Nucleic Acids Research, 2025, **53**, gkaf331
https://doi.org/10.1093/nar/gkaf331
Nucleic Acid Enzymes

Laboratory evolution of the bacterial genome structure through insertion sequence activation

Yuki Kanai [1], Atsushi Shibai [2], Naomi Yokoi [2], Saburo Tsuru [3,*], Chikara Furusawa [2,3,*]

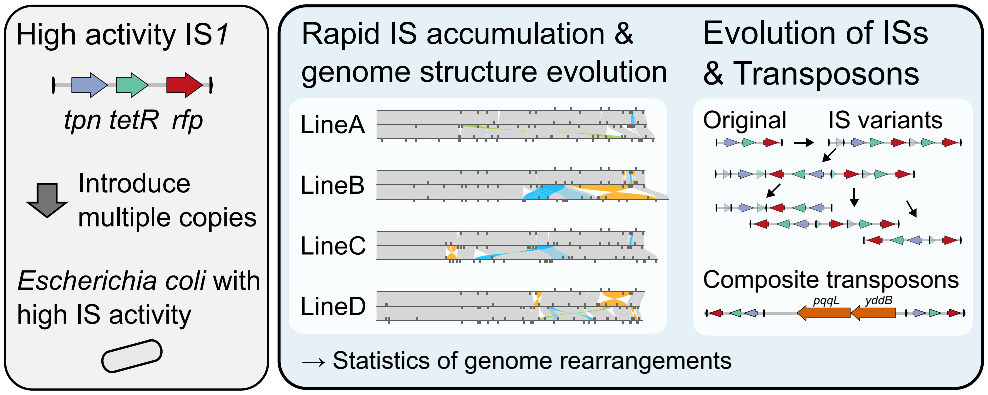
[1] Department of Biological Sciences, The University of Tokyo, 7-3-1 Hongo, 113-0033 Tokyo, Japan
[2] Center for Biosystems Dynamics Research, RIKEN, 6-7-1 Minatojima-minamimachi, Chuo-ku, 650-0047 Kobe, Japan
[3] Universal Biology Institute, The University of Tokyo, 7-3-1 Hongo, 113-0033 Tokyo, Japan

[*] To whom correspondence should be addressed. Email: chikara.furusawa@riken.jp
Correspondence may also be addressed to Saburo Tsuru. Email: tsuru@ubi.s.u-tokyo.ac.jp

Abstract

The genome structure fundamentally shapes bacterial physiology, ecology, and evolution. Though insertion sequences (IS) are known drivers of drastic evolutionary changes in the genome structure, the process is typically slow and challenging to observe in the laboratory. Here, we developed a system to accelerate IS-mediated genome structure evolution by introducing multiple copies of a high-activity IS in *Escherichia coli*. We evolved the bacteria under relaxed neutral conditions, simulating those leading to IS expansion in host-restricted endosymbionts and pathogens. Strains accumulated a median of 24.5 IS insertions and underwent over 5% genome size changes within ten weeks, comparable to decades-long evolution in wild-type strains. The detected interplay of frequent small deletions and rare large duplications updates the view of genome reduction under relaxed selection from a simple consequence of the deletion bias to a nuanced picture including transient expansions. The high IS activity resulted in structural variants of IS and the emergence of composite transposons, illuminating potential evolutionary pathways for ISs and composite transposons. The extensive genome rearrangements we observed establish a baseline for assessing the fitness effects of IS insertions, genome size changes, and rearrangements, advancing our understanding of how mobile elements shape bacterial genomes.

Graphical abstract


High activity IS*1*
tpn tetR rfp
Introduce multiple copies
Escherichia coli with high IS activity
Rapid IS accumulation & genome structure evolution
LineA
LineB
LineC
LineD
Evolution of ISs & Transposons
Original
IS variants
Composite transposons
pqqL
yddB
→ Statistics of genome rearrangements

Introduction

The genome structure, including gene number and arrangement, critically impacts bacterial physiology, ecology, and evolution [1, 2]. A key driver of its evolution is insertion sequences (IS), ubiquitous DNA transposons in bacterial genomes, activating, disrupting, and reordering genes through recombination [3]. For instance, most genome rearrangements during the two decades of the long-term evolution experiment of *Escherichia coli* (LTEE) were mediated by ISs [4]. In nature, ISs are particularly active in bacterial endosymbionts and pathogens, whose genomes can harbor hundreds of IS copies [5, 6]. The stable and possibly nutrient-rich environments within hosts, combined with frequent population size bottlenecks during transmission between hosts, relax selection, allowing the accumulation of weakly deleterious mutations, including the increase of IS copies [7]. ISs can disrupt genes upon insertion, and the proliferated ISs can delete sequences through tandem deletions and homologous recombinations, leading to genome reduction, even down to one-tenth of the size of free-living relatives [8].

Despite the abundance of sequence data, inferring the dynamics, including the intermediate steps of genome

Received: November 22, 2024. **Revised:** April 8, 2025. **Editorial Decision:** April 10, 2025. **Accepted:** April 14, 2025

structure evolution, can be difficult [9]. Moreover, disentangling the potential factors driving these dynamics is fundamentally challenging, as the conditions of evolution in nature cannot be controlled. Laboratory evolution offers an alternative approach, allowing researchers to directly observe evolution under controlled conditions, track the dynamics, and test evolutionary hypotheses [10], including the role of ISs in genome reduction.

Previous studies struggled to observe IS-mediated evolution in the laboratory primarily due to its slow pace. The model bacterium *E. coli* has a genome that includes almost all the genes of the genome-reduced aphid endosymbiont *Buchnera aphidicola* [11] and has pathogenic strains harboring hundreds of ISs [12]. This suggests that *E. coli* has the potential for intense IS-mediated reductive genome evolution in the laboratory. However, IS transposition in *E. coli* typically occurs only once per year (every few thousand generations) [12–14]. Plague attempted to simulate natural IS expansions by evolving *E. coli* for 4000 generations under relaxed selection but failed to detect any IS expansion [15]. To observe the IS-mediated genome evolution of *E. coli*, previous studies either required decades, as in the LTEE [14], or involved evolving hundreds of lines in parallel, as in the mutation accumulation (MA) experiment by Foster's group [13]. These studies highlight the need for a more efficient method to study IS-mediated genome evolution under well-defined conditions.

Here, we introduce a method to rapidly observe IS-mediated genome evolution in the laboratory. Plague attributed their failure to observe IS expansion to the low IS activity of the wild-type strain of *E. coli* compared to bacteria experiencing an increase in IS activity in nature [15]. Simulating the conditions of such IS expansion, we inserted multiple copies of high-activity ISs into the genome of *E. coli* and evolved 44 lines under a stable, nutrient-rich, and low-population-size condition. This approach enabled us to observe extensive IS-mediated genome structure evolution in just 10 weeks. Our study provides a reference for studying the fitness effects of IS insertions, genome size changes, and rearrangements, and demonstrates a powerful method for experimentally investigating bacterial genome structure evolution under controlled laboratory conditions.

Materials and methods

Reagents

We used the following reagents and instruments:

For cell cultures during the MA, we used LB broth, Miller (NACALAI TESQUE, Japan, 20068-75), LB agar, Miller (NACALAI TESQUE, Japan, 20069-65), anhydrotetracycline hydrochloride (aTc; Sigma–Aldrich, USA, 37919), chloramphenicol (Wako, Japan, 034-10572), dimethyl sulfoxide (Wako, Japan, 043-07216), and ethanol (99.5) (Wako, Japan, 057-00456).

The master stock of aTc was prepared at 5 mg/ml in dimethyl sulfoxide, and the working stock was prepared by diluting to 100 μM with ethanol and stored at $-20\,^{\circ}$C. Chloramphenicol was dissolved in ethanol and stored at $-20\,^{\circ}$C.

For handling nucleic acids and DNA sequencing, we used KOD One PCR Master Mix Blue (Toyobo, Japan, KMM-201), exonuclease V (RecBCD) (New England Biolabs, USA, M0345S), Rapid Barcoding Kit 96 (Oxford Nanopore Technologies, UK, SQK-RBK110.96), Native Barcoding Kit 24 V14 (Oxford Nanopore Technologies, UK, SQK-NBD114), In-Fusion Snap Assembly Master Mix (Takara Bio, Japan, 638948), Phenol/Chloroform/Isoamyl alcohol (25:24:1) (NIPPON Genetics, Japan, 311-90151), and Wizard HMW DNA Extraction Kit (Promega, USA, A2920).

We used the following instruments: Infinite F200 multimode plate reader (Tecan, Switzerland), Blue/green LED transilluminator (NIPPON Genetics, Japan, LB-16BG), FACSAria III (BD, USA), Flongle Flow Cell R9.4.1 (Oxford Nanopore Technologies, UK, FLO-FLG001), Minion Flow Cell R9.4.1 (Oxford Nanopore Technologies, UK, FLO-MIN106D), MinION Mk1B (Oxford Nanopore Technologies, UK, MIN-101B), and Minion Flow Cell R10.4.1 (Oxford Nanopore Technologies, UK, FLO-MIN114).

Biological resources

We used the following strains and plasmids: *E. coli* MDS42 [16], BW25113 *recA* (JW2669-KC, NBRP-E.coli at NIG, Japan) [17], pKD46 [18], pCP20 [19], and pAJM.011 [20].

The following plasmids were constructed in this study: pKD46_tetR and pYK-2X8 (plasmid with IS*1*-YK2X8). The maps are provided in Supplementary Fig. S1, and sequences are provided as in the "Data availability" section.

Preparation of strains and plasmids

To observe IS-mediated evolution in the laboratory, we designed a high-activity IS, IS*1*-YK2X8 (Fig. 1A), based on IS*1*, one of the most active IS in wild-type *E. coli* [13]. We made the following modifications to IS*1*. The wild-type transposase gene of IS*1* (*tpn*) has a frameshift in its coding sequence, significantly reducing its activity [21]. The A_6C mutation was introduced to the transposase gene to fix the frameshift to recover IS activity [21]. *tpn* was expressed from a strong inducible promoter based on $P_{\mathrm{LtetO\text{-}1}}$ [9]. *tetR* from pAJM.011 repressed the promoter [20], preventing unintended IS activity. aTc was added to culture media to cancel the repression and induce the expression of *tpn* to promote IS activity. Starting from an increased copy number of IS would likely increase the rate of IS-mediated genome evolution. To increase the copy number of IS based on fluorescence intensity, as described later, we introduced *rfp* (mScarlet-I) from pYK-1N5 [9]. Furthermore, transcription at the ends of the IS interferes with transposition [22]. To avoid promoter activities, we inserted strong terminators at the ends (*rrnB* T1 and L3S3P21 [23]). Inadvertent introduction of promoters was avoided by checking for potential promoter sequences computationally [24]. To avoid recombination within the IS*1* as much as possible, we checked for potentially unstable repetitive sequences using the EFM calculator [25]. For example, *tpn* overlaps with the terminal sequence of ISs called the inverted repeat (IR). The recombination between the IR and *tpn* may result in the unwanted loss or duplication of *tetR* and *rfp*. We reassigned the codons of the 3′-end of *tpn* to remove this homology. The DNA sequence of IS*1*-YK2X8 is provided with the custom scripts in the fasta directory and the plasmid map is provided in Supplementary Fig. S1.

A copy of IS*1*-YK2X8 was introduced into the genome of an IS-free strain of *E. coli*, MDS42 [16]. We used this strain to prevent native ISs from interfering with the IS*1*-YK2X8 activity. The IS was introduced using lambda red recombination [18] with plasmid pKD46_tetR (Supplementary Fig. S1A). The plasmid was synthesized by In-Fusion reaction, integrat-

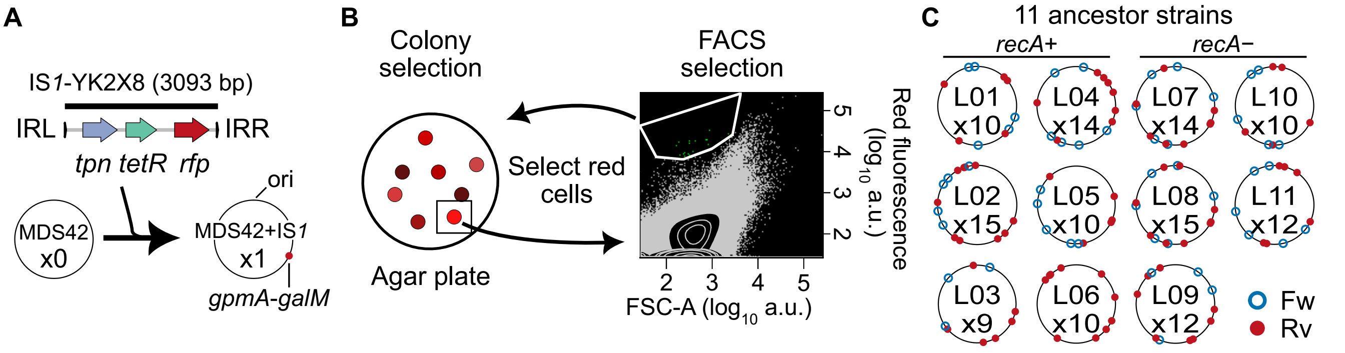


Figure 1. The construction of ancestor strains with multiple copies of high-activity ISs. (**A**) To achieve rapid IS-mediated evolution, we constructed *E. coli* strains with high IS activity. First, a copy of an inducible IS*1*-YK2X8 was introduced into the *gpmA-galM* locus of IS-less *E. coli* MDS42 by lambda red recombination. *tpn*: transposase gene, *tetR*: tet repressor gene, *rfp*: red fluorescent protein gene, IR: (L: left; R: right) inverted repeat. (**B**) Then, IS was accumulated in the genomes by selecting cells with brighter red fluorescence by FACS and by picking brightest colonies under blue/green LED. For FACS, bright cells within manually drawn gates (polygon) were collected. "FSC-A" indicates forward scatter value, used as a proxy for cell size. (**C**) The positions and numbers of IS copies in genomes of the 11 ancestor strains used for the subsequent MA experiment. The numbers after "x" indicate the copy numbers. The circular genomes are drawn to scale with the origin of replication (ori) at the top. Fw and Rv: ISs located on the forward and reverse strands, respectively.

ing the *tetR* cassette from pAJM.011 [20] into pKD46 [18]. pKD46_tetR was used instead of the typical pKD46 [18] to enhance the efficiency of lambda red recombination by repressing leaky IS*1* transposase expression, which may kill cells with IS*1* integrated. The recombination was performed as follows. A *gpmA*-IS-*galM* cassette with ∼500 bp overhang homologous to the *gpmA* and *galM* genes of MDS42, was prepared on plasmid pYK-2X8 (Supplementary Fig. S1B). The homologous sequences are adjacent in the MDS42 genome, and no sequence is deleted upon integration. The polymerase chain reaction (PCR) primers to amplify the *gpmA*-IS-*galM* cassette can also amplify a shorter sequence between the *gpmA* and *galM* genes from the genome (Supplementary Table S1). To avoid the amplification of genomic *gpmA* and *galM* that do not contain the IS, the plasmid was cured of residual genome DNA pre-PCR by exonuclease V treatment that deletes linear DNA for 30 min, following the manufacturer's protocol. Approximately 1 µg of PCR product was transformed into 80 µl of electrocompetent cells prepared from log-phase MDS42 cells cultured in Super Optimal Broth (SOB) with 0.1% (w/v) arabinose. After post-culture at 37 °C in SOB with catabolite repression (SOC) medium, cells were plated on LB agar plates with 12.5 µg/ml chloramphenicol at 42 °C to cure pKD46_tetR. IS insertion was confirmed by PCR using primers outside the homologous sequences, and the loss of the plasmid was confirmed by the loss of ampicillin resistance.

IS*1* has higher activities at lower temperatures [26]. Therefore, in the subsequent experiments, cells were cultured at 32 °C when inducing IS activity and at 37 °C otherwise to avoid unintended IS activity.

IS accumulation using fluorescence as a proxy for IS copy number

To further facilitate IS-mediated evolution, we prepared IS-accumulated strains of *recA*$^+$ and *recA*$^-$*E. coli*. *recA*$^+$ strains were prepared through cycles of fluorescence-activated cell sorting (FACS)-based selection and colony-based selections, using fluorescence as a proxy for IS copy number (Fig. 1B).

First, we selected cells by FACS. Four colonies obtained through the genome integration described in the previous section were transferred to LB medium with 12.5 µg/ml chloramphenicol and 100 nM aTc, and incubated at 32 °C with continuous shaking for one overnight. The antibiotic chloramphenicol was added to the media to avoid contamination of cultures by other bacteria. Approximately 100 cells exhibiting the brightest ∼0.01% fluorescence were isolated using FACS into 200 µl of phosphate-buffered saline (PBS) and cultivated on LB agar plates at 37 °C for two overnights. Twelve bright colonies under a blue/green LED transilluminator were picked and cultured in a rich defined medium based on standard M9 (Supplementary Table S2) and were considered independent evolutionary lines for the subsequent steps of IS accumulation.

Cells were further subject to cycles of FACS-based and colony-based selection. Colonies were diluted in the modified M9 medium with varying concentrations of aTc (0, 1, 10, and 100 nM) and incubated at 32 °C for two overnights. The cell cultures with the strongest induction without growth defects were diluted into PBS, and the 0.01% brightest cells were collected by FACS into 200 µl of PBS. The cell mixes were spread onto LB agar plates and incubated at 37 °C to obtain isolated colonies. The three brightest colonies were identified using a custom ImageJ script on photographs taken under a blue/green LED transilluminator, mixed into the modified M9 medium, and used for subsequent rounds of selections. A total of four rounds of FACS-based IS accumulation were performed.

recA$^-$ strains were prepared similarly. *recA* was deleted before IS accumulation from MDS42 by integrating the *kanR* cassette, which was amplified from the *recA*$^-$ strain of the Keio collection [17], using lambda red recombination [18]. The *kanR* cassette was then removed by FLP recombinase expressed from pCP20 [19]. A copy of IS was introduced into the *recA*$^-$ strain by lambda red recombination with pKD46_tetR. IS accumulation was performed with nine rounds of FACS-based IS accumulation.

The genotype of the cells was identified by long-read sequencing as described below but with a low read depth just to identify the presence of ISs. Among the multiple strains obtained, we chose the strains that seemed to have accumulated ISs based on the sequencing results. We excluded strains that ended up amplifying ISs through tandem duplication, possibly allowing the strains to increase *rfp* while avoiding transposition-induced growth defects (example in Supplementary Fig. S2).

Mutation accumulation experiment

The IS-accumulated strains were subjected to MA experiments in a nutrient-rich medium with aTc induction. For each passage, fresh LB agar plates were prepared by spreading aTc mixed in 100 µl of sterile water on pre-prepared LB agar plates containing 12.5 µg/ml chloramphenicol, resulting in a final aTc concentration of 3 nM. We employed a low induction concentration to avoid growth defects due to high IS activity. Chloramphenicol was added to avoid contamination of other bacteria.

Single visible colonies were randomly selected for each lineage and streaked onto new plates to obtain isolated colonies, as previously described [27]. Plates were incubated at 32 °C for 3–4 days with shading to prevent aTc degradation (two passages per week). To avoid selection for faster or slower growth, we randomly picked colonies with diameters between 2 and 3 mm. When unavailable, colonies were selected randomly.

The numbers of generations were inferred using a previously established regression between viable cell count and colony size [27]. For colonies within the target size range, we assumed a 2.5 mm diameter (27.2 generations). For colonies with atypical sizes, generations were calculated based on measured diameters.

Genome sequencing

To identify the genome sequences of the evolved strains, we employed long-read sequencing. Genomes were sequenced at three points: before MA, after eight passages, and after twenty passages. Raw reads were prepared as follows. During the MA, glycerol stocks were prepared every four passages from the same colonies used for the next passage. These stocks were created by incubating cells in LB medium with chloramphenicol and storing at −80 °C. After the MA, these stocks were restreaked on agar plates, single colonies were picked, and cultured overnight in LB medium with chloramphenicol at 37 °C. We did not add aTc to these post-MA cultures to avoid IS activity. Genome DNA was extracted from the cells using the standard phenol–chloroform method using the reagents of the Promega Wizard HMW DNA extraction kit and Phenol/Chloroform/Isoamyl alcohol (25:24:1). The extracted DNA was read using the Rapid Barcoding Kit 96 and sequenced using R.9.4.1 minion and flongle flow cells or using Native Barcoding Kit 24 V1 with R10.4.1 minion flow cells. We used v5.1.0–v5.7.5 of the MinKNOW software to run the experiments. Fastq files containing the DNA sequences were generated by ONT Guppy (v6.1.5–v6.5.7) super-accurate mode.

The reads containing sequences from individual DNA molecules were assembled to draft genomes as follows. Reads <1000 bp were filtered out using Filtlong (v0.2.0), and the remaining reads were assembled with Flye (v2.9) [28]. The median N50 of post-filter reads was 22.3 kbp, and the median read depth was × 30.3 (Supplementary Table S3). Some genomes failed to assemble automatically into a single circular contig using Flye due to the presence of repeats >100 kbp (e.g., L02-4 passage 20, Fig. 3A). Such genomes were manually resolved by visualizing the assembly graph using Bandage (v0.9.0) based on the network of contigs and read depths [29]. When resolving these repeats, we chose the organization resulting in the most parsimonious number of rearrangement events. Due to this limitation, inversions between the repeats could have been missed, and we might have underestimated their numbers.

The draft genomes were refined before the analysis of structural variations (SVs). Assembled draft genomes were polished using medaka (v1.6.0) as follows. Fastq reads were mapped to the draft genome using minimap2 (v2.24) [30]. Variants were called using medaka consensus and medaka variant. Polished sequences were generated using bcftools (v1.15) [31] `consensus` command, applying the called variants in VCF format to the draft genome.

To assess the quality of the polished genomes, we mapped assembly reads back to polished genomes using minimap2. We used cutadapt (v4.0) [32] to remove potentially chimeric reads before mapping. To identify potential misassemblies, we calculated the ratio between hard or soft clipped reads (clipping length >100 bp) and average read depth for each 1 kbp window. A missed tandem duplication would ideally produce a ratio of 1/1.5 (∼0.67). Imagine a genome with a tandem duplication >1 kbp with exactly 1× coverage and the duplication is missed in the assembly. In the window that includes the left end of the duplication, we expect 1× coverage of reads without clipping and an additional 1× coverage of reads only in the right side of the end. If we further assume that the end is at the center of the window, the average coverage would be 1.5× with one clipped read. Thus, we focused on loci with ratios >0.6, manually verified read alignments in IGV [33], and corrected assemblies when possible (Supplementary Table S3).

There were some genomes with high clipping ratio even after the manual curation. The ancestor genomes exhibited high ratios where we manually introduced SVs that were commonly found in multiple evolved genomes (see "Identification of IS insertion sites" section, Supplementary Table S4). Some genomes showed high ratios near IS copies inserted by lambda red recombination because sequences adjacent to the initial IS inadvertently matched the sequencing kit adapter (RBK110.96), causing supporting reads to be filtered by cutadapt. These clippings do not indicate misassemblies, as without filtering, the clipping ratio were typically within the normal range. Finally, reads of lines L01-4 (passage 20), L05-2 (passage 20), L05-4 (passage 8), and L06-1 (passage 8) did not support a single genome structure. We manually fixed these genomes so that the genome structure minimizes the number of sites with high clipping ratios while being consistent with other genomes of the same line.

To assess the validity of identified ISs we also checked that all ISs in genomes of passages 8 and 20 had at least one supporting read that start aligning from at least 100 bp upstream to 100 bp downstream of the IS, and within the sequence starting from 20 bp upstream to 20 bp downstream of the IS, the alignments do not have any gap, indel, or clipping >10 bp. As examples of read alignments on ISs, reads covering IS variants in Fig. 5B are shown in Supplementary Fig. S4.

Note that we did not polish the genomes by short-read sequences, and the analyzed genomes contain false point mutations and small indels. Nevertheless, the impact of these errors on SV analysis appears negligible, as evidenced by the general consistency of results among descendants of the same ancestor and the contiguity of the assembled genomes excluding those with repeats over 100 kbp.

Identification of IS insertion sites

To detect ISs even if they have altered structures as in Fig. 5, we adopted the following procedure. Sequences significantly (e-value <1e − 10) matching the IS were identified by blastn [34] (v2.14.0+), using the sequence of IS*1*-YK2X8 as the query. The blastn hits of candidate IS sequences were clustered, permitting a gap of up to 20 bp. Hits in the same cluster were assumed to be parts of the same IS. Among the IS candidate sequence clusters, those that contained at least 300 bp of continuous IS matches were classified as ISs. The direction of IS was assigned based on the largest match of IS in the cluster.

We then analyzed the "IS insertion site," the position of IS insertion based on the coordinates of a reference predecessor genome (Supplementary Fig. S5A). We analyzed such sites to connect ISs detected in descendant genomes to those in previous sequencing rounds and identify those that are newly inserted, even under the presence of structural mutations. For every IS detected in the descendant genome (e.g., post-twentieth passage), we identified the IS insertion sites in the genome of the predecessor (e.g., post-eighth passage), the ancestor (pre-MA), and MDS42 (the common ancestor). Sequences of ISs, including the 100 bp sequences flanking the IS in the genome of the descendants, were searched in predecessor genomes by blastn. We used the best hit in terms of e-value to avoid over-counting the insertion events when ISs were inserted into duplicated regions. This will generally give two loci for each IS, one for each end, because IS*1* typically forms 8 or 9 bp target site duplications upon insertion [3]. If there were duplications or inversions between ISs, the loci of the two ends would not be found near each other. To identify unique insertion sites in the coordinates of the predecessor genomes, we clustered the blastn hits of all ends of ISs in both the predecessor and descendant genomes, permitting a gap of up to 20 bp by agglomerative clustering (single linkage, scikit-learn [35], v0.24.2), and chose the center of the cluster as the IS insertion site. We allowed up to 20 bp gaps to allow loci separated by the target site duplications to be clustered together (Supplementary Fig. S5A).

We quantified IS insertion events by identifying unique IS insertion sites in the reference genome that were exclusive to descendant genomes. Specifically, IS insertion sites were categorized based on the presence of the sites across generations: "original" (present in both ancestor and descendant genomes), "lost" (present only in the ancestor), and "new" (present only in the descendant) (Supplementary Fig. S5B). This procedure prevents overestimation of insertion events when, for example, an IS undergoes tandem duplication, as the duplicated copies would share identical insertion sites in the predecessor genome (Supplementary Fig. S6C).

We recognize potential insertion overcounting when new IS elements insert into previously duplicated loci; blastn hits of flanking sequences may produce multiple identical e-value hits, potentially causing misassignment of insertion sites. Manual inspection revealed only one such case in L04-3 (passage 20) where an IS at 2.16 Mbp inserted into a locus duplicated in passage 8. However, insertion sites were correctly identified at the same predecessor genome locus at 2.32 Mbp, with no resulting overcounting.

The detection of IS insertion sites revealed some errors in the assembled sequences, which were manually corrected if necessary to avoid over-counting IS-related events. Several genomes originating from the same parental strain exhibited identical structural mutations related to ISs, perhaps due to either heterogeneity within the parental strain or mutation events between sequencing the ancestor strain and the start of MA. To prevent the over-representation of IS-related events, we manually incorporated those identical mutations into the pre-MA genomes (Supplementary Table S4). Besides, the post-MA genomes of L07 and L08 had identical IS insertion sites, likely due to cross-contamination before MA. Also, the genome of L04-1 after the 20 passages was highly diverged from the genome at the eighth passage. While the genomes of the two time-points were similar compared to genomes from other lines, many ISs inserted in the eighth passage were lost in the twentieth. Hence, we excluded the comparison between the eighth and twentieth passages of L04-1 from the analysis of IS insertion sites. Thus, the per-line event counts of SVs shown in Table 1 likely show slightly conservative estimates.

Visualization of structural variations

Genome structure changes were visualized using SyRI (v1.6.3) [36] and plotsr (v1.1.1) [37] (Fig. 3A and Supplementary Figs S8–S11). To circumvent the difficulty of identifying SVs in the presence of repetitive ISs, we removed the ISs from the genome sequences before running SyRI. Genomes of consecutive rounds of sequencing were aligned by minimap2. Since we did not polish the assemblies by short-read sequences, the preset `asm20` was used to accept mismatches due to small errors in the assembled genomes. Further options `-H -f 100 -r1k,10k -rmq=no` was used for compatibility with SyRI. SyRI was run using the output bam file with options `–allow-offset 20 –cigar –no-qc`. We considered that an SV had a corresponding IS if at least one of its ends was within 20 bp of an IS. The identified coordinates of the boundaries of SVs were transformed back to the original sequences containing IS using a custom script. We used the SV annotation by SyRI for inversions, but for deletions and duplications, we used the copy number changes, as explained in "Detection of deletions and duplications" section.

Rearrangements of at least 100 bp were visualized using plotsr. We preprocessed the output of SyRI to make it compatible with plotsr using a custom script as follows. We reassigned the genomes in the configuration file for plotsr into chromosomes and the generations to genomes to visualize the rearrangements in one plot. Blastn matches of ISs >300 bp were exported as BED files with a format compatible with plotsr. As plotsr did not display rearrangements within a synteny block, we used a custom script to reveal the nested rearrangements inside synteny blocks. Also, we reclassified CPG (Copy Gain) or CPL (Copy Loss) events into duplications or deletion, respectively, as they were also missed by plotsr. Some mutations were classified as highly diversified regions (HDRs) by SyRI and were excluded from the visualization. This includes the complex rearrangement depicted in Fig. 3D, as can be seen from the lack of bands connecting the regions in Supplementary Fig. S10.

Analysis by SyRI and plotsr revealed some errors in the assembled genomes. Among regions identified as HDRs, we manually fixed the regions that were contaminated by

Table 1. Events during the ten weeks of MA

	This study							LTEE[c]		Wild-type MA[e]	
Event	*recA*$^+$	/line	*recA*$^-$	/line	Total	/line	(IS)	Total	(IS)	Total	(IS)
IS increase	380	15.8	135	6.8	515	11.7	–	150.5	–	–	–
Insertion[a]	644	26.8	258	12.9	902	20.5	–	240[d]	–	758	–
(Simple Ins.)[a]	260	10.8	106	5.3	366	8.3	–	–	–	–	–
Deletion[a]	254	10.6	109	5.5	363	8.3	358	82	57	146	98
Duplication[a]	30	1.3	10	0.50	40	0.91	40	9	4	0	0
(Transposition)[a]	20	0.83	5	0.25	25	0.57	25	0	0	0	0
Inversion[ab]	42	1.8	12	0.60	54	1.2	54	19	15	–[f]	–[f]

Columns with (IS) indicate IS-mediated events among the total. Note that the number of insertions can exceed the net increase in IS count. For example, when an IS inserts and subsequently triggers a deletion that includes a pre-existing IS, the total IS count remains unchanged despite the new insertion event. a,b: Omitted data from L04-1 Passage 20 and reversions of inversions, respectively. c: Data from 12 lines after 40000 generations [4, 14]. d: "IS-related" mutations in a previous study [14]. e: Data from 520 independent MA lines of wild-type *E. coli* and its 12 repair-deficient mutants (total ∼2.2 million generations) [13]. f: Not detectable using short-read sequencing [13].

low-complexity repeats (Supplementary Table S3). We reran the above analyses if any modification was made to the genomes.

Analysis of the distribution of IS insertion sites

Hotspots of IS insertions were detected using a sliding window approach. We used 100 kbp-sized windows with a step size of 1 kbp. The significance of the IS insertion frequency in each region was determined using binomial tests.

We discriminated simple insertions from complex insertions entailing recombination, such as ISs that underwent intrachromosomal transposition via the cointegration pathway [38]. Simple insertions were identified in two steps (Supplementary Fig. S6A and B). First, for each IS in the descendant genome, we identified whether the blastn match of the sequence flanking the IS to the predecessor genome supports simple insertion. Specifically, this was judged by (i) 100 bp sequences flanking the IS have matches over 80 bp; (ii) the position where ISs are inserted based on the blastn are clustered into the same IS insertion site as described above; (iii) no significant match to the predecessor IS was found in the site (otherwise, it should be a part of a predecessor IS); and (iv) the strands the flanks match are consistent with a simple insertion event. Then, if at least one IS among the descendant ISs supporting a new IS insertion site satisfies these criteria, the new insertion is classified as a simple insertion. This criterion was also applied to identify eight simple insertions of composite transposons. We did not include them in the simple insertions in the hotspot analyses, but included them when analyzing the "local hopping" of ISs to increase the number of simple insertion events per line. Examples of simple insertions and complex insertions are shown in Supplementary Figs S6B–D and S7.

Following a previous analysis of MA of wild-type *E. coli* [13], we analyzed the distance distribution of newly inserted ISs relative to pre-existing ones. This involved comparing the distribution of distances from the newly inserted ISs to their nearest pre-existing ISs with randomly generated distance distributions. The random distributions were generated from 10 000 simulated insertion events per genome in pre-MA and post-eight passage genomes. The expected probability of simple insertions within 10 kbp of pre-existing ISs was calculated by taking the median of 11 medians, each derived from a simulated distribution based on IS positions in the pre-MA genomes of the 11 parent strains.

Detection of deletions and duplications

To identify deletions and duplications, we analyzed the copy number changes. First, we identified the copy number changes based on the MDS42 genome sequence (Supplementary Fig. S12). Genome sequences were mapped against the sequence of MDS42 but with the IS inserted between *gpmA* and *galM*, simulating the post-lambda red recombination state. We used minimap2 to align the genome of interest to the reference MDS42 genome. The per-base-pair depths were computed with the `depth` command of pysam [31] (v0.21.0). Small variations in copy numbers, likely due to sequencing errors, were omitted by smoothing the depths. First, bases were grouped together with neighboring bases with identical depths. To get a conservative estimate of the number of copy number changes, we merged groups <20 bp to their closest neighboring groups on the 5′ side that were≥20 bp.

To identify the numbers and size distribution of deletions and duplications, the same analysis was performed but by comparing the genome sequences from two consecutive genome sequencing rounds. We assumed that duplications occurred with the above-identified groups as units. The number of duplication events to achieve the observed copy number changes was assumed to be the copy number minus one. These assumptions can potentially lead to overcounting of duplications if nested or adjacent regions were duplicated, but no group seemed to be affected, checking by manual inspection. To avoid the distribution of duplications being distorted by tandem duplications with more than two copies, each duplicated group is counted only once in Fig. 3C regardless of the depth. In contrast, for counting the number of duplications as in Table 1, we counted the numbers including multiple duplications of the same region. Composite transpositions were identified by manual inspection of the duplicated regions.

Identification of essential genes

To compare the genome size changes of our study with those of the LTEE [39], we adjusted the genome sizes by the essential gene contents. For the essential genes of REL606, the ancestor strain of the LTEE, ideally, we would have used the list of genes essential in DM25 medium used in the LTEE. However, the criteria of essentiality in available datasets for REL606 [40] were not comparable with the criteria used for other strains [17, 41]. Thus, we considered genes essential in the LB medium we used for our MA experiment and assumed that essential genes in the standard K12 strain MG1655 are essential in both MDS42 and REL606. All essential genes were found in MDS42 by matching the bnumbers and names. For REL606, all essential genes were found by matching the bnumber or the gene names, or by manually finding the remaining orthologs by blastn.

Essential genes of our evolved strains were identified by mapping the genes of MDS42 to the evolved genomes using blastn with the default setting. We assumed the match was valid if the e-value was <1e–10 and the match spanned at least 95% of the gene or the match had at most 5 bp not matched. If multiple copies of an essential gene were detected in a genome, we assumed that the gene was not essential, as they are redundant and deleting or disrupting the gene would not be lethal. We assumed that these essential genes were the only essential sequences in the genomes to calculate the total length of essential sequences in each genome.

Estimation of the expected length of deletions

We estimated the expected median length of deletions based on the positions of essential genes and ISs in the genomes. Our analysis assumed that: (i) Deletions start at the ends of IS elements. (ii) Deletions can extend up to, but not include, the nearest essential gene. (iii) Non-IS ends of deletions are uniformly distributed across nonessential regions (adopted as a null model). (iv) The growth rate and deletion rate is independent of the distribution of ISs and essential genes in the genome.

For a integer-valued random variable l, its median l^* satisfies the following equation:

$$0.5 = \sum_{l=1}^{l^*} p(l),$$

given the probability density function $p(l)$.

Let us assume that we had only one genome j and calculate the density function $p(l;\, j)$, the probability that a deletion extends l bp from an IS end in genome j. The genome has IS ends labelled with i ($i = 1, 2, \ldots, n_j$). The distance to the nearest essential gene from the i-th end of the j-th genome is denoted as d_{ij}. Due to assumptions (i)–(iii), $p(l;\, j)$ is proportional to the number of IS ends that are at least l bp apart from the nearest essential gene $\sum_i [d_{ij} \geq l]$, where $[d_{ij} \geq l]$ indicates that if $d_{ij} \geq l$ then the value is 1, otherwise 0. This gives

$$p(l;\, j) = \frac{\sum_i [d_{ij} \geq l]}{\sum_{k=20} \sum_i [d_{ij} \geq k]} = \frac{\sum_i [d_{ij} \geq l]}{\sum_i (d_{ij} - 20)}.$$

The sum is taken from 20 bp, as we ignored deletions <20 bp in our analysis. From this probability density function, the median expected length of deletions in the genome l_j^* satisfies the following equation:

$$0.5 = \sum_{l=20}^{l_j^*} p(l;\, j).$$

To calculate the expected length through all genomes, we weighted the IS ends by the number of passages N_j between the genome and the next sequencing round based on assumption (iv) ($p(l,\, j) \propto N_j p(l;\, j)$). Now that we have the probability density function $p(l, j)$, the expected median length of deletions l^* can be calculated by solving the following equation:

$$0.5 = \sum_{l=20}^{l^*} \sum_j p(l,\, j) = \sum_{l=20}^{l^*} \frac{\sum_j N_j \frac{\sum_i [d_{ij} \geq l]}{\sum_i (d_{ij} - 20)}}{\sum_j N_j}.$$

When the exact integer l^* was not found, we interpolated the value of l^* by linear interpolation between the two closest integers. Bootstrap confidence intervals were calculated by resampling the set of IS ends.

Notes on IS loss

We identified 22 cases of potential IS loss events by comparing the genomes of two consecutive rounds of sequencing. We did a mapping, basically the reverse of the analysis of IS insertion sites. A total of 3000 bp sequences flanking each side of a copy of IS in the predecessor genome (e.g., post-eighth passage) were searched in the descendant genome (e.g., post-twentieth passage) by blastn. The hits with the lowest e-value were considered for each side of the IS. An IS in the predecessor genome was annotated as lost if the matched end positions of the flanks had a gap <100 bp (much smaller than ISs) and the two matches were on the same strand.

Among the ISs that were annotated as lost, we excluded ISs with some reads supporting the nonexistence of the IS in the predecessor genomes by manually checking the read alignments on IGV (Supplementary Fig. S3). This identified three cases of heterogeneity in the existence of ISs (L01-3.2.10, L02-4.2.7, and L08-4.2.1, notation as in Fig. 5). We considered them as false positives in the detection of IS loss.

Note that the detection of IS loss by this criterion may overestimate IS loss if the predecessor genome contained a duplication where only one copy included an IS. If the IS-containing duplicated region was deleted while retaining the IS-free copy, this would be incorrectly annotated as IS loss, when it should be considered as a deletion. Manual inspection of the genomes revealed no such cases.

We observed that among the 17 ISs that were considered lost between the eighth and twentieth passages, 15 were newly inserted during the first eight passages. The proportion is significantly higher than expected from the fraction of ISs in the genomes of the eighth passage that were newly inserted by simple insertions (one-sided binomial test, $P = 4.7 \times 10^{-12}$). While this might be because ISs inserted during MA are more likely to be lost than those inserted during the FACS-based selection, we believe that these "losses" are indicative of ISs being inserted after the MA when we prepared the samples for sequencing.

While the proportion of these insertions is small compared to the total number of insertions, and the majority of the mutations were consistent between two rounds of sequencing, one should note that some mutations could apparently be "reverted" for the same reason.

Detection of IS variants

The representative IS variants shown in Fig. 5 were identified by classifying all detected ISs according to size using 100 bp bins by hierarchical clustering connecting farthest neighbors (complete linkage and scikit-learn). The first IS detected with the most common length within each bin was chosen as representative of IS variants of that size range. Genes within ISs were annotated by blastn against the IS sequences with options `-evalue 1e-5 -max_target_seqs 100000`.

Statistical analyses

Paired *t*-tests, Wilcoxon rank-sum test, and correlations were calculated using R stats (v4.3.1). Analysis of Variance (ANOVA) and χ^2 tests was performed using the `ANOVA` function from the R car package (v3.1.2). Linear and generalized linear models (GLMs) were performed using the `lm` and `glm` functions of R stats, respectively. We did not include the interaction terms in the models. Most models included intercepts, except for two analyses: the correlation between the number of generations and the number of IS insertion sites; and the correlation between IS insertion count and genome size changes. These intercepts were excluded because there should be no IS insertions at zero generations, and zero IS insertions would imply no evolution.

To show the extent of IS insertions in our study, we compared the number of IS insertions in the *recA*$^+$ lines and the MA experiment performed by Foster's group [42]. In their study, it required 111 passages to reach 3080 generations [42], and we assumed that one passage was performed every weekday.

Binomial tests were performed using the `binom.test` function of R stats. Sliding window analyses were performed using a custom script using `binom.test`. Results were adjusted for multiple test comparisons using the Benjamini–Hochberg method via the `p.adjust` function in the R stats package. For assessing the significance of the correlation between the frequency of essential genes and insertions within regions, we performed the same analysis but against nonoverlapping windows of 100 kbp.

K-sample Anderson–Darling tests were used to compare the observed distance distribution against distances from 10 000 randomly generated insertion events using `KSampleADTest` of the HypothesisTests package (v0.11.0) in Julia. From the obtained *P*-values, Fisher's combined probability tests were performed using the `sumlog` function from the R metap package (v1.1).

Bootstrap confidence intervals were calculated using the `boot` function of boot (v1.3.28) in R and `bootstrap` function of scipy (v1.9.3) in Python with 1000 bootstrap samples using the bias-corrected and accelerated (BCa) method.

To analyze how essential genes affect insertion frequency, we compared genomes across consecutive sequencing rounds. We divided predecessor genomes into 10 kbp windows and analyzed whether the presence of essential genes (Ess) and IS elements in the windows (IS), and *recA* in the genomes (recA) affect the rate of IS insertions (INS) per nonessential non-IS loci per passage (GLM, negative binomial, and log link). The number of nonessential non-IS loci (*l*) and the number of passages between two genomes (*N*) were included as offsets, assuming that they have a proportional effects on the number of IS insertions. In other words, the model we used is expressed as follows:

$$\mathbf{log}(\frac{\mathbf{INS}}{lN}) \sim \beta_0 + \beta_1\mathbf{Ess} + \beta_2\mathbf{IS} + \beta_3\mathbf{recA},$$

where βs indicate the predicted coefficients.

Novel programs, software, and algorithms

The custom scripts used for the analysis is available on figshare (https://doi.org/10.6084/m9.figshare.27800133).

Results

The construction of ancestor strains

We designed a derivative of IS*1*, a major IS of *E. coli* (IS*1*-YK2X8, Fig. 1A). The IS was tailored to induce rapid IS-mediated structural rearrangements in *E. coli* as follows. A high-activity mutant of the transposase gene (*tpn*) [21] was placed under a strong inducible promoter (P_{Tet}). The promoter is repressed by the TetR repressor expressed from *tetR* inserted within the IS and is derepressed when aTc is supplemented to the growth medium. To facilitate the rapid observation of IS-mediated genome evolution, we introduced multiple copies of IS*1*-YK2X8. This was achieved by introducing a copy of *rfp* in the IS and using red fluorescence as a proxy of IS copy number. After introducing a copy of IS*1*-YK2X8 to an IS-less *E. coli* strain MDS42 [16], we repeatedly selected for brighter red fluorescence via FACS and picking colonies under blue/green LED (Fig. 1B).

We prepared six strains with *recA* and five without *recA* to observe the potential effect of genetic backgrounds on IS-mediated genome evolution (Fig. 1C). *recA* is the major enzyme in the recombinational repair system in bacteria [43]. The loss of related genes is a characteristic of genome reduction in endosymbionts [5], and its loss affects the rate of recombination between ISs [44]. Even so, the effect of *recA* loss on IS-mediated genome evolution was not investigated in Foster's group's MA of wild-type *E. coli*, despite running the experiment with knockouts of other repair pathways [13].

Expansion of ISs under relaxed selection

We evolved the strains under conditions simulating the relaxed selection considered to cause IS expansion in nature (Fig. 2A). When streaked on agar plates, *E. coli* forms colonies derived from single cells. Propagating a single colony every passage reduces the effective population size (N_e), allowing the accumulation of weakly deleterious mutations by genetic drift. Laboratory evolution using this method is called MA experiments [45]. We subjected the IS-accumulated strains to MA (reducing N_e), used the nutrient-rich Luria-Bertani medium (allowing more genes to be lost), and achieved high activity of IS by adding aTc to the medium. To systematically avoid selection based on growth rates, we randomly selected colonies within a predefined size range, avoiding both unusually large and small colonies (see "Mutation accumulation experiment" section). Following this procedure, 97% of selected colonies ranged between 1.5 and 3.0 mm in diameter. We evolved four evolutionary lines from every eleven IS-accumulated strains (44 lines). We denote the lines as L01-2: the second line derived from the first ancestor strain, L01. To capture the dynamics of genome evolution, we sequenced genomes at two time-points: post-eight and post-twenty passages.

After approximately a median of 541 generations (20 passages, 10 weeks, 95% CI assuming normality: 539–541), the numbers of ISs reached 18–35 copies in *recA*$^+$ strains and 13–29 copies in *recA*$^-$ strains. ISs increased by an average of 11.7 copies per genome (Fig. 2B and Table 1). The *recA*$^+$ strains showed an mean increase of 15.8 (9–23) copies, while the *recA*$^-$ strains showed an increase of 6.8 (0–17) copies. To analyze the difference due to the presence of *recA*, we fitted a GLM with IS copy number increase after 20 passages as the response variable, and the presence of *recA* and initial IS count as predictors (Poisson, log-link). The best fit

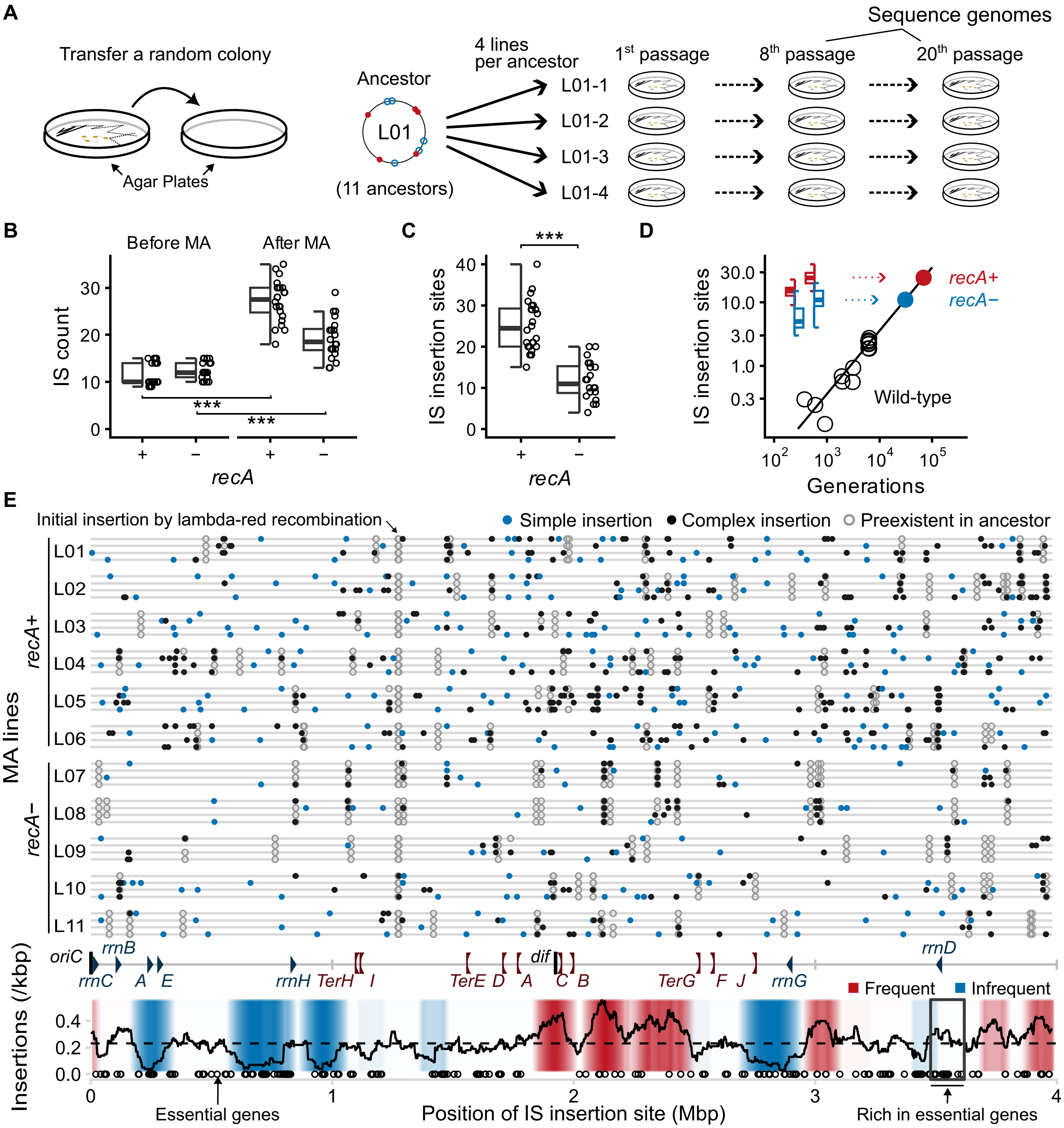


Figure 2. IS expansion after 10 weeks of neutral evolution. (**A**) Overview of the MA experiment. Single-cell bottlenecks were imposed on every passage to achieve relaxed neutral evolution. Genomes were sequenced after eight and twenty passages. (**B** and **C**) IS counts and new insertion sites after MA. $***: P < 2 \times 10^{-16} (\chi_1^2 = 161$, fold change $= 2.4$, 95% CI[2.1, 2.8]), $P = 7.6 \times 10^{-8} (\chi_1^2 = 29$, FC $= 1.5$, 95% CI[1.3, 1.8]), $P < 2 \times 10^{-16} (\chi_1^2 = 106$, FC $=$ 0.47, 95% CI[0.40, 0.54]) for (B) χ^2 tests on Poisson GLM (log-link) with the lines and before/after MA as predictors (*recA*$^+$: $n = 24$, *recA*$^-$: $n = 20$) and (C) χ^2 test on Poisson GLM (log-link) with *recA*+/- as predictors ($n = 44$), respectively. (**D**) Comparison of the number of new insertion sites with a previous MA of wild-type *E. coli* [13]. Linear regression (black line) was used to estimate the number of generations required for wild-type MA to reach insertion site counts observed in our study (filled colored circles). We used the median generations for each passage for the *x*-axis of the boxplots. Inter-line differences were at most 16 generations and negligible at this scale. Boxplots show the medians (center line), quartiles (box limits), and 1.5× interquartile ranges (whiskers). (**E**) Distribution of IS insertions after twenty passages of MA. Upper: Positions of IS insertion sites based on the coordinates of MDS42 + IS*1*. Each horizontal line corresponds to one of the 44 evolutionary lines. Closed circles indicate new insertions, where blue and black circles indicate simple and complex insertions (those associated with recombinations), respectively. Open circles indicate pre-existing ISs before MA. Lower: IS insertion frequency throughout the genome. The line graph represents the mean insertion frequencies within 100 kbp windows sampled every 1 kbp ($n = 3981$). Color depths indicate the number of windows with significant IS insertion accumulation or depletion. The box around 3.5 Mbp indicates the essential gene-rich IS-empty zone identified in the LTEE [14].

was 10.6 (×2.2 for *recA*⁺) × $0.96^{\text{(initial IS count)}}$ copy increase per genome (Supplementary Fig. S13). χ^2 tests indicated that the increase due to *recA*⁺ was significant ($\chi_1^2 = 66.8$, $P = 3.0 \times 10^{-16}$), whereas the pre-MA IS count was not ($\chi_1^2 = 3.1$, $P = 0.078$). Thus, RecA activity was the primary determinant of IS proliferation, independent of the initial abundance of IS.

To compare our observed copy number increase with Foster's group's MA of wild-type *E. coli* without assembled genomes [13], we analyzed the number of insertions based on new IS insertion sites: unique inserted positions in the predecessor genome coordinates (Supplementary Fig. S5). Comparing the genomes before MA, eight, and twenty passages, we detected 902 IS insertion sites (Table 1). To highlight the rapid IS expansion observed, we also compared the genomes after 20 passages directly with the ancestor genomes. Remarkably, a median of 26.8 sites per line was detected in *recA*⁺ (Fig. 2C), a number comparable to roughly 6.9×10^4 (95% CI [6.3×10^4, 7.6×10^4]) generations or 9.5 years of MA using wild-type *E. coli* [13] (Fig. 2D; linear regression, $n = 15$, $R^2 = 0.97$, 3.6×10^{-4} insertions per generation, 95% CI [3.2×10^{-4}, 3.9×10^{-4}]). However, IS copy-number increase per passage decreased with the number of pre-existing ISs (log-link GLM with Poisson distribution, FC = 0.93, 95% CI [0.91, 0.94], $\chi_1^2 = 71$, $P < 2 \times 10^{-16}$, Supplementary Fig. S13B). This decline could be attributed to two factors: the increasing fitness cost of ISs to the host and the higher copy number of *tetR*, which could have reduced transposase expression.

The distribution of IS insertions

The significant enrichment or depletion of IS insertions in specific genomic regions suggests either inherent bias in insertion site preference or selective pressures acting on genomes with ISs in these locations. Mapping the insertion sites to the MDS42 genome revealed a predominantly random global distribution of IS insertions (Fig. 2E), consistent with Foster's group's MA [13]. Taking 100 kbp sliding windows throughout the genome, some windows exhibited significant IS insertion frequencies ($\alpha < 0.05$ in two-sided binomial test with Benjamini–Hochberg (BH) correction for false discovery rate (FDR), Fig. 2E, lower). IS*1* can cause recombination upon insertion [38]. We classified an insertion as simple if the insertion sites due to the two ends of the IS in a descendant were within 20 bp of each other in the predecessor genome with no signs of complex rearrangements; we classified the rest as complex insertions (typically those entailing deletions or inversions, Supplementary Fig. S6). Focusing on the 366 simple insertions, no window showed significant bias (two-sided binomial test with BH correction for FDR, $P \geq 0.62$, Supplementary Fig. S14). However, this lack of significance may reflect reduced statistical power due to the smaller number of simple insertions, particularly for detecting coldspots (Supplementary Figs S15 and S16). An exception was the significantly frequent simple insertions in the terminal 1 Mbp region (115/366, two-sided binomial test, $P = 0.011$).

Following a previous analysis of Foster's MA [13], we next analyzed the distance distribution of IS insertions relative to pre-existing copies of ISs. This confirmed that new ISs were inserted significantly closer to pre-existing ISs than expected by random chance, or "local hopping" of ISs [13] (Supplementary Fig. S17, Fisher's combined probability test, $\chi_{174}^2 = 738$, $P = 1.5 \times 10^{-70}$). However, we noticed that excluding the complex insertions, the distribution of new IS insertion sites was not significantly different from the random distribution (Fisher's combined probability test, $\chi_{174}^2 = 192$, $P = 0.16$). The trend was the same when we reanalyzed Foster's MA data [13] (Supplementary Fig. S18). While the local hopping of ISs was evident for total insertions in IS*1*A and IS*5*A (Anderson-Darling (AD) test, IS*1*A: $P = 0.0035$, IS*5*A: $P = 2.1 \times 10^{-9}$), the distribution of simple insertions was indistinguishable with the random distribution (AD test, IS*1*A: $P = 0.65$, IS*5*A: $P = 0.056$). This suggests that the local hopping of ISs was due to complex insertions such as those involving deletions (Supplementary Fig. S6).

Comparing our study with the LTEE of *E. coli* with larger population sizes ($N_e = 3.3 \times 10^7$) [14] can clarify how selection shapes IS insertion distributions. In a previous study, an IS-empty zone rich in essential genes was identified in the LTEE [14] (3.480 to 3.615 Mbp in Fig. 2E, lower). To evaluate how selection affects IS frequency in this zone, we first analyzed insertion sites mapped to the original MDS42 genome and found no significant depletion (one-sided binomial test, $P = 0.94$). We then controlled for differences in the proportion of essential gene because the strains we used are derived from MDS42, which has fewer nonessential genes than the REL606 ancestor used in the LTEE. Using essential gene assignments from the Profiling of Escherichia coli Chromosome (PEC) database [41], we compared insertion frequency per nonessential sequence. The previously identified IS-empty zone showed significantly lower insertion frequency compared to the rest of the genome in the LTEE but not in our study (one-sided binomial test, LTEE: $P = 0.011$, This study: $P = 0.98$), supporting the role of purifying selection in the LTEE.

A GLM further supports that neighboring essential genes have a minor effect on IS target site preference (see "Statistical analyses" section); no significant depletion of IS insertions near essential genes was observed beyond the expected depletion due to lethal disruptions of essential genes. The presence of *recA* had significant positive effects on the total insertion rate per nonessential and non-IS locus, consistent with our previous analysis (fold-change = 2.1; $\chi_1^2 = 99$, $P < 2 \times 10^{-16}$). The presence of pre-existing ISs strongly increased the total insertion rate (fold-change = 5.2; $\chi_1^2 = 227$, $P < 2 \times 10^{-16}$), consistent with local hopping. The presence of essential genes had a small but significant negative effect on the total insertion rate (fold-change = 0.83, 95%CI [0.70, 0.97]; $\chi_1^2 = 5.7$, $P = 0.017$). However, when focusing on simple insertions and performing the same analysis, the presence of essential genes had negligible effects on the insertion rate (fold-change = 0.95, 95%CI [0.76, 1.2]; $\chi_1^2 = 0.18$, $P = 0.67$). This suggests that the slight negative effect on total insertion rate reflects the depletion of complex insertions near essential genes. This depletion likely results from the loss of cells where insertions triggered lethal deletions spanning essential genes, rather than from IS target site preference in simple insertions.

IS-driven genome structure evolution

Given the IS expansion, we next studied the changes in the genome structure (Fig. 3A and Table 1). We detected a total of 457 large-scale rearrangements, including 54 inversions, 363 deletions, and 40 duplications. Ninety-nine percent of the rearrangements were associated with ISs, signifying that we successfully observed IS-mediated genome evolution in the laboratory.

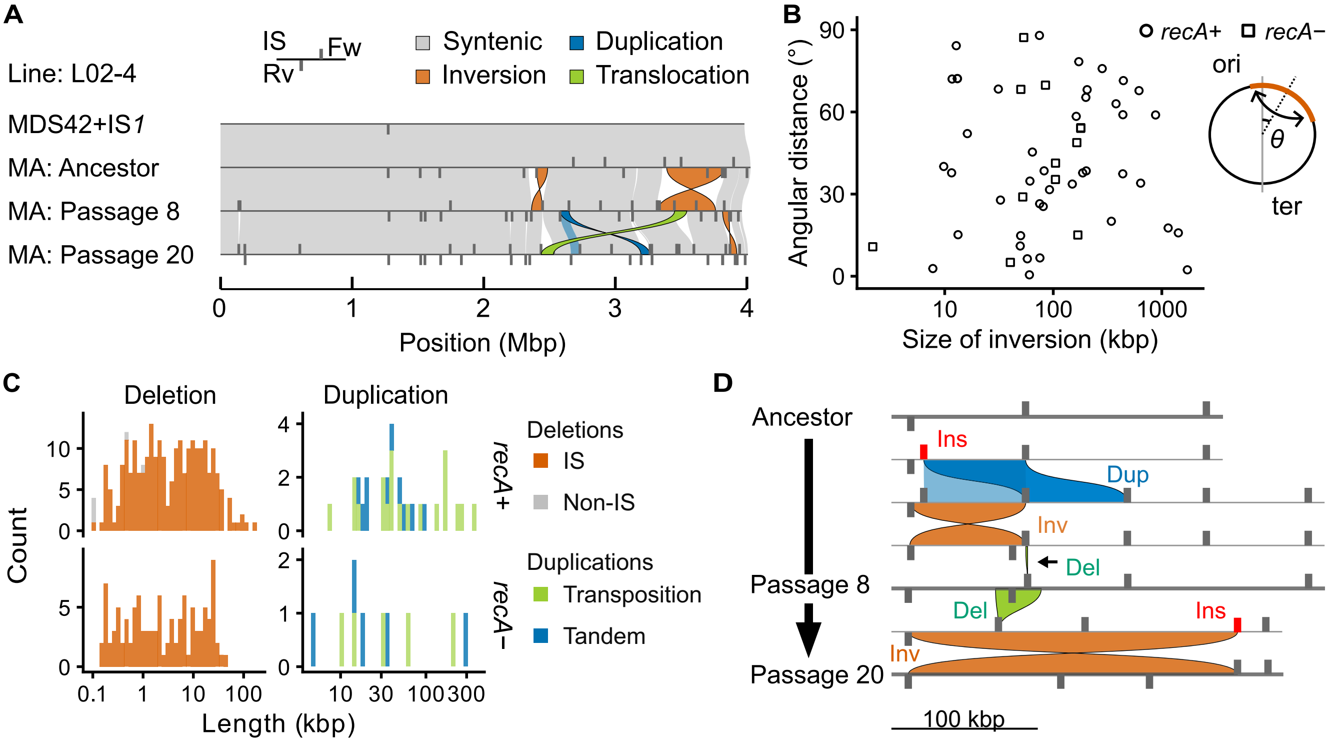


Figure 3. Rearrangements detected after 10 weeks of evolution. (**A**) Genome structure evolution of the IS*1*-integrated MDS42 strain, before and after 8 and 20 passages of MA (line L02-4). Genomes are aligned with the origin of replication set at zero. Gray bars represent the positions and orientations of ISs, with the top and bottom bars indicating the direction of the transposase gene. Homologous regions are connected by bands, with colors indicating the type of SV identified by SyRI [36]. Results for other lines are shown in Supplementary Figs S9–S11. (**B**) Inversion sizes and absolute angular distances of inversion centers from the replication origin (θ). θ = 0 indicates symmetry with respect to the ori-ter axis. (**C**) Length distribution of deletions and duplications. For deletions, red bars indicate those with at least one corresponding IS; gray bars indicate those without. For duplications, multiple copies of the same region are counted once. Note that for both deletions and duplications, the lengths do not include the ISs themselves. (**D**) Hypothetical route of a complex rearrangement observed in L05-4 during the MA. The depicted events represent one of the most parsimonious evolutionary routes inferred from the alignment of the three genomes. Note that the ISs are not drawn to scale.

The median size of inversions was 84 kbp (range: 2.2 kbp–1.7 Mbp, Fig. 3B). All three inversions exceeding 1 Mbp were nearly symmetric to the ori-ter axis, consistent with observations of large inversions in nature [46].

Deletions ranged from 102 bp to 163 kbp (Fig. 3C), with the largest deletion being a deletion of a duplicated region (L04-3, Supplementary Fig. S9). Despite the general randomness of IS insertions, the observed deletion sizes were much shorter than expected, reflecting a log-normal like long-tailed distribution (Supplementary Fig. S19). If instead, we assume a uniform distribution of deletions from pre-existing ISs to random loci that do not contain any essential genes in between as a null model, the expected median deletion size would be 30 kbp (bootstrap 95% confidence interval: 28–31 kbp), but the actual median deletion size was only 2516 bp. Nevertheless, the large number of deletions resulted in deletions spanning 1.3 Mbp in total, corresponding to 34% of the genome (Supplementary Fig. S12). As expected, no essential genes in the PEC and Keio datasets were deleted [17, 41]

Duplications had a median size of 40 kbp (range: 4830 bp–372 kbp, Fig. 3C). Surprisingly, the majority of the duplications resulted from transpositions of composite transposons with a copy of IS at each end (25/40), rather than tandem duplications (Table 1).

The interplay of IS-related mutations drove extensive genome rearrangements through insertions, inversions, deletions, and duplications. Line L05-4 exemplifies this rapid genomic change, displaying a complex combination of these mutations (Fig. 3D). Despite the complexity, our frequent sequencing ensured that unresolved complex rearrangements were rare and unlikely to have significantly affected the overall mutation statistics presented.

The rearrangements led to genome size changes ranging from 174 kbp reduction to 369 kbp increase or −4.4% to 9.2% (Fig. 4A). However, overall, deletions were offset by the genome size increase from duplications and IS insertions, resulting in negligible changes in the median genome sizes ($recA^+$: 8903 bp, 0.22%, $recA^-$: −4845 bp, −0.12%). To account for the low nonessential gene content of MDS42 compared to the ancestor of LTEE and the larger size of ISs used in our study, we adjusted the sizes of deletions based on the nonessential gene content of the parent strains used in the two studies (MA ancestors vs. REL606), and the sizes of IS insertions were reduced to that of the wild-type IS*1*. In addition, for the LTEE, we focused on the nonpoint-mutator strains, as the point-mutator strains had decelerated the pace of IS insertions [14]. We found that the two studies had a similar number of both median IS insertions ($recA^+$: 24.5, LTEE: 23.8) and median genome size reduction ($recA^+$: −0.96%, $n = 24$, bootstrap 95% confidence interval: (−1.7, 1.3); LTEE: −0.82%, n = 6, bootstrap 95% confidence interval: (−1.3, −0.035); Fig. 4B). While the genome size changes strongly correlated with the number of IS insertions in the LTEE (linear regression, $R^2 = 0.85$, $P = 0.0033$, $n = 6$), our study showed a large

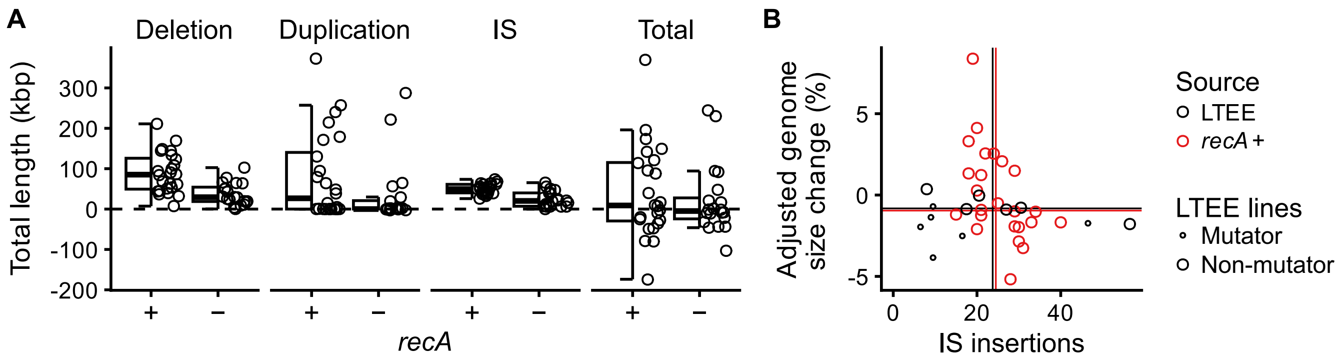


Figure 4. Genome size changes after 10 weeks of evolution. (**A**) Total length changes due to the structural mutations. Each data point signifies the overall length variation for a particular type of mutation in an evolutionary lineage. *IS* shows the total length of ISs, while *Total* shows the overall genome size changes. Boxplots show the medians (center line), quartiles (box limits), and 1.5× interquartile ranges (whiskers). (**B**) Genome size changes after adjustments to compare with the LTEE. Lines indicate medians for *recA*$^+$ strains in this study and nonmutator strains in the LTEE [39].

variation in genome size changes even among strains with similar numbers of IS insertions, obscuring the correlation (linear regression, $R^2 = 0.01$, $P = 0.62$, $n = 24$). The largest contributor to the variance in genome size changes in our study was the length of duplications, as indicated by ANOVA among the factors considered: deletions ($F_{1,20} = 8.4 \times 10^5$), duplications ($F_{1,20} = 3.9 \times 10^6$), and IS insertions ($F_{1,20} = 2.7 \times 10^3$) ($n = 24$).

The evolution of gene order within the IS

Unexpectedly, various structural variants of ISs were detected (Fig. 5A). While some of the IS variants were present in the ancestor genomes (L01, L02, L03, L09, and L10), new IS variants emerged in 10/20 lines with pre-existing IS variants and 12/24 lines without pre-existing IS variants. Various mechanisms formed the variants. For instance, L10-3.3.0 (underlined in Fig. 5A, IS named as in the figure caption) likely formed through a deletion of the sequence between two ISs: L10-3.2.0 and L10-3.2.1. Six IS variants formed through the insertion of one IS within another (Fig. 5B and Supplementary Fig. S20). While two variants (L02-1.1.12 and L10-3.1.8) were present in the ancestor genomes and thus may have resulted from selection to increase IS copy number, at least four IS variants formed during the MA. We tested whether this number coincided with the expected number based on the fraction of ISs within nonessential sequences. While this fraction increased during the MA (mean: from 1.1% to 2.1%), the observed number of IS-within-IS insertions matched the expected numbers using fractions in both the ancestral (2.7) and post-MA (5.4) genomes. This lack of self-avoidance aligns with previous findings: While some composite transposons exhibit "target immunity," where transposase avoids self-insertion, ISs typically lack this property [47]. The nested arrangement due to self-insertion brings homologous sequences into proximity, potentially leading to frequent homologous recombination within an IS, resulting in the further formation of variants (e.g., L10-2.2.20, dashed underlined in Fig. 5A). Furthermore, the transposition of subsequences from within the nested ISs led to the formation of additional variants (Fig. 5B).

ISs can function in pairs to transpose intervening genes as composite transposons. A classical example is Tn*9*, where a pair of IS*1* elements mobilizes the chloramphenicol resistance gene *cat* [48]. We repeatedly observed similar phenomenon with IS*1*-YK2X8 elements transposing intervening genes ("Transposition" in Fig. 3C). A notable example occurred in line L05-4 (Fig. 5C and Supplementary Fig. S10). Here, a pair of IS*1*-YK2X8 were inserted into sequences flanking *pqqL* and *yddB* by the eighth passage. The pair of IS containing a 7.5 kbp sequence within the *pqqL* operon was then transposed as an 13.7 kbp composite transposon into four different loci by the twentieth passage (Supplementary Fig. S21). Another example was found in line L04-4 (Supplementary Fig. S9). By the eighth passage, two copies of a 41 kbp sequence between *yohK* and *yehB* were found with a pair of ISs at their ends. By the twentieth passage, two additional copies of the sequence were found, resulting in a total of four copies in the genome (Supplementary Fig. S22). One copy was inserted by simple insertion, while the insertion of the other copy entailed an inversion between the newly inserted locus and the original locus.

Discussion

Bacterial genomes are dynamic and can undergo drastic changes. However, two key aspects of genome evolution remain elusive despite accumulating sequence data. First, intermediate evolutionary steps are often obscured, as we previously suggested for the IS-mediated evolution of operons [9]. Second, the specific conditions driving evolutionary phenomena are difficult to identify in complex and varying natural environments. Laboratory evolution allows researchers to disentangle such intricacies through controlled experiments [10] but has been limited by the typically slow pace of genome structure evolution [4, 13, 15].

To address these challenges, we developed an *E. coli* strain with high IS activity (Fig. 1) and demonstrated rapid IS-mediated genome evolution under relaxed selection in the laboratory, simulating the natural evolution of symbionts and pathogens. In just ten weeks, we observed numerous IS insertions, IS-mediated duplications, and deletions (Fig. 2, Fig. 3), contributing to at most −4.4% to 9.2% genome size changes (Fig. 4). This suggests that, given the right conditions, bacterial genomes can rapidly evolve their structure in a relatively short time, similar to pathogens that rapidly adapt to new hosts and antibiotics [49, 50].

Long-read sequencing enabled us to reveal numerous IS insertions and IS-mediated structural mutations (Table 1). IS variants and composite transposons were also observed

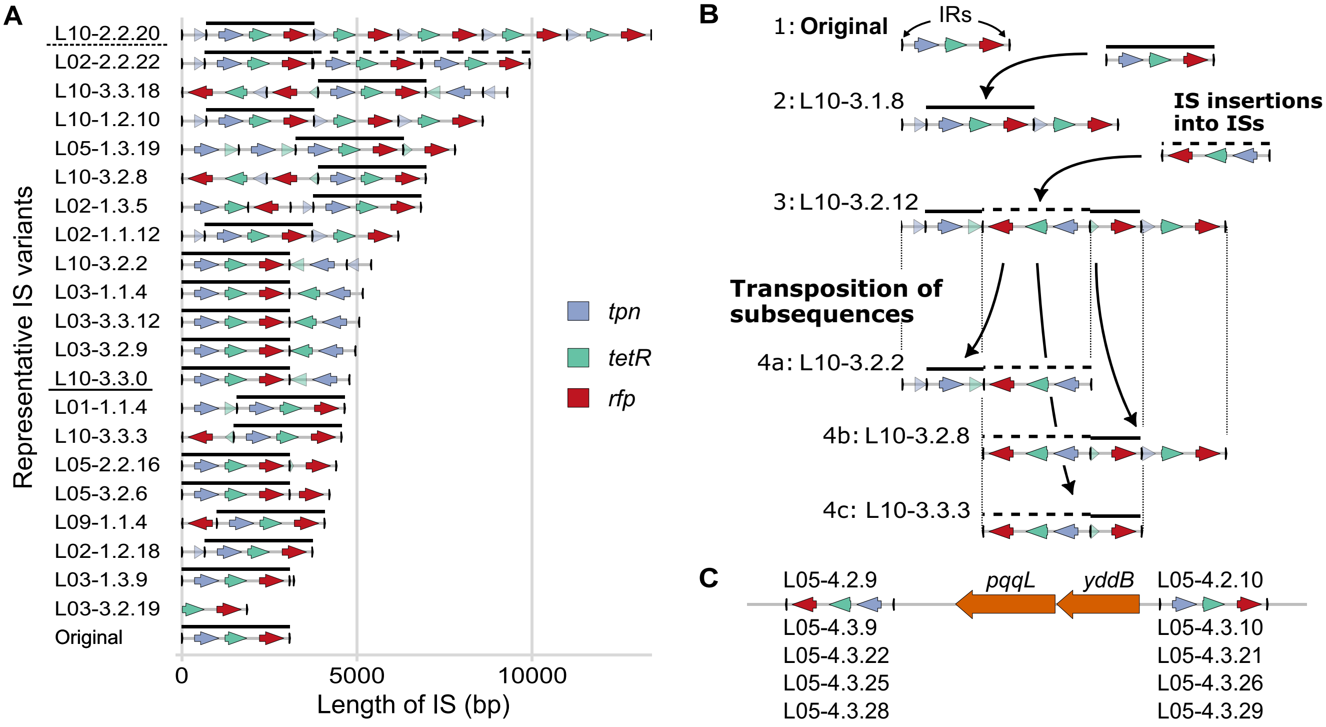


Figure 5. Representative IS variants detected after 10 weeks of evolution. (**A**) The representative variants. Each IS is uniquely labeled, e.g., L01-2.3.4 represents the fourth IS in the genome of the third DNA sequencing (first: before MA, second: passage 8, third: passage 20) of line L01-2. Full-length ISs are indicated by the lines above genes. Disrupted genes are shown in translucent colors. (**B**) Example of IS variant formation. Steps 1–4 represent a hypothetical scenario inferred from the structures of ISs. The lines above genes distinguish sequences derived from inserted ISs from those from the original IS. (**C**) Example of an active composite transposon that led to the partial duplication of the *pqqL* operon. The IDs indicate the ISs involved in the composite transposon.

(Fig. 5), showcasing our method's potential to study both the evolution of bacterial genomes and the underexplored evolution of ISs themselves [51].

By tracking evolution under controlled conditions with accelerated timescales, our approach allows the investigation of fundamental questions surrounding bacterial genome evolution. Our experiment fills the gap of the lack of experimental studies demonstrating extensive genome evolution under relaxed selection, and comparing our study with previous studies clarifies notions that previously were only vaguely suggested.

For instance, our experiment supports the notion that bacterial genomes are biased toward deletions [52–54]. In the laboratory, the LTEE showed a decrease in genome sizes [39], but it has remained unclear whether the decrease may reflect selective forces or mutation biases. Despite the effective population size of our study ($N_e \sim 10$) being much smaller than that of the LTEE ($N_e = 3.3 \times 10^7$) [55], the decrease in genome sizes in our study and the LTEE were remarkably similar (Fig. 4B). While some deletions in the LTEE, such as the parallel deletion found in the ribose operon (median 5881 bp) [4] were beneficial [56], this comparison suggests that genome reduction in the LTEE largely reflects genetic hitchhiking of deletions with beneficial mutations [4, 39].

In contrast to Foster's MA study [13] where duplications went undetected, our work reveals that duplications are major contributors to genome size variation (Fig. 4). This means that, like other mobile elements, IS can drive both genome reduction and expansion [52, 57]. Majority of duplications we observed occurred through composite transposition (Fig. 3C), suggesting that the limited IS copy number in wild-type *E. coli* may have constrained duplication events in previous studies. While wild-type strains contain only six copies of IS*1* separated by hundreds of kilobases [13], our ancestral strains carried over ten copies of IS*1*-YK2X8 (Fig. 1B), facilitating the composite transposition that requires proximal IS pairs [58]. While previous studies focused on easily detectable small deletions, which can lead to an impression that genomes without gene influx simply shrink due to the deletion bias [5], our results with duplications suggest a more nuanced picture. The observed interplay between frequent small deletions and rare large duplications suggests that genome size evolution can follow complex trajectories, potentially including transient expansions (Fig. 4B).

The four-fold higher deletion count compared to Foster's study [13] provides a clearer picture of deletion size distribution. Consistent with previous findings, we observed ISs inserted significantly closer to pre-existing ISs than expected (Supplementary Fig. S17), and deletion sizes significantly skewed toward shorter lengths than predicted by our null model of random insertions followed by nonlethal deletions between pre-existing ISs (Supplementary Fig. S19). However, our larger sample size revealed the longer tail of the deletion size distribution, not evident in the previous study [13] but consistent with findings in human cancer cell lines [59]. This skewed distribution contrasts with inversion sizes (Fig. 3B), which had a median of 84 kbp, similar to the median deletion sizes expected from the null model (30 kbp) and the median distance between random IS insertions and pre-existing ISs (104 kbp, median at empirical cumulative distribution function (ECDF) = 0.5 across 12 lines, Supplementary Fig. S17). We identified that local IS hopping was significant only when including complex insertions, mostly involving deletions (Table 1), in both our study (Supplementary Fig. S17) and in IS*1* and IS*5* in Foster's MA study [13] (Supplementary Fig. S18). Overall, we speculate from these observations a bimodal deletion size distribution, with peaks ∼1 kbp (adjacent IS deletions) and 10 kbp (deletions between ISs). Further studies are required to understand the mechanisms determining this distribution.

Some loci seem to have had a higher frequency of IS-related mutations than others, as we observed a locus with multiple structural mutations in a short period (Fig. 3D). We speculate that, in addition to the frequency of composite transpositions, the high ratio of deletions versus insertions in our study compared to wild-type (40% versus 19%, Table 1) may reflect the different IS positions in the ancestors. Strains with high IS activity or strains with recent IS expansions might better represent IS-mediated genome evolution in nature than wild-type *E. coli* often used.

In contrast, we found no preference of simple insertion sites at either 100 or 10 kbp windows, but two caveats should be considered. First, the lack of significant insertion frequencies in 100 kbp windows might be attributed to limited statistical power due to the small number of simple insertions. In particular, increasing the number of passages or lines for at least three-fold is required to reliably detect significant coldspots of simple insertions (Supplementary Fig. S16A). Second, we found that insertions near existing ISs tend to be complex. One interpretation could be that simple insertions near existing ISs are rapidly followed by deletions, converting simple insertions into complex events. Even if the initial positions of ISs were hotspots (which is likely given that we let ISs accumulate autonomously), subsequent simple insertions in these hotspots might have been converted into complex events, leading to an apparent lack of significant hotspots for simple insertions. Although our strategy of allowing ISs to insert throughout the genome autonomously effectively mimics natural IS expansion, future studies should also consider inserting ISs at random positions with a mechanism independent of IS to eliminate this bias.

Through comparing *recA*$^{\pm}$ strains, we unexpectedly observed reduced IS activity in *recA*-deficient strains (Fig. 2B). This effect could partially be due to synthetic lethality, which makes a deletion caused by IS insertion lethal only in the *recA*-deficient strains [60]. However, synthetic lethality alone is unlikely to explain the observed 2.2-fold reduction, as it would require approximately half of all genes to exhibit synthetic lethality with *recA* loss. Instead, we speculate that a more major factor is the *recA*-deficient strains' inability to repair double-strand breaks triggered by IS transposase [44] making cells with higher IS copy numbers less viable. Genome-reduced endosymbiotic bacteria often lack recombination genes [5]. While this is commonly thought to promote genome degradation [61], our findings suggest it might also lower IS activities, potentially preventing deleterious genome rearrangements.

Surprisingly, we observed multiple IS variants and composite transposons (Fig. 5). Most ISs, consistent with our findings, lack strong target site preferences [3]. However, some ISs exhibit nonrandom insertion patterns, preferring specific sequences or avoiding self-insertion [3]. The observation of the formation of IS variants through nested insertions (Fig. 5B) suggests that target preference for specific sites within an IS, if present, might facilitate the formation of IS variants, enhancing its evolvability, and should be further investigated in future studies.

The high IS activity allowed us to observe the emergence of composite transposons, key elements in spreading antibiotic resistance and pathogenicity [62]. In particular, sequences containing the *pqqL* operon and the locus between *yohK* and *yehB* were each found at four different loci by the twentieth passage. These sequences had pairs of inverted IS*1*-YK2X8 elements at their ends, indicating active transposition as composite transposons (Fig. 3C; Supplementary Figs. S21 and S22). Well-known composite transposons are typically <10 kbp, and in our study, simple IS insertions were observed on average 8.3 times per line (Table 1), and 5.5% of simple insertions are expected to be found within 10 kbp (according to the null-model simulations as in Supplementary Fig. S17). We believe that similar experiments, for instance, under selective conditions relevant to pathogenicity, may rapidly generate composite transposons with clinically relevant functions. Furthermore, investigating the conditions and mechanisms that enhance the formation of composite transposons would also be an intriguing direction for further studies.

Surprisingly, lines without *recA* showed 2- to 4-fold fewer deletions, duplications, and inversions than lines with *recA* (Table 1). This contradicts previous findings that *recA* deficiency typically decreases recombination rates by 10-fold [44, 63]. One possibility is that the modified IS sequence with the inducible promoter and larger size may have favored *recA*-independent transposase-mediated recombinations [64] over the *recA*-dependent pathways used by wild-type IS*1*. This could have maintained high recombination rates regardless of *recA*, reducing differences between the two genetic backgrounds. Alternatively, the excessive IS activity in *recA*$^{+}$ strains might have selected for cells with reduced recombination. With 26.8 new IS insertions per line, occurring more than once per passage (Table 1), many cells containing multiple mutations could have been lost in subsequent passages due to fitness costs.

A general caveat of our study is that engineering IS*1* could have led to relative rates of different structural mutation types that differ from those in wild-type *E. coli*. For instance, recombination rates typically increase with size [63]. Our modified IS, three times larger than wild-type IS*1*, may have caused higher recombination rates relative to simple insertions than in wild-type strains. The frequent composite transpositions, absent from the other laboratory evolution studies [4, 13], may have resulted from the modifications we made to the IS. Future studies should examine how mutations in ISs, including their sizes, influence recombination rates across different genetic backgrounds. Additionally, the modifications may have altered self-insertion rates, affecting both the frequency and diversity of observed IS variants. The larger size of our ISs should have increased the likelihood of self-insertions. In contrast, fusing the two open reading frames of IS*1* transposase to increase its activity [21] may have disrupted the transposase's *cis*-preference [65], whereby it preferentially acts on its source IS, potentially reducing self-insertion rates. Furthermore, since we modified IS regulation to accelerate genome evolution, identifying natural conditions that trigger IS expansion is necessary to determine the relevance of our results to the early stages of host restriction we aimed to simulate.

There are four additional limitations worth noting. First, our proposed method to accelerate genome structure evolution requires improvement, as the observed genome evolution remained minor compared to natural evolution, which can involve megabase-pair changes in genome sizes and hundreds of IS copies [5]. Second, the bioinformatics pipeline requires further development. Although we automated much of the analysis, numerous steps demanded manual intervention, including resolving complex genome rearrangements, identifying the parsimonious events that led to nested duplications, verifying IS insertion site assignments in duplicated regions, and classifying duplication types. Third, the MA protocol and genome sequencing could have been improved to obtain more accurate genome structures. Observed IS insertion rates exceeded expectations, with more than one insertion per passage in *recA*$^+$ strains (Table 1), suggesting colonies contained cells with diverse mutations. Also, due to the high IS activity, cells during post-MA incubation may have led to some cells in the population having additional mutations. Heterogeneity due to these factors necessitated manual resolution of misassemblies in some genomes. Future studies might mitigate this by streaking colonies onto plates without inducers at the passage that is to be sequenced and directly extracting DNA from these colonies. Furthermore, more frequent sequencing and improving the accuracy and length of long-read sequencing would help resolve these issues by facilitating inference of parsimonious evolutionary paths as well as more accurate detection of complex rearrangements and IS variants. The improved sequencing would also allow for the detection of point mutations within the ISs, further elucidating the mechanisms by which ISs evolve. Fourth, MA studies cannot completely eliminate the effects of selection [66]. We minimized selection on growth rate by imposing stabilizing selection on colony sizes, maintaining ∼26.5–27.7 generations per passage, in addition to bottlenecks during transfers. Nevertheless, we cannot exclude the possibility that certain combinations of deleterious or beneficial mutations were fixed by selection.

An important direction for future research is to apply our method to clarify mechanisms underlying the coexistence of ISs in bacterial genomes, maintained through horizontal transfer, transposition, loss, and fitness effects. Comparative genomic analyses suggest ISs copy numbers are balanced primarily by gain through horizontal transfer and loss through deletion, with episodes of extreme expansion [67–69]. While our study provides direct evidence that ISs can expand without selection for beneficial insertions (Fig. 2B), it is limited to mimicking the transient proliferation phase within the long-term evolutionary dynamics of ISs. Future studies extending our method to rapidly simulate genome evolution under varying IS activities, copy numbers, horizontal transfer rates, and selection pressures could verify conclusions from comparative studies and elucidate mechanisms of long-term IS maintenance in bacterial genomes.

The slow pace of IS-mediated evolution has been a bottleneck in studying genome evolution, leading to the reliance on comparative studies. Organisms with high IS activity are often slow growers [70] or pathogenic [12]. Artificial genome rearrangement methods, such as protoplast fusion [71, 72] or site-specific recombination on synthesized genomes [73] do not reflect typical natural processes. Applying our method to rapidly growing, safe, and experimentally tractable organisms like *E. coli* could serve as a proxy for studying the evolutionary dynamics of unculturable or pathogenic species. We believe our accessible method to observe genome evolution under controlled laboratory conditions stimulates further studies toward an experimentally grounded understanding of the principles behind bacterial genome evolution.

Acknowledgements

The text has been edited with the assistance of language models, ChatGPT (4o), Claude (Sonnet), and Github Copilot (v1).

Author contributions: Yuki Kanai (Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Software, Validation, Visualization, Writing—original draft, Writing—review & editing), Atsushi Shibai (Methodology, Resources), Naomi Yokoi (Investigation), Saburo Tsuru (Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing—review & editing), and Chikara Furusawa (Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing—review & editing)

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

This work was supported by the Japan Society for the Promotion of Science [21J20693 to Y.K., 18H02427 to S.T., 24K21985 to S.T., 22K21344 to C.F., 23K27164 to C.F., 24H01798 to C.F]; the Japan Science and Technology Agency [JPMJER1902 to C.F.]; the Japan Agency for Medical Research and Development [JP223fa627001 to Y.K.]; and a Grant-in-Aid from the Institute for Fermentation, Osaka [G-2025-2-046]. Funding to pay the Open Access publication charges for this article was provided by the Institute for Fermentation, Osaka [G-2025-2-046].

Data availability

The raw sequencing data and the assembled genomes are available at DDBJ (PRJDB17574). The plasmid sequence and custom scripts used for the analysis is available on figshare (https://doi.org/10.6084/m9.figshare.27800133). The genome sequence can also be found with the codes.

References

1. Rodríguez-Gijón A, Nuy JK, Mehrshad M *et al.* A genomic perspective across Earth's microbiomes reveals that genome size in archaea and bacteria is linked to ecosystem type and trophic strategy. *Front Microbiol* 2022;**12**:761869. https://doi.org/10.3389/fmicb.2021.761869
2. Touchon M, Rocha EPC. Coevolution of the organization and structure of prokaryotic genomes. *Cold Spring Harb Perspect Biol* 2016;**8**:a018168. https://doi.org/10.1101/cshperspect.a018168
3. Siguier P, Gourbeyre E, Chandler M. Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiol Rev* 2014;**38**:865–91. https://doi.org/10.1111/1574-6976.12067
4. Raeside C, Gaffé J, Deatherage DE *et al.* Large chromosomal rearrangements during a long-term evolution experiment with

Escherichia coli. *mBio* 2014;**5**:e01377–14. https://doi.org/10.1128/mBio.01377-14

5. McCutcheon JP, Moran NA. Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol* 2012;**10**:13–26. https://doi.org/10.1038/nrmicro2670
6. Dekker JP. Within-host evolution of bacterial pathogens in acute and chronic infection. *Annu Rev Pathol Mech Dis* 2024;**19**:203–26. https://doi.org/10.1146/annurev-pathmechdis-051122-111408
7. Moran NA, Plague GR. Genomic changes following host restriction in bacteria. *Curr Opin Genet Dev* 2004;**14**:627–33. https://doi.org/10.1016/j.gde.2004.09.003
8. Plague GR, Dunbar HE, Tran PL *et al.* Extensive proliferation of transposable elements in heritable bacterial symbionts. *J Bacteriol* 2008;**190**:777–9. https://doi.org/10.1128/jb.01082-07
9. Kanai Y, Tsuru S, Furusawa C. Experimental demonstration of operon formation catalyzed by insertion sequence. *Nucleic Acids Res* 2022;**50**:1673–86. https://doi.org/10.1093/nar/gkac004
10. Kawecki TJ, Lenski RE, Ebert D *et al.* Experimental evolution. *Trends Ecol Evol* 2012;**27**:547–60. https://doi.org/10.1016/j.tree.2012.06.001
11. Shigenobu S, Watanabe H, Hattori M *et al.* Genome sequence of the endocellular bacterial symbiont of Aphids *Buchnera* Sp. APS. *Nature* 2000;**407**:81–6. https://doi.org/10.1038/35024074
12. Hawkey J, Monk JM, Billman-Jacobe H *et al.* Impact of insertion sequences on convergent evolution of *Shigella* species. *PLOS Genet* 2020;**16**:e1008931. https://doi.org/10.1371/journal.pgen.1008931
13. Lee H, Doak TG, Popodi E *et al.* Insertion sequence-caused large-scale rearrangements in the genome of *Escherichia coli*. *Nucleic Acids Res* 2016;**44**:7109–19. https://doi.org/10.1093/nar/gkw647
14. Consuegra J, Gaffé J, Lenski RE *et al.* Insertion-sequence-mediated mutations both promote and constrain evolvability during a long-term experiment with bacteria. *Nat Commun* 2021;**12**:980. https://doi.org/10.1038/s41467-021-21210-7
15. Plague GR, Dougherty KM, Boodram KS *et al.* Relaxed natural selection alone does not permit transposable element expansion within 4,000 generations in *Escherichia coli*. *Genetica* 2011;**139**:895–902. https://doi.org/10.1007/s10709-011-9593-x
16. Pósfai G, Plunkett G, Fehér T *et al.* Emergent properties of reduced-genome *Escherichia coli*. *Science* 2006;**312**:1044–6. https://doi.org/10.1126/science.1126439
17. Baba T, Ara T, Hasegawa M *et al.* Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the keio collection. *Mol Syst Biol* 2006;**2**: 2006.0008. https://doi.org/10.1038/msb4100050
18. Datsenko KA, Wanner BL. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc Natl Acad Sci USA* 2000;**97**:6640–5. https://doi.org/10.1073/pnas.120163297
19. Cherepanov PP, Wackernagel W. Gene disruption in *Escherichia coli*: Tc^R and Km^R cassettes with the option of Flp-catalyzed excision of the antibiotic-resistance determinant. *Gene* 1995;**158**:9–14. https://doi.org/10.1016/0378-1119(95)00193-a
20. Meyer AJ, Segall-Shapiro TH, Glassey E *et al. Escherichia coli* "Marionette" strains with 12 highly optimized small-molecule sensors. *Nat Chem Biol* 2019;**15**:196–204. https://doi.org/10.1038/s41589-018-0168-3
21. Sekine Y, Ohtsubo E. Frameshifting is required for production of the transposase encoded by insertion sequence 1. *Proc Natl Acad Sci USA* 1989;**86**:4609–13. https://doi.org/10.1073/pnas.86.12.4609
22. Chandler M, Galas DJ. Cointegrate formation mediated by Tn*9*. II. Activity of IS*1* is modulated by external DNA sequences. *J Mol Biol* 1983;**170**:61–91. https://doi.org/10.1016/s0022-2836(83)80227-7
23. Chen YJ, Liu P, Nielsen AAK *et al.* Characterization of 582 natural and synthetic terminators and quantification of their design constraints. *Nat Methods* 2013;**10**:659–64. https://doi.org/10.1038/nmeth.2515
24. LaFleur TL, Hossain A, Salis HM. Automated model-predictive design of synthetic promoters to control transcriptional profiles in bacteria. *Nat Commun* 2022;**13**:5159. https://doi.org/10.1038/s41467-022-32829-5
25. Jack BR, Leonard SP, Mishler DM *et al.* Predicting the genetic stability of engineered DNA sequences with the EFM calculator. *ACS Synth Biol* 2015;**4**:939–43. https://doi.org/10.1021/acssynbio.5b00068
26. Reif HJ, Saedler H. IS*1* is involved in deletion formation in the gal region of *E. coli* K12. *Mol Gen Genet* 1975;**137**:17–28. https://doi.org/10.1007/BF00332538
27. Maeda T, Shibai A, Yokoi N *et al.* Mutational property of newly identified mutagen L-glutamic acid γ-hydrazide in *Escherichia coli*. *Mutat Res Mol Mech Mutagen* 2021;**823**:111759. https://doi.org/10.1016/j.mrfmmm.2021.111759
28. Kolmogorov M, Yuan J, Lin Y *et al.* Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* 2019;**37**:540–6. https://doi.org/10.1038/s41587-019-0072-8
29. Wick RR, Schultz MB, Zobel J *et al.* Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* 2015;**31**:3350–2. https://doi.org/10.1093/bioinformatics/btv383
30. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;**34**:3094–100. https://doi.org/10.1093/bioinformatics/bty191
31. Danecek P, Bonfield JK, Liddle J *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* 2021;**10**:giab008. https://doi.org/10.1093/gigascience/giab008
32. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* 2011;**17**:10–12
33. Robinson JT, Thorvaldsdóttir H, Winckler W *et al.* Integrative genomics viewer. *Nat Biotechnol* 2011;**29**:24–6. https://doi.org/10.1038/nbt.1754
34. Camacho C, Coulouris G, Avagyan V *et al.* BLAST+: architecture and applications. *BMC Bioinform* 2009;**10**:421. https://doi.org/10.1186/1471-2105-10-421
35. Pedregosa F, Varoquaux G, Gramfort A *et al.* Scikit-Learn: machine learning in Python. *J Mach Learn Res* 2011;**12**:2825–30
36. Goel M, Sun H, Jiao WB *et al.* SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol* 2019;**20**:277. https://doi.org/10.1186/s13059-019-1911-0
37. Goel M, Schneeberger K. plotsr: visualizing structural similarities and rearrangements between multiple genomes. *Bioinformatics* 2022;**38**:2922–6. https://doi.org/10.1093/bioinformatics/btac196
38. Biel SW, Berg DE. Mechanism of IS*1* transposition in *E. Coli*: choice between simple insertion and cointegration. *Genetics* 1984;**108**:319–30. https://doi.org/10.1093/genetics/108.2.319
39. Tenaillon O, Barrick JE, Ribeck N *et al.* Tempo and mode of genome evolution in a 50,000-generation experiment. *Nature* 2016;**536**:165–70. https://doi.org/10.1038/nature18959
40. Couce A, Caudwell LV, Feinauer C *et al.* Mutator genomes decay, despite sustained fitness gains, in a long-term experiment with bacteria. *Proc Natl Acad Sci USA* 2017;**114**:E9026–35. https://doi.org/10.1073/pnas.1705887114
41. Kato Ji, Hashimoto M. Construction of consecutive deletions of the *Escherichia coli* chromosome. *Mol Syst Biol* 2007;**3**:132. https://doi.org/10.1038/msb4100174
42. Lee H, Popodi E, Tang H *et al.* Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing. *Proc Natl Acad Sci USA* 2012;**109**:E2774–83. https://doi.org/10.1073/pnas.1210309109
43. Wiktor J, Gynnå AH, Leroy P *et al.* RecA finds homologous DNA by reduced dimensionality search. *Nature* 2021;**597**:426–9. https://doi.org/10.1038/s41586-021-03877-6
44. Reams AB, Kofoid E, Kugelberg E *et al.* Multiple pathways of duplication formation with and without recombination (RecA) in

Salmonella enterica. *Genetics* 2012;**192**:397–415. https://doi.org/10.1534/genetics.112.142570
45. Elena SF, Lenski RE. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nat Rev Genet* 2003;**4**:457–69. https://doi.org/10.1038/nrg1088
46. Darling AE, Miklós I, Ragan MA. Dynamics of genome rearrangement in bacterial populations. *PLOS Genet* 2008;**4**:e1000128. https://doi.org/10.1371/journal.pgen.1000128
47. Mahillon J, Chandler M. Insertion sequences. *Microbiol Mol Biol Rev* 1998;**62**:725–74. https://doi.org/10.1128/mmbr.62.3.725-774.1998
48. Alton NK, Vapnek D. Nucleotide sequence analysis of the chloramphenicol resistance transposon Tn*9*. *Nature* 1979;**282**:864–9. https://doi.org/10.1038/282864a0
49. Armbruster CR, Marshall CW, Garber AI *et al.* Adaptation and genomic erosion in fragmented *Pseudomonasaeruginosa* populations in the sinuses of people with cystic fibrosis. *Cell Rep* 2021;**37**:109829. https://doi.org/10.1016/j.celrep.2021.109829
50. He S, Chandler M, Varani AM *et al.* Mechanisms of evolution in high-consequence drug resistance plasmids. *mBio* 2016;**7**:10.1128/mbio.01987-16. https://doi.org/10.1128/mBio.01987-16
51. Kanai Y, Tsuru S, Furusawa C. Insertion sequences: simple mobile elements with rich ecological and evolutionary structures. *Curr Opin Syst Biol* 2023;**36**:100481. https://doi.org/10.1016/j.coisb.2023.100481
52. Moran NA. Microbial minimalism: genome reduction in bacterial pathogens. *Cell* 2002;**108**:583–6. https://doi.org/10.1016/S0092-8674(02)00665-7
53. Kuo CH, Moran NA, Ochman H. The consequences of genetic drift for bacterial genome complexity. *Genome Res* 2009;**19**:1450–4. https://doi.org/10.1101/gr.091785.109
54. Sela I, Wolf YI, Koonin EV. Theory of prokaryotic genome evolution. *Proc Natl Acad Sci USA* 2016;**113**:11399–407. https://doi.org/10.1073/pnas.1614083113
55. Lenski RE, Rose MR, Simpson SC *et al.* Long-term experimental evolution in *Escherichia coli*. I. Adaptation and divergence during 2,000 generations. *Am Nat* 1991;**138**:1315–41. https://doi.org/10.1086/285289
56. Cooper VS, Schneider D, Blot M *et al.* Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of *Escherichia coli* B. *J Bacteriol* 2001;**183**:2834–41. https://doi.org/10.1128/JB.183.9.2834-2841.2001
57. Siozios S, Nadal-Jimenez P, Azagi T *et al.* Genome dynamics across the evolutionary transition to endosymbiosis. *Current Biology* 2024;**34**:5659–70. https://doi.org/10.1016/j.cub.2024.10.044
58. Morisato D, Way JC, Kim HJ *et al.* Tn*10* transposase acts preferentially on nearby transposon ends in vivo. *Cell* 1983;**32**:799–807. https://doi.org/10.1016/0092-8674(83)90066-1
59. Li Y, Roberts ND, Wala JA *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* 2020;**578**:112–21. https://doi.org/10.1038/s41586-019-1913-9
60. Kouzminova EA, Rotman E, Macomber L *et al.* RecA-dependent mutants in *Escherichia coli* reveal strategies to avoid chromosomal fragmentation. *Proc Natl Acad Sci USA* 2004;**101**:16262–7. https://doi.org/10.1073/pnas.0405943101
61. Dale C, Wang B, Moran N *et al.* Loss of DNA recombinational repair enzymes in the initial stages of genome degeneration. *Mol Biol Evol* 2003;**20**:1188–94. https://doi.org/10.1093/molbev/msg138
62. Vandecraen J, Chandler M, Aertsen A *et al.* The impact of insertion sequences on bacterial genome plasticity and adaptability. *Crit Rev Microbiol* 2017;**43**:709–30. https://doi.org/10.1080/1040841X.2017.1303661
63. Oliveira PH, Lemos F, Monteiro GA *et al.* Recombination frequency in plasmid DNA containing direct repeats—predictive correlation with repeat and intervening sequence length. *Plasmid* 2008;**60**:159–65. https://doi.org/10.1016/j.plasmid.2008.06.004
64. Braedt G. Recombination in *recA* cells between direct repeats of insertion element IS*1*. *J Bacteriol* 1985;**162**:529–34
65. Duval-Valentin G, Marty-Cointin B, Chandler M. Requirement of IS*911* replication before integration defines a new bacterial transposition pathway. *EMBO J* 2004;**23**:3897–906. https://doi.org/10.1038/sj.emboj.7600395
66. Heilbron K, Toll-Riera M, Kojadinovic M *et al.* Fitness is strongly influenced by rare mutations of large effect in a microbial mutation accumulation experiment. *Genetics* 2014;**197**:981–90. https://doi.org/10.1534/genetics.114.163147
67. Bichsel M, Barbour AD, Wagner A. Estimating the fitness effect of an insertion sequence. *J Math Biol* 2013;**66**:95–114. https://doi.org/10.1007/s00285-012-0504-2
68. Iranzo J, Gómez MJ, de Saro FJL *et al.* Large-scale genomic analysis suggests a neutral punctuated dynamics of transposable elements in bacterial genomes. *PLOS Comput Biol* 2014;**10**:e1003680. https://doi.org/10.1371/journal.pcbi.1003680
69. Iranzo J, Cuesta JA, Manrubia S *et al.* Disentangling the effects of selection and loss bias on gene dynamics. *Proc Natl Acad Sci USA* 2017;**114**:E5616–24. https://doi.org/10.1073/pnas.1704925114
70. Miller SR, Abresch HE, Ulrich NJ *et al.* Bacterial adaptation by a transposition burst of an invading IS element. *Genome Biol Evol* 2021;**13**:evab245. https://doi.org/10.1093/gbe/evab245
71. Patnaik R, Louie S, Gavrilovic V *et al.* Genome shuffling of Lactobacillus for improved acid tolerance. *Nat Biotechnol* 2002;**20**:707–12. https://doi.org/10.1038/nbt0702-707
72. Zhang YX, Perry K, Vinci VA *et al.* Genome shuffling leads to rapid phenotypic improvement in bacteria. *Nature* 2002;**415**:644–6. https://doi.org/10.1038/415644a
73. Dymond JS, Richardson SM, Coombes CE *et al.* Synthetic chromosome arms function in yeast and generate phenotypic diversity by design. *Nature* 2011;**477**:471–6. https://doi.org/10.1038/nature10403.

Received: November 22, 2024. **Revised:** April 8, 2025. **Editorial Decision:** April 10, 2025. **Accepted:** April 14, 2025