating the structural mutation rate (Figure 2B). Positive selection can facilitate this process by providing intermediate adaptive steps towards operon formation (Figure 2A, (7)). For experimental verification, we constructed a plasmid that mimics the result of the ‘insertion’ step of the proposed model (Figure 3), transformed it into an IS-less *E. coli* strain (Figure 4), and allowed it to evolve under positive selection for kanamycin resistance and intense red fluorescence. We found that cells showed fluorescence indicative of IS-mediated deletions, especially when IEE was present (Figure 5). The structural mutations due to the transposition of IS caused rapid evo-

lution that led to the formation of novel operons (Figure 6) overnight in 200 μ l cultures.

The IDE model is consistent with our understanding of the formation and maintenance of operons

The IDE model is consistent with how operons in prokaryotes are thought to be under selection for better gene coregulation (5,13–15,45) because adaptation for coregulation can begin once an operon has formed (7).

The IDE model is also consistent with our current understanding of how operons tend to form. For instance: (i) When new operons form, genes that are common across bacterial species are likely to be upstream of others (9), probably because this would preserve their sophisticated promoters even after deletion by ISs. (ii) When genes are added to existing operons, genes tend to be appended or prepended rather than inserted (9), probably because genes are added by ISs deleting redundant sequences.

The advantages of the IDE model

The IDE model can form operons not only (i) under positive selection but also (ii) under neutral evolution. This is because (iii) IS can locally accelerate the structural mutation rate during the ‘deletion’ step of operon formation. Also, (iv) it can work together with other mechanisms (Table 2, Supplementary Figure S7) to form the operon-rich genomes of prokaryotes.

Gradual and adaptive formation of operons. The IDE model can incorporate the advantage of the ‘selfish operon model’ that states that operons can form without relying on rare events that directly cluster two genes into operons ((7), Figure 2A). In our experimental demonstration, positive selection enabled us to rapidly acquire cells in the various phases of operon formation (Figure 5E).

Formation of suboptimal operons under weak selection. The expression levels of genes in operons are generally suboptimal (48), and genes in operons consisting of genes that are not functionally related tend to have functions important for cell growth (44). Our model is in line with these observations, as it explains operon formation without relying on the often-assumed selection for better gene coregulation. Consistently, various tendencies of operon formation, including the tendencies above, can be explained by assuming that operons form by selection for the preservation of genes rather than the selection for their coregulation (9,12).

This explains an observation that inspired our new model, that is, some endosymbionts that have experienced extreme population bottlenecks (49) have small but operon-rich genomes (50,51). The selection for better coregulation would be too weak in these organisms for operons to form. Because insect symbionts generally reduce their genome size through a burst and a subsequent loss of ISs (18,52), in line with our study, ISs may have formed their operon-rich genomes.

Our model improves the ‘persistence model’ (12), which currently relies upon the selection for coregulation to explain the formation of operons (4,13). Unlike the ‘persistence model,’ we have attributed the large tandem deletions

Table 2. Models of operon formation and mutation mechanisms referred to in this study

	Regulatory Models ^(a)	Selfish Operon Model ^(b)	Persistence Model ^(c)
	<p>Better Coregulation</p> <ul style="list-style-type: none"> • Stochastic Expression • Protein Complex • Balance Molarity etc. 	<p>HGT</p> <p>Operon-level Selection</p>	<p>Tandem Deletions</p> <p>Selection for increased robustness</p>
What it explains Selection: Intermediate state Selection: End product Mutations enhancing operon formation Hot spot of operon formation Reference	Why operons are prevalent - Coregulation - - (5, 6, 7, 8)	How gene clusters form Efficient horizontal gene transfer - Horizontal gene transfer Mobile elements (7)	Why gene clusters are prevalent Robustness against tandem deletion - Tandem deletions - (12)
Recombination Model^(d)	Scribbling Pad Model^(e)	SNAP Hypothesis^(f)	IDE Model^(g)
<p>Selection against breaking gene pairs</p>	<p>Rapid mutation on plasmid "Scribbling Pad"</p>	<p>Duplication and Selection</p>	<p>Insertion</p> <p>IS</p> <p>Deletion</p> <p>Excision</p>
Why gene clusters are prevalent - Maintaining gene pairs Random recombination Phages (11)	Where operons form - Recombination on plasmids Plasmids (46)	How gene order evolves Stronger expression of duplicated genes - Duplication - (47)	How operons form Stronger expression of dormant genes Stabilization of adaptive operons IS-mediated IDE and recombination Around ISs This study

(a) Regulatory models. Once an operon has formed, it can evolve complex regulation through positive selection (right arrow). Operon structures are selected because it is easier for them to evolve into conducting complex regulation compared to that of multiple promoters. **(b) Selfish operon model.** The closer a pair of genes are, the more likely they are transferred together by horizontal gene transfer (HGT, right arrow). Frequent HGT, thus, causes selection at the gene cluster level to gradually shorten sequences in between, which the lower limit is operon formation. **(c) Persistence model.** Genomes in which genes important for survival ('persistent' genes) are dispersed are statistically more vulnerable to tandem deletions than genomes in which these genes are clustered (cross indicates lethal deletions). This negative selection against sequences between 'persistent' genes under frequent tandem deletions results in operon-rich genomes with closer spacings between the 'persistent' genes. **(d) Recombination model.** In genomes that undergo frequent recombination, operons are selected because co-functional gene sets are less likely to be disrupted by recombination. **(e) Scribbling pad model.** Plasmids have high structural mutation rates compared to chromosomes. Sets of genes in the chromosomes are transferred to plasmids and form operons. **(f) SNAP Hypothesis.** Frequent reordering of genes is explained by a rapid two-step process: duplication (horizontal bar) of beneficial genes in a new niche, which rapidly reaches fixation, and subsequent loss of redundant or superfluous genes. Sometimes, this rearrangement forms operons beneficial in the new niche. **(g) IDE model.** ISs 'catalyze' operon formation by IDE reactions. An IS Inserts itself between two genes via transposition, and Deletes adjacent sequences bringing the two genes closer together. Finally, Excision of IS results in operon formation without leaving traces of the IS.

that cluster genes to the activity of ISs (17,18). The activity of ISs alters the expression of surrounding genes (41). This enables cells with new IS insertions to become adaptive. In addition, this ensured that the premature loss of ISs after the deletion was maladaptive (Figure 2A). It seems that the structure of ISs that have evolved for their proliferation may have made them a source of mutation particularly suitable for the formation of operons.

Accelerated local structural mutation rate. Unlike other mechanisms of large deletions (53) that may lead to operon formation, a prominent feature of the IDE model is that it relies upon the activity of ISs. ISs not only undergo the IDE reactions but actively transpose and are both drivers (54) and major targets of homologous recombination (55). Therefore, they can accelerate the formation of operons by transiently increasing the local structural mutation rates surrounding the IS ((17,30), Figure 2B). How the increased

mutation rate is kept local and transient is critical for our model because the rate of recombination can exceed the upper limit of global mutation rates set by the population size (56). In contrast, previous models based on random recombination ('recombination model' (11)) have been refuted because the rate of random recombination in prokaryotes seems to be too low to account for operon formation (7). Moreover, a reproducible experimental demonstration of operon formation (Figure 5), unprecedented in any other models of operon formation, was possible because of the rapid mutation by IS3.

In addition to enabling multiple pairs of beneficial genes to be tested for undergoing the deletion step (Figure 2B), recombination can also help ISs to find loci where they can "creep" toward genes to form new operons without encountering essential genes. Within prokaryotic genomes, some loci are easier to have only dispensable genes between two beneficial genes such as: pathogenicity islands (13),

plasmids ('Scribbling pad hypothesis' (46)), and clusters of nonessential genes (12,16). ISs can actively form such loci by promoting recombination. For instance, duplication by IS-mediated recombination can create redundant copies of essential genes, making them nonessential ('SNAP hypothesis,' (47), Supplementary Figure S7E). In addition, recombination of two copies of ISs can form plasmids (24) and become hot spots for undergoing the IDE reactions (Supplementary Figure S7D).

The IDE model complements other models of operon formation. The operon-rich genomes of prokaryotes may have been shaped through the coexistence of ISs in the prokaryotic genome, whereby operons rapidly form through the IDE model. Combination of this model with other models such as the 'SNAP hypothesis' and 'Scribbling pad hypothesis' can facilitate both the formation and enrichment of operons (Supplementary Figure S7). On the other hand, the formation of some operons is better explained by other models. For instance, new operons can form by replacing the genes of preexisting operons (9,47) or by duplication (10). Post the formation of operons, beneficial operons can acquire sophisticated regulation and get fixed ('regulatory model'). These operons persist across diverse prokaryotes due to the 'persistence model' and 'selfish operon model.'

The activity of transposase may decrease the expression of genes surrounding an IS

An important characteristic of ISs is that they alter the expression of surrounding genes through their internal promoters (41). We observed that the increased activity of the IS3 transposase significantly decreased the expression of neighboring genes encoding fluorescent proteins (Figure 4). However, this did not seem to be due to the deletion of *rfp* and *yfp* (Supplementary Figure S4). Moreover, no growth defect was observed (Supplementary Figure S2B), implying that the reduced copy number of plasmids due to their excision was not the cause because pYK-IN5 is a low copy number plasmid and decreasing its copy number should have been detrimental to growth.

Rather, we speculate that the decreased fluorescence could have been due to the nicking of DNA by the transposase, the binding of transposase to the IRs, the formation of a protein DNA complex called 'Figure-eight' (40), or the increased transcription itself strengthening the supercoiling of DNA (57). While solving the dominant cause of the decreased fluorescence is out of our scope, we speculate that ISs may act as global regulators, with IS transposase acting as transcription repressors and IS inverted repeats as operator sequences. Supporting this, ISs have been found in orientations that interfere with neighboring genes in clinical samples of *Staphylococcus aureus* (58).

Future extensions of our experimental model

This study has demonstrated experimentally, perhaps for the first time, the conditions under which new operons can form rapidly and reproducibly in the laboratory. Using our experimental method as a basis, we believe that future research will be able to address questions about operon for-

mation that have remained unexplored. We propose the following extensions.

First, the efficiency and mechanisms of the deletion and excision steps require further analysis. In general, large deletions can occur at significant rates (53) via mechanisms such as break repairs by homologous recombination and end-joining using microhomologies (27) and replication errors. In accordance with our study, ISs promote deletions around the IRs (20–22,25,28,55). ISs seem to be a major source of deletion (17,55), but various mechanisms may contribute to operon formation. How different genetic backgrounds, such as the presence and absence of IEE, might influence the relative contributions of each mechanism to operon formation remains poorly understood (17). FACS, which was used in our study, is not particularly suitable for quantifying low operon formation rates, as seen in the *-iee* condition after a day of evolution. Extending our system to compare the rates of deletions and excisions with and without ISs or IEE would facilitate our understanding of the contribution of the IDE model in establishing extant operon structures.

Second, we believe that extending our experimental model using chromosomal constructs would better illustrate the strength of the IDE model. In genomes with multiple copies of an IS in the chromosomes, recombination among the copies likely enhances operon formation by the IDE model (Figure 2B). Since new operons seem to have formed after decades of laboratory evolution of *E. coli* (59), using chromosomal constructs may result in the formation of diverse new operon structures. Such an experiment would clarify the extent to which the IDE model can form new operons in prokaryotic genomes and might also clarify phenomena that may have occurred using a plasmid-based system, such as changes in fluorescence intensity due to the heterogeneity among multiple copies.

Finally, to demonstrate operon formation, we evolved operons under a simple fitness landscape: the stronger the expression of *rfp*, the fitter the cells were. However, a notable strength of the IDE model is that operon formation can be neutral, beneficial, or even deleterious. Future studies might examine whether the high local mutation rates due to IS activity may even render deleterious routes of operon formation to occur frequently enough to be observed.

Validation of operon formation by the IDE model is difficult in nature

A major limitation of our study is that the proposed model still needs validation through identification of examples in nature. This is important as ISs are regarded as drivers of operon destruction (19,23), although this view may be biased because studies have focused more on the evolution of conserved operons than on their formation.

A potential difficulty to find examples of operon formation in nature is that IS can be thought of as a genetic "catalyst" that locally accelerates the mutation rate and transforms a genome into a new genome with an additional operon (Figure 2). As a catalyst, the activity of IS is virtually traceless. As a result, previous studies have overlooked the role of deletions by IS in operon formation because, based on parsimony of events, operon formation by the IDE model cannot be distinguished from large tandem deletions without ISs (9). Indeed, partial sequences of ISs may re-

main as read *g* of Figure 5D, but these sequences would be rapidly lost, as they are unlikely to have any function. We believe that to find cases where operons are formed by our model and to determine whether ISs are the major drivers of operon formation, analyzing ongoing evolutions in detail (28,55,60,61) to detect the intermediate steps of operon formation is essential.

We postulate that these steps may potentially be identified in organisms that are under reductive evolution (62), which ISs are known to facilitate (18,63). When genomes expand, ISs can potentiate newly acquired genes to form operons by activating them. As genome size decreases, ISs delete redundant sequences and excise themselves, forming new operons. Supporting this idea, a systematic study of genome size evolution in cyanobacteria showed that operon-rich genomes tend to have experienced genome reduction (14). In contrast, larger genomes with many ISs are relatively poor in operons (4), perhaps because they are yet to experience an upsurge in operon formation. Future studies on, for example, the evolution of pathogenic *E. coli* with multiple ISs that degrade their genomes (64) and are excised by IEE (65), or organisms with genome size evolution artificially accelerated using hyper-active transposons (66) may provide a clearer picture of the coevolution of prokaryotic genomic and regulatory architecture with ISs, as suggested by our study.

DATA AVAILABILITY

Nanopore sequencing data have been deposited with SRA (BioProject PRJNA768397). The raw flow cytometry data are available in FlowRepository (FR-FCM-Z4LV).

SUPPLEMENTARY DATA

[Supplementary Data](#) are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank the following people for generously sharing their reagents: The mScarlet-I sequence was generated by mutating the mScarlet sequence, which was a gift from Dr Ryudo Ohbayashi. The Tn5 transposase sequence used as *-tpn* control was a gift from Dr Ronald Chalmers. The Marionette Sensor Collection was a gift from Dr Christopher Voigt (Addgene #1000000137). pCDSSara-L18 was a gift from Dr Peter Schultz.

Author contributions: Conceptualization and methodology: Y.K. and S.T.; data curation, formal analysis, investigation, validation and visualization: Y.K.; funding acquisition: Y.K., S.T. and C.F.; project administration: S.T.; resources and supervision: S.T. and C.F.; writing – original draft: Y.K.; writing – review & editing: Y.K., S.T. and C.F.

FUNDING

Japan Society for the Promotion of Science [21J20693 to Y.K., 18H02427 to S.T., 17H06389 to C.F., 19H05626 to C.F.]; Japan Science and Technology Agency [JPMJER1902 to C.F.]. Funding for open access charge: the Japan Science and Technology Agency.

Conflict of interest statement. None declared.

REFERENCES

- Wolf, Y.I., Rogozin, I.B., Kondrashov, A.S. and Koonin, E.V. (2001) Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res.*, **11**, 356–372.
- Koonin, E.V. and Wolf, Y.I. (2008) Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.*, **36**, 6688–6719.
- Jacob, F. and Monod, J. (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.*, **3**, 318–356.
- Núñez, P.A., Romero, H., Farber, M.D. and Rocha, E.P. (2013) Natural selection for operons depends on genome size. *Genome Biol. Evol.*, **5**, 2242–2254.
- Price, M.N., Huang, K.H., Arkin, A.P. and Alm, E.J. (2005) Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome Res.*, **15**, 809–819.
- Sneppen, K., Pedersen, S., Krishna, S., Dodd, I. and Semsey, S. (2010) Economy of operon formation: Cotranscription minimizes shortfall in protein complexes. *MBio*, **1**, 4–6.
- Lawrence, J.G. and Roth, J.R. (1996) Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics*, **143**, 1843–1860.
- Shieh, Y.W., Minguez, P., Bork, P., Auburger, J.J., Guilbride, D.L., Kramer, G. and Bukau, B. (2015) Operon structure and cotranslational subunit association direct protein assembly in bacteria. *Science*, **350**, 678–680.
- Price, M.N., Arkin, A.P. and Alm, E.J. (2006) The life-cycle of operons. *PLoS Genet.*, **2**, 0859–0873.
- Fani, R., Brilli, M. and Liò, P. (2005) The origin and evolution of operons: the piecewise building of the proteobacterial histidine operon. *J. Mol. Evol.*, **60**, 378–390.
- Stahl, F.W. and Murray, N.E. (1966) The evolution of gene clusters and genetic circularity in microorganisms. *Genetics*, **53**, 569–576.
- Fang, G., Rocha, E.P. and Danchin, A. (2008) Persistence drives gene clustering in bacterial genomes. *BMC Genomics*, **9**, 4.
- Touchon, M. and Rocha, E.P. (2016) Coevolution of the organization and structure of prokaryotic genomes. *Cold Spring Harb. Perspect. Biol.*, **8**, a018168.
- Memon, D., Singh, A.K., Pakrasi, H.B. and Wangikar, P.P. (2013) A global analysis of adaptive evolution of operons in cyanobacteria. *Antonie Van Leeuwenhoek*, **103**, 331–346.
- Pál, C. and Hurst, L.D. (2004) Evidence against the selfish operon theory. *Trends Genet.*, **20**, 232–234.
- Hosseini, S.R. and Wagner, A. (2018) Genomic organization underlying deletional robustness in bacterial metabolic systems. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, 7075–7080.
- Lee, H., Doak, T.G., Popodi, E., Foster, P.L. and Tang, H. (2016) Insertion sequence-caused large-scale rearrangements in the genome of *Escherichia coli*. *Nucleic Acids Res.*, **44**, 7109–19.
- Siguié, P., Gourbeyre, E. and Chandler, M. (2014) Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiol. Rev.*, **38**, 865–891.
- Brandis, G. (2021) Reconstructing the evolutionary history of a highly conserved operon cluster in *Gammaproteobacteria* and *Bacilli*. *Genome Biol. Evol.*, **13**, evab041.
- Reif, H.J. and Saedler, H. (1975) IS1 is involved in deletion formation in the gal region of *E. coli* K12. *Mol. Gen. Genet.*, **137**, 17–28.
- Sekine, Y., Eisaki, N. and Ohtsubo, E. (1994) Translational control in production of transposase and in transposition of insertion sequence IS3. *J. Mol. Biol.*, **235**, 1406–1420.
- Jilk, R.A., Makris, J.C., Borchardt, L. and Reznikoff, W.S. (1993) Implications of Tn5-associated adjacent deletions. *J. Bacteriol.*, **175**, 1264–1271.
- Itoh, T., Takemoto, K., Mori, H. and Gojobori, T. (1999) Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Mol. Biol. Evol.*, **16**, 332–346.
- Snyder, L.M.H.T., Peters, J.E. and Champness, W. (2013) In: *Molecular Genetics of Bacteria*. ASM Press, Washington, D.C.
- Kusumoto, M., Ooka, T., Nishiyama, Y., Ogura, Y., Saito, T., Sekine, Y., Iwata, T., Akiba, M. and Hayashi, T. (2011) Insertion sequence-excision enhancer removes transposable elements from

- bacterial genomes and induces various genomic deletions. *Nat. Commun.*, **2**, 152.
26. Price, M.N., Huang, K.H., Alm, E.J. and Arkin, A.P. (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res.*, **33**, 880–892.
 27. Chayot, R., Montagne, B., Mazel, D. and Ricchetti, M. (2010) An end-joining repair mechanism in *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 2141–2146.
 28. Ooka, T., Ogura, Y., Asadulghani, M., Ohnishi, M., Nakayama, K., Terajima, J., Watanabe, H. and Hayashi, T. (2009) Inference of the impact of insertion sequence (IS) elements on bacterial genome diversification through analysis of small-size structural polymorphisms in *Escherichia coli* O157 genomes. *Genome Res.*, **19**, 1809–1816.
 29. Chain, P.S., Hu, P., Malfatti, S.A., Radnedge, L., Larimer, F., Vergez, L.M., Worsham, P., Chu, M.C. and Andersen, G.L. (2006) Complete genome sequence of *Yersinia pestis* strains antiqua and Nepal516: Evidence of gene reduction in an emerging pathogen. *J. Bacteriol.*, **188**, 4453–4463.
 30. Cooper, V.S., Schneider, D., Blot, M. and Lenski, R.E. (2001) Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of *Escherichia coli* B. *J. Bacteriol.*, **183**, 2834–2841.
 31. Pósfai, G., Plunkett, G., Fehér, T., Frisch, D., Keil, G.M., Umenhoffer, K., Kolisnychenko, V., Stahl, B., Sharma, S.S., de Arruda, M., Burland, V., Harcum, S.W. and Blattner, F.R. (2006) Emergent properties of reduced-genome *Escherichia coli*. *Science*, **312**, 1044–1046.
 32. Shoji, S., Dambacher, C.M., Shajani, Z., Williamson, J.R. and Schultz, P.G. (2011) Systematic chromosomal deletion of bacterial ribosomal protein genes. *J. Mol. Biol.*, **413**, 751–761.
 33. Blundell-Hunter, G., Tellier, M. and Chalmers, R. (2018) Transposase subunit architecture and its relationship to genome size and the rate of transposition in prokaryotes and eukaryotes. *Nucleic Acids Res.*, **46**, 9637–9646.
 34. Bindels, D.S., Haarbosch, L., Van Weeren, L., Postma, M., Wiese, K.E., Mastop, M., Aumonier, S., Gotthard, G., Royant, A., Hink, M.A. *et al.* (2016) mScarlet: a bright monomeric red fluorescent protein for cellular imaging. *Nat. Methods*, **14**, 53–56.
 35. Nagai, T., Ibata, K., Park, E.S., Kubota, M., Mikoshiba, K. and Miyawaki, A. (2002) A variant of yellow fluorescent protein with fast and efficient maturation for cell-biological applications. *Nat. Biotechnol.*, **20**, 87–90.
 36. Lutz, R. and Bujard, H. (1997) Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I₁-I₂ regulatory elements. *Nucleic Acids Res.*, **25**, 1203–1210.
 37. Meyer, A.J., Segall-Shapiro, T.H., Glassey, E., Zhang, J. and Voigt, C.A. (2019) *Escherichia coli* ‘Marionette’ strains with 12 highly optimized small-molecule sensors. *Nat. Chem. Biol.*, **15**, 196–204.
 38. Silander, O.K., Nikolic, N., Zaslaver, A., Bren, A., Kikoin, I., Alon, U. and Ackermann, M. (2012) A genome-wide analysis of promoter-mediated phenotypic noise in *Escherichia coli*. *PLoS Genet.*, **8**, e1002443.
 39. Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.
 40. Chandler, M., Fayet, O., Rousseau, P., Hoang, B.T. and Duval-valentin, G. (2015) Copy-out–paste-in transposition of IS911: a major transposition pathway. *Microbiol. Spectr.*, **3**, <https://doi.org/10.1128/microbiolspec.MDNA3-0031-2014>.
 41. Vandecraen, J., Chandler, M., Aertsen, A. and Van Houdt, R. (2017) The impact of insertion sequences on bacterial genome plasticity and adaptability. *Crit. Rev. Microbiol.*, **43**, 709–730.
 42. Charlier, D., Piette, J. and Glansdorff, N. (1982) IS3 can function as a mobile promoter in *E. coli*. *Nucleic Acids Res.*, **10**, 5935–5948.
 43. Fehér, T., Bogos, B., Méhi, O., Fekete, G., Csörgő, B., Kovács, K., Pósfai, G., Papp, B., Hurst, L.D. and Pál, C. (2012) Competition between transposable elements and mutator genes in bacteria. *Mol. Biol. Evol.*, **29**, 3153–3159.
 44. Rogozin, I.B., Makarova, K.S., Murvai, J., Czabarka, E., Wolf, Y.I., Tatusov, R.L., Szekeley, L.A. and Koonin, E.V. (2002) Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Res.*, **30**, 2212–2223.
 45. Lim, H.N., Lee, Y. and Hussein, R. (2011) Fundamental relationship between operon organization and gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 10626–10631.
 46. Norris, V. and Merieau, A. (2013) Plasmids as scribbling pads for operon formation and propagation. *Res. Microbiol.*, **164**, 779–787.
 47. Brandis, G. and Hughes, D. (2020) The SNAP hypothesis: chromosomal rearrangements could emerge from positive Selection during Niche Adaptation. *PLoS Genet.*, **16**, e1008615.
 48. Price, M.N., Deutschbauer, A.M., Skerker, J.M., Wetmore, K.M., Ruths, T., Mar, J.S., Kuehl, J.V., Shao, W. and Arkin, A.P. (2013) Indirect and suboptimal control of gene expression is widespread in bacteria. *Mol. Syst. Biol.*, **9**, 660.
 49. Moran, N.A. (2002) Microbial minimalism: Genome reduction in bacterial pathogens. *Cell*, **108**, 583–586.
 50. Brinza, L., Calevro, F., Duport, G., Gaget, K., Gautier, C. and Charles, H. (2010) Structure and dynamics of the operon map of *Buchnera aphidicola* sp. strain APS. *BMC Genomics*, **11**, 666.
 51. Güell, M., van Noort, V., Yus, E., Chen, W.-H., Leigh-Bell, J., Michalodimitrakis, K., Yamada, T., Arumugam, M., Doerks, T., Kühner, S. *et al.* (2009) Transcriptome complexity in a genome-reduced bacterium. *Science*, **326**, 1268–1271.
 52. McCutcheon, J.P. and Moran, N.A. (2012) Extreme genome reduction in symbiotic bacteria. *Nat. Rev. Microbiol.*, **10**, 13–26.
 53. Koskiniemi, S., Sun, S., Berg, O.G. and Andersson, D.I. (2012) Selection-driven gene loss in bacteria. *PLoS Genet.*, **8**, e1002787.
 54. Reams, A.B., Kofoid, E., Kugelberg, E. and Roth, J.R. (2012) Multiple pathways of duplication formation with and without recombination (RecA) in *Salmonella enterica*. *Genetics*, **192**, 397–415.
 55. Raeside, C., Gaffé, J., Deatherage, D.E., Tenaillon, O., Briska, A.M., Ptashkin, R.N., Cruveiller, S., Médigue, C., Lenski, R.E., Barrick, J.E. *et al.* (2014) Large chromosomal rearrangements during a long-term evolution experiment with *Escherichia coli*. *MBio*, **5**, e01377-14.
 56. Lynch, M., Ackerman, M.S., Gout, J.F., Long, H., Sung, W., Thomas, W.K. and Foster, P.L. (2016) Genetic drift, selection and the evolution of the mutation rate. *Nat. Rev. Genet.*, **17**, 704–714.
 57. Yeung, E., Dy, A.J., Martin, K.B., Ng, A.H., Del Vecchio, D., Beck, J.L., Collins, J.J. and Murray, R.M. (2017) Biophysical constraints arising from compositional context in synthetic gene networks. *Cell Syst.*, **5**, 11–24.
 58. Yebra, G., Haag, A.F., Neamah, M.M., Wee, B.A., Richardson, E.J., Horcajo, P., Granneman, S., Tormo-Más, M.Á., de la Fuente, R., Fitzgerald, J.R. *et al.* (2021) Radical genome remodelling accompanied the emergence of a novel host-restricted bacterial pathogen. *PLoS Pathog.*, **17**, e1009606.
 59. Blount, Z.D., Barrick, J.E., Davidson, C.J. and Lenski, R.E. (2012) Genomic analysis of a key innovation in an experimental *Escherichia coli* population. *Nature*, **489**, 513–518.
 60. Consuegra, J., Gaffé, J., Lenski, R.E., Hindré, T., Barrick, J.E., Tenaillon, O. and Schneider, D. (2021) Insertion-sequence-mediated mutations both promote and constrain evolvability during a long-term experiment with bacteria. *Nat. Commun.*, **12**, 980.
 61. He, S., Chandler, M., Varani, A.M., Hickman, A.B., Dekker, J.P. and Dyda, F. (2016) Mechanisms of evolution in high-consequence drug resistance plasmids. *MBio*, **7**, e01987-16.
 62. Wolf, Y.I. and Koonin, E.V. (2013) Genome reduction as the dominant mode of evolution. *BioEssays*, **35**, 829–837.
 63. van Dijk, B., Bertels, F., Stolk, L., Takeuchi, N. and Rainey, P.B. (2021) Transposable elements promote the evolution of genome streamlining. *Phil. Trans. R. Soc. B*, **377**, 20200477.
 64. Hawkey, J., Monk, J.M., Billman-Jacobe, H., Pálsson, B. and Holt, K.E. (2020) Impact of insertion sequences on convergent evolution of *Shigella* species. *PLoS Genet.*, **16**, e1008931.
 65. Toro, M., Rump, L.V., Cao, G., Meng, J., Brown, E.W. and Gonzalez-Escalona, N. (2015) Simultaneous presence of insertion sequence excision enhancer and insertion sequence IS629 correlates with increased diversity and virulence in Shiga toxin-producing *Escherichia coli*. *J. Clin. Microbiol.*, **53**, 3466–3473.
 66. Verniyk, V., Karcagi, I., Timár, E., Nagy, I., Györkei, Á., Papp, B., Györfy, Z. and Pósfai, G. (2020) Exploring the fitness benefits of genome reduction in *Escherichia coli* by a selection-driven approach. *Sci. Rep.*, **10**, 7345.