A new route for integron cassette dissemination among bacterial genomes

Céline Loot, Gael A Millot, Egill Richard, Claire Vit, Baptiste Darracq, Vincent Parissi, Frédéric Lemoine, Théophile Niault, Jean Cury, Eduardo Pc Rocha, et al.

▶ **To cite this version:**

Céline Loot, Gael A Millot, Egill Richard, Claire Vit, Baptiste Darracq, et al.. A new route for integron cassette dissemination among bacterial genomes. 2022. pasteur-03826688

HAL Id: pasteur-03826688
https://pasteur.hal.science/pasteur-03826688

Preprint submitted on 24 Oct 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A new route for integron cassette dissemination among bacterial genomes

Céline Loot[1#], Gael A. Millot[2], Egill Richard[1,3], Claire Vit[1,3], Baptiste Darracq[1,3], Vincent Parissi[4,5], Frédéric Lemoine[2], Théophile Niault[1,3], Jean Cury[6], Eduardo PC Rocha[7] and Didier Mazel[1]

[1]Institut Pasteur, Unité Plasticité du Génome Bactérien, CNRS UMR3525, 75724 Paris, France.
[2]Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, F-75015 Paris, France
[3]Sorbonne Université, Collège doctoral, F-75005, Paris, France
[4]University of Bordeaux, Fundamental Microbiology and Pathogenicity Laboratory, CNRS, UMR 5234, SFR TransBioMed, Bordeaux, France.
[5] Viral DNA Integration and Chromatin Dynamics Network (DyNAVir), France.
[6]Université Paris-Saclay, Centre National de la Recherche Scientifique, Inria, Laboratoire de Recherche en Informatique UMR 8623, Orsay, France
[7]Institut Pasteur, Université Paris Cité, CNRS UMR3525, Microbial Evolutionary Genomics, 75724 Paris, France.
[#]Correspondent footnote: celine.loot@pasteur.fr
Tel: +33 1 4061 3287

Running title: Integron cassette dissemination in bacterial genomes

Key words: Bacterial evolution, antibiotic resistance, site-specific recombination, Integron, cassette dynamics, *attC* sites, genome-wide mapping.

Abstract

Integrons are genetic elements found exclusively in bacteria. They are well known for their role in disseminating antibiotic resistance genes among pathogens and more generally for enabling bacteria to rapidly adapt to changing environmental conditions. Integrons constitute a natural system to capture, stockpile, shuffle, express and disseminate genes embedded in cassettes. All these events are governed by the integron integrase through site-specific recombination between integron *att* sites (*attC* and *attI* sites). Here, we demonstrate that integron integrase can efficiently catalyze the insertion of cassettes in bacterial genomes, outside the *att* sites carried by the integron system. Surprisingly, analysis of more than 5×10^5 independent clones revealed that the genome recombination sites differ greatly, in terms of sequence and structure, from both classical *attC* and *attI* recombination sites. We named these

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

new sites *attG*. Notably, among these a few are driving integration at very high rates. We also showed that, once inserted in genomes, cassettes can be expressed if located near a bacterial promoter. Moreover, even if occurring at low frequency, genome inserted cassettes can be excised precisely or imprecisely, inducing in this latter case, chromosomal modifications. These results unveil a new route for antibiotic resistance dissemination and expand the role of integrons in bacterial genome evolution.

Introduction

Bacteria exchange and recombine DNA that accelerate their evolution and adaption to new stresses and environments. Understanding evolvability of bacterial human pathogens is crucial to fight infectious diseases. One of the most striking examples of their evolvability is the rapid emergence of multi resistant bacteria after the introduction of antibiotics in clinical practice. The integron system, discovered in the late 80s within Gram-negative pathogen bacteria, constitutes an active player of this antimicrobial resistance (AMR) (1). Integrons are natural genetic "toolboxes" able to stockpile, shuffle and express adaptive functions embedded in cassettes (2). Their evolutionary success relies on the diversity of these functions, among which one finds hundreds of antibiotic resistance genes. Anthropogenic pressures such as use of antibiotics have led to the selection of mobilization events in integrons, such as their association with transposons and conjugative plasmids. These so-called Mobile Integrons (MIs) have now disseminated in many environmental bacterial niches and are currently prevalent in hospitals worldwide. They have been found associated to 190 AMR cassettes comprising resistance determinants against almost all antibiotic families. Importantly, MIs are only the tip of the iceberg since much larger Sedentary and Chromosomal Integrons (SCIs) have been found in approximately 17% of the genomes available in databases (3,4). Both MIs and SCIs share however the same general organization: a stable platform and a variable cassette array (Figure 1). The platform is composed of three key elements: the *intI* gene coding for a specific type of

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

tyrosine recombinase, the integron integrase, the primary *attI* recombination site in which promoterless cassettes are inserted, and a P_C promoter oriented to direct transcription of proximal cassettes (2). The rest of the cassettes (more distal) represents a low-cost memory of valuable functions for the cell which can potentially be expressed through the reordering of the cassette array. As said above, a cassette is a mobilizable element that generally contains a promoterless coding sequence (CDS) and is ended by an IntI-recognizable recombination site called the *attC* site. The cassette reordering is ensured by two types of recombination events, namely cassette excision and cassette insertion. Cassette excision occurs through recombination between two *attC* sites while integration mostly occurs through recombination between an *attC* site (carried by an excised circular cassette) and the *attI* site (Figure 1). A key point of the integron system is that *attC* sites share little sequence conservation. They are instead recognized by the integrase through their single-stranded secondary structure that can be formed thanks to its remarkably conserved palindromic organization (5). The two strands of *attC* sites do not have the same propensity for recombination. The bottom strand (bs) is on average 1000 times more recombinogenic than the top strand (ts) (5). This strand selectivity is essential to the integration of cassettes in the correct orientation relatively to the P_C promoter and hence to the expression of the promoterless gene they contain. Importantly, cassette shuffling is highly regulated: the expression of the integrase ensuring cassette recombination is controlled by the bacterial stress (SOS) response, itself triggered by some widely used antibiotics (6-8). By controlling the expression of IntI, bacteria can reshuffle the integron cassettes in moments of stress. All these features make the integron a unique recombination system.

As a result of the recent burst of metagenomic sequencing and bioinformatics tools development, a substantial number of cassettes were identified in many bacteria species from very different environments and integron cassette databases are constantly fed at an increasingly high rate (9). In MIs, many of the identified cassettes provide resistance to antibiotics, whereas

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

cassettes found in SCIs are involved in mobility, metabolism, biofilm formation, bacteriophage resistance or host surface polysaccharide modification (2,4,9). Here, we focused on the ability of bacteria to access new functions and to build a repertoire of cassettes pertinent for their lifestyle.

Early studies on integrons showed the recombination properties of cassettes in the proper *att* sites of the integron. The ability of integrons to disseminate cassettes outside these specific sites, *i.e.* within genomes, has been overlooked for decades as being of little significance (10-12). Despite this, we have identified here, in more than 9000 fully sequenced bacterial genomes available in genomic databases, numerous isolated cassettes (CDSs associated with the *attC* site with no other *att* integration sites nearby). We named these cassettes SALIN, for Single *attC* site lacking integron-integrase and proposed that they may result from cassette integration by recombination events in genomes. By performing a series of *in vivo* experiments, we confirmed the high propensity of integron cassettes to disseminate in bacterial genomes. We also demonstrated that these cassettes can be expressed when inserted near endogenous bacterial promoters. We then constructed libraries of cassettes inserted into the *E. coli* genome. Deep sequencing of these libraries has shown that cassettes can target a very large number of unique sites and has allowed to characterize these new insertion sites. Surprisingly, they differ greatly, in terms of sequence and structure, from both classical *attC* and *attI* recombination sites. We named these new sites, "*attG"* sites for attachment site of the genome.

These results revisit the classical model of cassette recombination and reveal a new and efficient pathway for cassette dissemination extending the role of integrons in bacterial evolution.

MATERIALS AND METHODS

Bacterial strains, plasmids and primers

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The different plasmids and strains that were used in this study are described in Table S1 and S2. Primers used to construct the different strains and vectors are listed in Table S3.

Media

Escherichia coli and *Vibrio cholerae* were grown in Luria Bertani broth (LB) at 37°C. *E. coli* strains containing a plasmid with a thermo-sensitive origin of replication were grown at 30°C. In the case of *E. coli*, antibiotics were used at the following concentrations: carbenicillin (Carb), 100 µg/ml, kanamycin (Km), 25 µg/ml, chloramphenicol (Cm), 25 µg/ml, spectinomycin (Sp), 50 µg/ml. Diaminopimelic acid (DAP) was supplemented when necessary to a final concentration of 0.3 mM. To induce the P_{bad} promoter, L-arabinose (Ara) was added to a final concentration of 2mg/ml; to repress it, glucose (Glc) was added to a final concentration of 10mg/ml. To induce the P_{tet} promoter, anhydrotetracycline (aTc) was added to a final concentration of 1µg/ml. *V. cholerae* strains were cultivated in the same conditions and with the same antibiotic concentrations except for Cm and Sp, that were supplemented at a final concentration of 5 µg/ml and 100 µg/ml respectively. When *V. cholerae* strains were cultivated in presence of glucose, the later concentration of Sp was increased 2-fold (200 µg/ml).

Suicide conjugation assays

The conjugation assays were based on that of Vit et al, 2021 (13). Conjugation ensures the delivery of one of the recombination substrates (*att* sites) carried by a pSW23T suicide plasmid into a recipient cell expressing the IntI integrase (carried by pBAD43 or pBAD18 plasmids) and containing, or not, a second *att* recombination site (*attI* or *attC* sites) carried on pSU38, pBAD43 or pTOPO derivate plasmids. From the donor *E. coli* ß2163 strain, the pSW23T plasmid is delivered in a single stranded form into a recipient *E. coli* MG1655 or *V. cholerae* N16961. This plasmid contains an RP4 origin of transfer (*oriT*RP4) oriented in such way to deliver either the reactive bottom strand of the *attC* recombination sites or the top one. Recipient strains cannot sustain replication of this suicide pSW plasmid and the only way for the pSW

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

vector to be maintained in recipient cells is to recombine with *att* sites contained in the recipient strain or to insert in the bacterial genome. The frequency of cassette recombination is measured by comparing the number of recombined cells having acquired the resistance marker carried by the pSW23T vector, and the total number of recipient cells. The donor strains were grown overnight in LB media supplemented with Chloramphenicol (Cm) (resistance of the pSW plasmid), Kanamycin (Km) (resistance of the ß2163 strain) and DAP (since ß2163 donor strain requires DAP to grow in rich medium), the recipient strain was grown overnight in LB media supplemented with appropriate antibiotics depending on the plasmids used and Glucose (Glc, to repress the integrase gene). Both donor and recipient overnight cultures were diluted 1/100 in LB with DAP or Arabinose (Ara) respectively and incubated until OD=0.7-0.8. 1ml of each culture were then mixed and centrifuged at 6000rpm for 6mins. The pellet was suspended in 100µl LB, spread on a conjugation membrane (mixed cellulose ester membrane from Millipore, 47mm diameter and 0.45µm pore size) over a LB+agarose+DAP+Ara Petri dish and incubated 3H for conjugation and recombination to take place. The membrane with the cells was then resuspended in 5ml LB, after which serial 1:10 dilutions were made and plate on LB+agarose media supplemented with appropriate antibiotics. The recombination frequency was calculated as the ratio of recombinant CFUs, obtained on plates containing Cm and the antibiotics corresponding to recipient cells, to the total number of recipient CFUs, obtained on plates containing only antibiotics corresponding to recipient cells. The overall recombination frequency is a mean of at least 3 independent experiments.

Note that we adapted this protocol when using *V. cholerae* as recipient strain in which plasmids are more easily lost in absence of antibiotic selection than in *E. coli*. In this case, after an overnight culture, recipient cells were diluted (1:100) and grown in presence of Sp and Ara (0.2%) respectively to maintain pBAD43 vector and to allow the expression of the integrase. The donor strain was grown in parallel in presence of DAP. When both donor and recipient

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

cultures reach an OD_{600nm} of 0.7-0.8, recipient *V. cholerae* strains were washed twice by centrifugation at 6000rpm for 6mins and resuspension of the pellet in 1ml of LB. 1ml of each donor and receptor cultures was then mixed and centrifuged at 6000rpm for 6mins. The obtained pellet was re-suspended in a droplet of LB and spread on a 0.45 µm filter placed on MH DAP, Ara plates and incubated at 37°C during 3h. From this part, we proceed as described for *E. coli*.

Conjugation assay in presence of integron

To determine the genome cassette insertion frequency in the presence of an integron, we performed a conjugation assay proceeding exactly as described above. We used the $attC_{aadA7}$-containing pSW plasmid (pD060) as donor plasmid and we constructed a temperature-sensitive replicating receptor plasmid containing an *attI1* site followed by a spectinomycin resistant $attC_{aadA1}$ cassette (pM335). The DH5α cell is used as host for cloning this plasmid. Once constructed, this vector was transformed into the MG1655 recipient strain containing the pBAD18::P_{BAD}-*intI1* plasmid (p3938, $Carb^R$). These donor and recipient strains were conjugated for 3h but at 30°C (to ensure the replication of the pM335 plasmid). The membrane with the cells was then resuspended in 2ml LB+Carb+Glc, divided in two parts, each spread on a new conjugation membrane over a LB+agarose+Carb+Glc Petri dish and incubated 20H at 30°C and 42°C. These incubation temperatures respectively favor and disfavor the pM335 plasmid maintenance. The recombination frequency of cassettes was calculated as the ratio of recombinant CFUs, obtained on Cm (to select cassette insertion), Carb (resistance of the used integrase-carrying plasmid) and Glc (to repress the *intI1* gene) containing plates, to the total number of recipient CFUs, obtained on Carb and Glc containing plates. Note that plates were respectively incubated at 30°C and at 42°C. The overall recombination frequency is a mean of at least 3 independent experiments.

Cassette expression assay

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

To test if *attC* cassettes can be expressed once inserted in genome, we performed a conjugation assay proceeding exactly as described above. We constructed a suicide plasmid vector containing the $attC_{aadA7}$ but adding a kanamycin (*km*) resistance gene without promoter just downstream the *attC* site (pN695-pN697, Cm^R). We also added two different RBS just upstream the *km* gene (RBS1, pN708-pN709 and RBS2, pN705-pN707, Cm^R). The Π1 cell is used as host for cloning these plasmids. Once constructed, these plasmids were transformed into the β2163 donor strain. These donor strains and the recipient MG1655 strain containing the pBAD43::P_{BAD}-*intI1* plasmid (pL294, Sp^R) were conjugated. The recombination frequency of cassettes was calculated as the ratio of recombinant CFUs, obtained on Cm (to select cassette insertion), Sp (resistance of the used integrase-carrying plasmid) and Glc (to repress the *intI1* gene) containing plates, to the total number of recipient CFUs, obtained on Sp and Glc containing plates. The recombination frequency of solely cassettes expressing the *km* resistance gene was calculated as the ratio of recombinant CFUs, obtained on Km, Cm, Sp and Glc containing plates, to the total number of recipient CFUs, obtained on Sp and Glc containing plates. The overall recombination frequency is a mean of at least 3 independent experiments.

Cassette excision assay

To test if *attC* cassettes can be excised once inserted in genome, we performed a conjugation assay proceeding exactly as described above. We constructed a new suicide plasmid vector containing the $attC_{aadA7}$ and the *ccdB* toxin gene under the control of the P_{BAD} promoter (pM779-pM781, Cm^R). The Π3813 cell, a *pir*+ CcdB resistant *E. coli* strain (14), is used as host for cloning this plasmid. Once constructed, this plasmid was transformed into the β3914 donor strain to perform conjugation. We also constructed a new vector expressing the integrase. This vector is a temperature-sensitive replicating vector, and we replaced the P_{BAD} promoter controlling the *intI1* gene expression by a P_{TET} promoter (pN435-pN437, Km^R). DH5α cells were used as host for cloning this plasmid. Once constructed, this plasmid was transformed into

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the MG1655*ΔrecA* recipient strain. Both donor and recipient strains were conjugated. Conjugation was performed by adding anhydrotetracycline (aTc) in place of Ara to express the integrase gene and at 30°C. We also used Glc to repress the *ccdB* toxin gene. Cassette insertion events in genome were selected by plating cells on Cm (to select cassette insertion), Sp (resistance of the used integrase-carrying plasmid) and Glc (to repress the *ccdB* toxin gene) containing plates. After that, 24 recombinants clones were picked and cultivated at 42°C to ensure the loss of the thermosensitive integrase expressing plasmid. The 24 obtained recombinant clones were transformed with the pBAD43::P_{TET}-*intI1* plasmid (pM888, Sp^R) and, as control, with the pBAD43::P_{TET} plasmid (pM889, Sp^R). Clones were grown for 8h in presence of aTc (to express the integrase gene), Sp and Glc. The aim of this step is to promote successful recombination event leading to cassette excision. The excision frequency of cassettes was calculated as the ratio of recombinant CFUs, obtained on Sp and Ara containing plates, to the total number of recipient CFUs, obtained on Sp and Glc containing plates. Note that in presence of Ara (to express the *ccdB* toxin gene), only the clones in which the P_{BAD}*ccdB*-containing plasmid has been excised can survive. We checked the Cm sensitivity of a large number of clones. The overall recombination frequency is a mean of at least 3 independent experiments.

Testing the hotspots as receptor and donor sites

To test if the hotspots, the median spot (MS) and the unique spot (US) can be used as donor or receptor sites, we performed conjugation assays proceeding exactly as described above.

To test the hotspots, MS and US sites as receptor sites, we used the $attC_{aadA7}$-containing pSW plasmid (pD060) as donor plasmid and we constructed receptor plasmids containing the different hotspots, MS and US. The Top 10 cell is used as host for cloning the sites. Once constructed, each vector was transformed into the MG1655 recipient strain containing the pBAD43::P_{BAD}-*intI1* plasmid (pL294, Sp^R). The donor strain and these recipient MG1655

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

strains were conjugated. The recombination frequency of cassettes was calculated as the ratio of recombinant CFUs, obtained on Cm (to select cassette insertion), Sp (resistance of the used integrase-carrying plasmid) and Glc (to repress the *intI1* gene) containing plates, to the total number of recipient CFUs, obtained on Sp and Glc containing plates. The overall recombination frequency is a mean of at least 3 independent experiments.

To test the hotspot sites as donor sites, we constructed suicide plasmid vectors containing the *ybhO* hotspot site delivering either the bottom strand (pO323-pO324) or the top one (pO321-pO322), the *aslB* hotspot site delivering either the bottom strand (pO749-pO750) or the top one (pO751) and the *pyrE* hotspot site delivering either the bottom strand (pO752) or the top one (pO753-pO755). The Π1 cells were used for cloning these plasmids. Once constructed, these plasmids were transformed into the β2163 donor strain. These donor strains and the recipient MG1655 *recA* strain containing the pBAD43::P_{BAD}-*intI1* plasmid (pL294, Sp^R) and the pSU38Δ::$attC_{aadA7}$ (pO371-pO372) were conjugated. The recombination frequency of cassettes was calculated as the ratio of recombinant CFUs, obtained on Cm (to select cassette insertion), Sp (resistance of the used integrase-carrying plasmid), Km (resistance of the used $attC_{aadA7}$-carrying plasmid) and Glc (to repress the *intI1* gene) containing plates, to the total number of recipient CFUs, obtained on Sp, Km and Glc containing plates. The overall recombination frequency is a mean of at least 3 independent experiments.

Analysis of recombination events and point localization

For each experiment, clones were picked and isolated on antibiotic containing plates. Recombination events were checked by Polymerase chain reaction (PCR) using the DreamTaq DNA polymerase (Fisher Scientific). All the PCRs were directly performed on, at least, eight randomly chosen bacterial clones per experiment. Some PCR reactions were purified using the PCR purification kit (Fisher Scientific) and sequenced to confirm the insertion point (Eurofins).

Insertion events in att sites carried on pSU38 vector

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

For analysis of co-integrates formation, we performed PCR reactions on randomly chosen clones per each experiment using SWend/MRV primers to confirm the bs recombination (when bs is injected), SWend/MFD primers to confirm the bs recombination (when ts is injected) and Swend/MRV primers to confirm the ts recombination (when the ts is injected). Recombination points were precisely determined by sequencing PCR products using MRV or MFD primers.

Insertion events in hotspot, MS and US sites carried on pTOPO vector

For analysis of co-integrates formation, we performed PCR reactions on randomly chosen clones per each experiment using SWbeg/MFD primers to confirm the *ybhO*, *aslB*, *ilvD*, *pyrE* and *yjhH* hotspot recombination and using SWbeg/MRV primers to confirm the *metC* hotspot, MS-*abgA* and US-*ygcE* recombination. Recombination points were precisely determined by sequencing PCR products using Swbeg primers.

Insertion events in att sites carried on pSC101ts vector

For analysis of co-integrates formation, we performed PCR reactions on randomly chosen clones per each experiment using SWbeg/o1714 primers to confirm the *attI1* recombination and using SWend/o1704 primers to confirm the $attC_{aadA1}$ recombination. Recombination points were precisely determined by sequencing PCR products using o1714 or o1704 primers.

Insertion events on bacterial genome

For analysis of insertions of *attC* containing-pSW23T plasmids in *E. coli* genome (at secondary sites), we performed random PCR. For these, we performed a first random PCR reaction using the o1863 degenerated and the o2405 primers. The o2405 primer hybridizes upstream of the *attC* sites on pSW23T plasmids. Due to the presence of degenerate nucleotides in the o1863 primer, low hybridization temperatures were used, first, 30°C during 5 cycles and after, 40°C during 30 cycles. The obtained amplified DNA fragments were subjected to a second PCR reaction to enrich for PCR products corresponding to cassette insertion. For this purpose, we used o1865 and o1388 primers. These primers hybridize respectively to the fixed part of the

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

degenerated o1863 primer and upstream (but closer than o2405) of the *attC* sites on pSW23T plasmids. Recombination points were precisely determined by sequencing PCR products using o1366. The o1366 primer hybridizes upstream (but closer than o1388) of the *attC* sites on pSW23T plasmids.

Excision events on bacterial genome

For analysis of excision events from insertions of *attC* containing-pSW23T plasmids in *E. coli* genome, we performed PCR. Note that to perform PCR, we designed appropriate primers based on the knowledge of cassette recombination points during insertion. We performed PCR on clone 7, 8 and 9. We used o6263/o6264 for clone 7, o6265/o6266 for clone 8 and o6267/o6268 for clone 9. Excision points were precisely confirmed by sequencing the PCR products.

Principles of insertion profiling by deep sequencing

Library preparation

Clones obtained from three independent conjugation assays were collected and subjected to genomic extraction using the kit DNeasy® Tissue Kit (Qiagen). DNA was mechanically fragmented using the Covaris method (DNA shearing with sonication). Adaptors are ligated to the fragmented DNA pieces. Nested PCR, including 30 rounds of amplification, was performed to amplify low-abundance junctions in the DNA population using first o1366 and o6036 (to amplify the cassette genome junction, Table S3) and second o6035 and o6036 (to reconstitute adaptors, Table S3). PCR-enriched junctions were deep-sequenced using next-generation sequencing (NGS) technologies (Illumina MiSeq v3 single-end 150 cycles).

Bioinformatics analysis

The Nextflow pipeline used to analyze the raw fastq files and generate the Figure 7 and S7 is available at https://gitlab.pasteur.fr/gmillot/14985_loot and is briefly described here. First, non-genomic sequences such as barcodes and linkers were trimmed from the raw fastq reads. Then, reads showing the *attC* sequence in 5', expected by the PCR-enriched junctions described

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

above, were selected and the *attC* sequence was trimmed such that the 5' end of reads corresponds to the insertion site in the genome. Only reads showing at least 25 nucleotides post-trimming were kept and aligned on the *E. coli* str. K-12 substr. MG1655 reference genome sequence (NCBI NC_000913.3) using the --very-sensitive option of Bowtie2 (15). Q20 mapped reads were selected and checked for absence of soft clipping in each extremity. In our experimental design, reads showing the same plasmid insertion site result from three non-exclusive processes: 1) same plasmid insertion site in two different bacteria, 2) bacteria clonal amplification and 3) DNA PCR enrichment. The MarkDuplicates-Picard tool of GATK (https://github.com/broadinstitute/gatk) could not be used to remove read duplicates, as eliminating read showing identical 5' end would also remove reads resulting from the first process. To alleviate such stringency, we considered as duplicates reads showing the same 5' and 3' extremities, and we analyzed libraries with and without removing the duplicates. Position of plasmid insertion was defined by the 5' extremity of forward and 3' extremity of reverse read alignments. Sequence around insertion sites were extracted using bedtools (https://bedtools.readthedocs.io/en/latest/) and the nucleotide consensus of these n sequences were visualized with the R package ggseqlogo (16). Random insertions were determined by deducing a sequence motif from the insertion consensus and by randomly select with replacement n positions among all the motif positions present in the reference genome.

Genomics analysis of cassette insertion sites

Data

The sequences and annotations of complete genomes were downloaded from NCBI RefSeq (accessed in February 2018, http://ftp.ncbi.nih.gov/genomes/refseq/bacteria/). Using the IntegronFinder program (https://github.com/gem-pasteur/Integron_Finder), we analyzed 9078 bacterial genomes, including 9744 replicons labeled as chromosomes and 7810 replicons

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

labeled as plasmids. IntegronFinder ensures an automatic and accurate identification of integrons, cassette arrays, and *attC* sites.

SALIN analysis in sequenced strains

SALIN elements are single *attC* sites lacking integrase. They are a subtype of CALINs where the cluster is composed of 1 *attC*. SALINs were detected with IntegronFinder v2 with the option "--calin-threshold 1" to not filter out single *attC* sites.

RESULTS

In silico detection of isolated cassettes in bacterial genomes

Several computational pipelines were designed to detect integrons and *attC* sites using bacterial metagenomic data (3,9,17). All these surveys detected large numbers of single *attC* sites corresponding to isolated cassettes, but these cassettes were systematically removed from the analysis as considered to be an experimental bias of the method and likely false positive results. We took advantage of the recent release of IntegronFinder 2.0 (4) to detect the presence of isolated cassettes in the 9078 bacteria complete genomes of the RefSeq NCBI database. By analogy with the previously described CALINs (Clusters of *attC* sites lacking integron-integrases (3), we named these isolated cassettes, SALINs, for Single *attC* site lacking integron-integrases. We obtained 1021 genomes containing SALINs, 1049 containing CALINs (more than 2 consecutive *attC* sites), 939 containing complete integrons and 156 containing In0 (Integrase without cassettes *i.e.,* no *attC* sites) (Figure 2A). Among the 1021 genomes containing SALINs, 741 contain only one SALIN and the remaining 280 more than one SALIN (Figure 2B). Interestingly, SALINs are more represented than CALINs since we found 1433 SALINs widespread in genomes and only 970 CALINs (Figure 2C). CALIN are thought to arise from integrons by integrase gene loss caused by deletions or pseudogenization events or by rearrangements of parts of the cassette array mediated by transposable elements (3). However, previous analysis showed that most CALIN (95%) are not close to recognizable *intI*

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

pseudogenes (3). Furthermore, we observe that many CALINs are not surrounded by transposable elements and tend to be small: among the 970 CALINs, 491 harbor only 2 cassettes and 206 only 3 (Figure 2C). This raises the possibility that small CALINs and SALINs could also originate from another type of events. Rather than being remnants of integrons, some of these elements could result from the insertion of integron cassettes in the genome (and not on *attI* sites).

Cassette insertions in *att* and genome sites

To validate our hypothesis, we tested the cassette insertion capability of the class 1 MI cassettes in either *att* or in genome sites. For this, we used our previously developed conjugation suicide assay to exclusively deliver the *attC*-containing plasmids on a single-stranded form in a recipient *E. coli* strain containing a vector expressing the integrase (Figure 3A) (5,13,18). This assay mimics the natural conditions in which cassettes are delivered through horizontal gene transfer. Once delivered in the recipient strain, this suicide plasmid cannot replicate and can be assimilated to a non-replicative integron cassette. Thus, the only way for these synthetic cassettes to be maintained is to be inserted in a functional replicon. Here, the donor strain contains a plasmid carrying the $attC_{aadA7}$ class 1 MI site and recipient strains contain a plasmid carrying the second *att* partner either an *attI1* or an *attC* (VCR or $attC_{ereA2}$) site (Figure 3A and Figure S1). The obtained rate of cassette recombination reached more than 10^{-1} using *attI1* and was slightly lower (around 10^{-2}) using *attC* sites, but clearly above the rates obtained without integrase expression (Figure 3B, left panel). We then checked where insertion occurred in recombinant clones by performing PCR and sequencing. In the presence of the *attI1* site, all cassette insertions tested (117/117) took place in the expected 5'AAC3' triplet of this site (Figure 3B, middle panel). However, this was not always the case using *attC* sites, resulting in the absence of PCR amplification (Figure 3B, middle panel). To determine where the donor

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

plasmid was inserted in these clones, we designed a random PCR approach using one primer hybridizing in the *attC*-containing donor plasmid and another degenerated primer (Materials and methods, and Figure S2). PCR amplicon of various sizes suggested different recombination sites for each tested clone (Figure S2) and sequencing these PCR products indicated that the donor plasmid was not inserted in the recipient plasmid but in the host genome (Figure 3B, right panel).

As mentioned above, we failed to detect genomic insertion when using the *attI1* site probably due to the very high propensity of this site to recruit cassettes. We therefore designed an experimental procedure to easily determine if cassettes can be inserted in genomes even in the presence of an *attI1* site. For this, we performed our conjugation assay using a receptor strain containing a temperature-sensitive replicating plasmid carrying a synthetic integron (*i.e.,* an *attI1* site followed by one cassette ended by an $attC_{aadA1}$ site) (Figure S3A). The conjugation and recombination reactions were performed in presence of the synthetic integron *i.e.,* in condition under which the plasmid can replicate (at 30°C). After that, recombinant clones were selected, *in parallel*, at 30°C and at 42°C. At 30°C, all cassette insertion events (in *attI1*, $attC_{aadA1}$ and in genome sites) are selected. At 42°C, insertion events corresponding to insertion in the recipient plasmid are counter selected and those corresponding to insertion in the chromosome are therefore enriched and more easily detectable. To determine the precise frequency of such recombination events, some clones were analyzed by random PCR. As expected, at 30°C, cassette insertions are mainly detected in the *attI1* site (more than 10^{-1}, Figure S3B and S3C, left panels), while at 42°C, the selected recombination events are mainly detected in the chromosome (close to 10^{-2}, Figure S3B and S3C, right panels). Altogether, these results prove for the first time that integron cassettes can be disseminated in host genome at very high frequency even in the presence of a resident integron.

Cassette insertions in bacterial genomes

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

To better define the propensity of cassettes arising from horizontal gene transfer to insert in bacterial genomes, we performed the same conjugation assay, but in an *E. coli* recipient strain devoid of recipient plasmid carrying *att* sites (Figure 4A). We tested the integration properties of five different *attC* sites ($attC_{aadA7}$, $attC_{oxa2}$, $attC_{ereA2}$, $attC_{dfrB2}$ or VCR sites) delivered as synthetic cassettes by donor plasmids during suicide conjugation. These *attC* sites were chosen from both MIs and SCIs and based on their high recombinogenic properties (Figure S1, (19)). All the tested sites showed a rate of recombination comprised between 10^{-2} and 10^{-4} (Figure 4B, left panel), far above what was obtained without integrase expression. Using the previously described random PCR and sequencing approaches, we confirmed that cassettes were inserted at several locations in the *E. coli* genome. Furthermore, the rate of insertion dropped to 10^{-6} when using the *attI1* site or the top strand of the $attC_{aadA7}$ site, equivalent to the rates obtained in the absence of IntI1. Thus, *attI* sites do not recombine in genomes at a significant extent, whereas *attC* sites can recombine in genomes in the same way that they recombine with the *attI* and *attC* sites, *i.e.,* as a single-stranded structured form made by the bottom strand, but not by the top one.

Replacing the integrase IntI1 by IntI2 or IntI3, respectively the integrases of the class 2 and class 3 MIs, has no effect on the high rate of insertion of the $attC_{aadA7}$ cassette, either in the *E. coli* genome (Figure 4B) or in a recipient plasmid carrying the *attI2*/*attI3* or $attC_{ereA2}$ canonical sites (Figure S4), indicating that the property to insert integron cassettes in the *E. coli* genome is not restricted to IntI1. We also demonstrated that IntI1 can mediate cassette insertions at high frequency (almost 10^{-3}) in another Gram-negative bacteria, the pathogenic *V. cholerae* strain (Figure 4, right panel). Interestingly, *V. cholerae* contains a massive chromosomal integron located on the chromosome 2 and harboring 179 *attC* containing cassettes (20). However, using random PCR (on 48 randomly chosen clones) and sequencing (of 15 randomly chosen PCR products), we did not detect any insertion events in these *attC* sites carried by the resident SCI,

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

showing that integron cassettes can be inserted in the *V. cholerae* genome at very high frequency even in the presence of a resident integron.

Genome inserted cassettes can be expressed

Most genes present in integron cassettes are promoterless. To determine if cassettes can be expressed when inserted in the genome, we added a promoterless kanamycin resistance gene preceded or not by a ribosome binding site (RBS) into the donor plasmid containing the $attC_{aadA7}$ site (Figure 5A). This matches previous observations in real cassettes, where some genes are preceded by a suitably spaced RBS while others are lacking it (2). We tested the two different RBS1 and RBS2 sites that can be naturally found in cassette integrons. For all tested cassettes, a high frequency (more than 10^{-3}) of genomic insertions was observed using Cm selective medium, as expected (Figure 5B). Using the double Cm and Km selective medium, the insertion rate was far above the one without integrase expression, as soon as the donor plasmid presents the *km* resistance gene, and especially when it is preceded by a RBS motif. Comparing the insertion rate using the RBS2 site in the presence of Cm (2×10^{-3}) or Cm and Km media (2×10^{-4}) indicated that 10% of inserted cassettes are expressed when the *km* gene is associated with this RBS2 site. Analysis of over 40 randomly chosen Km^R clones using random PCR and sequencing confirmed that the expressed cassettes were inserted near a resident promoter (Figure 5C). These results demonstrate that a large proportion of the cassettes inserted into the genomes can be expressed if they are located in the vicinity of a promoter, thus conferring a new phenotype on the bacteria.

Genome inserted cassettes can be excised

To test if the cassettes inserted in the genomes could be excised or become fixed, we added, in the $attC_{aadA7}$-containing donor plasmid, a *ccdB* gene encoding a bacterial toxin under control of the P_{BAD} promoter (Figure 6A). The *ccdB* gene was previously used as a potent counterselection marker in several commonly used applications (14,21). First, we performed the conjugation assay

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and selected the cassette insertion events as described above. Second, we randomly chose 24 recombinant clones and independently analyzed their frequency of cassette excision. For this, we removed the resident thermosensitive plasmid expressing the integrase by cultivating the clones at 42°C and then reintroduced plasmids containing or not the integrase gene. We then induced the integrase to mediate the excision reaction and selected clones on arabinose containing plates to express the CcdB toxin. In these conditions, only clones that have lost the *ccdB* gene, due to a cassette excision event, are selected while the others die. Among the 24 independent clones, excision events were notable for clone 8 and clone 9 (Figure 6B). PCR and sequencing analysis revealed that excision can occur "precisely", *i.e.*, at the insertion site, or "imprecisely", *i.e.,* outside the insertion site at another proximal genome site (Figure 6B and S5). In the latter case, this could lead to genome modifications (Figure S5).

Cassette insertion occurs in a broad number of genomic locations

Deciphering where and how integron cassettes are inserted in genomes is fundamental to understand their potential cost and impact on host evolution. We therefore performed a genome wide NGS mapping of insertion sites (Figure S6) using a library of around 50,000 recombinant clones obtained after insertion of the $attC_{aadA7}$ containing plasmids in the *E. coli* genome catalyzed by the IntI1 integrase. Genomic sequences flanking the insertions were extracted from each sequencing read, aligned to the *MG1655 E. coli* reference genome, and used to call precise insertion sites. At first glance, genomic insertion occurred in many positions in the *E. coli* genome (22,271 unique insertion sites, Figure S7), with a huge variation in site usage (Figure 7A-B). To better analyze these data, we first removed the duplicated reads from the 3,205,043 reads (Figure 7 and Figure S7), leading to 361,464 reads, corresponding to the potential number of insertions that occurred in 50,000 recombinant clones. A window size of 200,000 bps sliding every 100 bps along the genome (Figure 7C) uncovered preferential insertions near the origin of replication. To decipher this bias, we performed multiplex digital PCR (dPCR). For this,

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

genomic DNA was purified at different time during cell conjugation. From the beginning of the conjugation to 60 min, the time frame during which the majority of conjugation events occur, the *oriC*/*terC* ratio was found to be ~2, indicating that cells contain twice as many *oriC* than *terC* copies (Figure S8), thus providing twice more DNA sequences for cassette insertion near the *oriC* than near the *terC*. This explains, at least in part, why insertions are favored near the origin of replication.

Alignment of all the DNA sequences flanking the integration cutting sites revealed a short 5'GWT3' consensus sequence (Figure 7D). Such motif is present 338,348 times in the *E. coli* reference genome: 169,209 and 169,139 times in the top and bottom strand, respectively. From here, we randomly selected with replacement 361,464 sites among the 338,348 5'GWT3' sites present in the genome to use them as a "random control" of cassette insertion (Figure 7E-I). Then, effect of genomic features on insertion site usage was analyzed. No obvious bias of insertion was detected depending on the forward or reverse strand of the genome (Figure 7E, left panel), and the same for the leading or lagging strand template during replication (Figure 7E, right panel), even if oscillations of usage were uncovered when regrouping the number of events in a sliding window (Figure 7F). Notably, these oscillations did not follow the random usage of the 5'GWT3' motifs available in the genome (Figure 7F). The other notable but expected bias detected was a decrease of insertions in the essential genes (7% (295/4213) of the *E. coli* genes according to (22)), when compared to non-essential genes or when using random insertions (Figure 7G-H), but with no obvious effect of insertion in the same or opposite direction of transcription. Finally, insertion around transcription start sites (TSS) appeared shifted upstream of essential genes (-105 bps, Figure 7I). Altogether, we conclude that IntI1 dependent plasmid integration can occur all along the *E. coli* genome with no notable effect of genomic features, except the 5'GWT3' DNA sequence motif and the avoidance of essential genes. We also do not exclude a potential *oriC* attractivity that would need to be explored

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

further. Interestingly, same results were observed using IntI2 and IntI3 in the *E. coli* genome (Figure S7). The only remarkable difference was a 1 base larger 5'TGWT3' consensus insertion site for IntI2 that may explain the lower frequency of cassette insertion in genomes when mediated by this integrase (Figure 4B).

Cassette insertions occur in hotspots in the *Escherichia coli* genome

Several insertion hotspots were denoted when considering all the 3,205,043 reads (*i.e.,* including the duplicate ones, Figure 7A), whatever the IntI1, IntI2 or IntI3 integrase tested (Figure S7). Note that for IntI1 and IntI3, the strongest hotspot was the same, located in the *ybhO* gene (Figure S7, red boxes). However, duplicate reads can be artifacts coming from the experimental procedure used (bacteria clonal amplification and DNA PCR enrichment, see the Materials and Methods section). Thus, to experimentally validate that these insertion sites are indeed insertion hotspots, the six strongest hotspots resulting from the IntI1 experiment (corresponding to insertions in *ybhO*, *aslB*, *ilvD*, *pyrE*, *metC* and *yjhH* genes in decreasing order) were cloned in a recipient plasmid (Figure 8A). As a control, we used a genomic site used only once by IntI1 for cassette insertion, called US-*ygcE* (US for Unique Site) and another site used 15 times, *i.e.,* very close to the median of insertion site usage, called MS-*abgA* site (MS for Median Site). In each case, the cloned segment encompassed the 300 and 200 bps flanking the 5'GWT3' insertion point. The six hotspots showed a rate of recombination above 5 $\times 10^{-4}$, far above what was obtained without integrase expression, while no recombination event was detected for the control sites (Figure 8B). Notably, the highest insertion efficiency was obtained for the *ybhO* hotspot, the one showing the highest insertion usage (209,387). These results confirm that these hotspots are regions attracting cassette insertions and not the consequences of experimental bias.

The properties of genome insertion sites differ from *attC* and *attI* recombination sites

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

To further characterize the attractivity of genome sites for cassette insertion, we used the *ybhO* site as proxy of genome insertion sites and determined its minimal functional length. We constructed and tested several lengths of base pairs on each 5' and 3' side of the cutting position of the *ybhO* hotspot site (Figure 8C). Decreasing the lengths to 9 nucleotides on each side maintained the recombination frequency above 10^{-3} (construct 9-9, Figure 8C), defining a minimal 18 nt length-functional insertion, while smaller lengths hampered the recombination frequencies (constructs 9-8, 8-9, 10-0, 0-10, Figure 8C). These functional 9-9 lengths were confirmed for the *aslB* hotspot (Figure 8D). Alignment of the 10 base pairs on each 5' and 3' side of the cutting position of the six hotspots from Figure 8B uncovered a consensus sequence, keeping the highly conserved 5'GWT3' residues in positions -1 to 2, but also revealing two supplementary motifs, on either side of the cleavage site (Figure S9A). We validated the importance of the right motif 5'CRGM3' in positions 6 to 9. Indeed, replacing this motif by the 5'CCAG3' and 5'TCAG3' motifs in the *ybhO* hotspot decreased insertion frequencies by more than 2 logs compared to the *wt ybhO* 5'CAGC3' or the consensus 5'CAGA3' sequences (Figure S8B). Interestingly, performing the same position replacements in the right part of the *attI1* site did not hamper the recombination frequency (Figure S9B), indicating structural differences between the *attI* and genome sites. From these, we conclude that genome sites differ broadly, in terms of sequence and structure from both classical *attC* and *attI* recombination sites. We propose to name these new types of genomic integron sites "*attG*", where G stands for "genome".

A new way to generate array of cassettes

To determine if inserted cassettes in *attG* sites can attract new cassettes and reconstitute an *attI* like integron platform, we performed the same experiment as in Figure S9B, but we mimicked the insertion of three different cassettes in the *attG* site of *ybhO* hotspot before assessing the insertion frequency of cassettes in these synthetic genome sites. The tested cassettes were

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

VCA458_VCR, VCA462_VCR or *aadA7_attC*, depending on their 6 to 9 position motif that would reconstitute or not the 6 to 9 motif of the *ybhO* hotspot site (Figure S9A and S9C). Insertion frequency was above 10^{-4} when the 6 to 9 motif of the *ybhO* hotspot was maintained in the 5' part of the inserted cassette (Figure S9C), meaning that an array of cassettes could be constituted by multiple successive integration events if efficient *attG* sites are reconstituted.

attG sites are recombined as double-stranded forms and are single cleaved

Our results support the hypothesis that *attG* sites are bound and recombined by the integrase as a double-stranded form. Indeed, *attG* sites are equally frequent in the leading strand as in the lagging one, when the single-strand availability associated with the latter would favor folding of single-stranded structures (Figure 7E). Moreover, a secondary structure prediction software such as RNAfold program from ViennaRNA Package (23) (https://www.tbi.univie.ac.at/RNA/) did not reveal secondary structures in the tested *attG* sequence sites. To confirm the double-stranded nature of the *attG* sites, we cloned three hotspots in the donor plasmid of our conjugation assay, in both orientations, delivering either the top or the bottom strand as donor sites during conjugation, and we tested the ability of these *attG* sites to recombine in the $attC_{aadA7}$ receptor site present in the recipient plasmid (Figure 9A). If *attG* sites are recombined as a strand specific single-stranded form, a difference in the recombination rate would be expected while injecting the top or the bottom strand as shown for *attC* sites ((5), Figure 9B and 9C). In contrast, if *attG* are recombined as a double-stranded form and thus requires second strand synthesis to be effective, no difference of recombination frequency is expected whatever the injected strand as shown for the *attI* site ((5), Figure 9B and 9C). Obtained insertion frequencies varied between the tested hotspots but were similar regardless of their orientation (Figure 9A). These results confirm that *attG* sites are recombined as double stranded form, like for the *attI* sites. By performing PCR and sequencing of recombined products, we also confirmed that recombination occurs by a single cleavage of the double strand matrix. Indeed,

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

a double cleavage would lead to a heterogeneity of sequences at the insertion point as expected when the core sequences of *att* sites are different (Figure S10A). Here, we obtained a sequence homogeneity around the insertion site (Figure S10B). We also confirmed by sequence analysis and determination of the cassette insertion orientation that recombination takes place between the bottom strands of both the hotspots and *attC* receptor sites and at the expected recombination points (Figure S10C). Note that we sequenced more than 24 recombination products for each top and bottom injected strands and for all three tested hotspot sites (more than 144 sequences total, Figure 9A), illustrating the robustness of our results.

We therefore conclude that during cassette insertion in the genome, the *attG* sites recombine as double-stranded forms and that only their bottom strands are cleaved. The synaptic complex conformation is therefore expected to be the same that for an *attC* × *attI* reaction and to generate, after the 1st strand exchange, an atypical Holliday junction (due to the single-stranded nature of the *attC* site) resolved by a replicative way (24) (Figure S10).

DISCUSSION

The rearrangement of cassettes is central to the integron system. Cassette mobility within the integron takes place by site-specific recombination between *attI* and *attC* sites. This mechanism facilitates acquisition of cassettes in the integron platform and their consequently expression with minimal disturbance to the remaining genome. This unique genetic system allows bacteria to evolve in response to pressure such as antibiotic use.

Up to now, all performed insertion assays revealed an almost insignificant (when quantified, no more than 10^{-6}) rate of cassette insertions at secondary sites in plasmids and genomes (10-12,25,26). These rare insertion events have always been considered anecdotal and relegated to secondary status in the integron functioning. However, several natural cases of likely cassette insertion outside the integron were described. The single and complete *aadB* cassette has been found inserted at a secondary site into the IncQ plasmid RSF1010 (27) and into the pRAY

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

plasmid (28,29). In both plasmids, cassettes were found inserted just downstream of a known promoter ensuring the expression of the *aadB* antibiotic resistance gene. Here, by performing an extensive bioinformatics analysis of available sequenced bacterial genomes, we found many isolated cassettes inserted in bacterial genomes. We called these cassettes, SALIN, for Single *attC* sites lacking integron-integrase and we took this observation as a starting point to study and identify a so far hidden cassette dissemination route in bacterial genomes. Using an assay mimicking the natural conditions in which the acquisition of cassettes occurs through horizontal gene transfer, we demonstrated that integron cassettes can disseminate outside the integron platform, at several positions in host genomes and at a high frequency, *i.e.,* close to the frequency obtained using canonical *attI and attC* sites. We unveiled that integrase insertion sites have a very small 5'GWT3' consensus sequence meaning that the cassette insertional landscape can be very large. As example, in the 4,641,652 bps of the MG1655 *E. coli* genome, this represents 338,348 theoretically targetable sites. We used insertion hotspots to determine the characteristics of the genome sites and demonstrated that these sites are structurally different from both *attC* and *attI* sites, *i.e.,* with a cleavage point located in their central part, getting them closer to classical double-stranded core recombination sites such as *dif* sites (30). We therefore called these new sites, *attG* sites. As these *attG* sites are recombined as a double-stranded form, we suggest that *attI* sites could have derived from the sequence of one of these hotspots through coevolution with an integrase. These results illustrate the very high flexibility of the integron integrases, already known for its ability to recombine sites in single-stranded or double-stranded forms (5), to perform single or double cleavage (31) and now, to catalyze site-specific or "random" recombination. Interestingly, another site-specific recombinase, the lambda (λ) integrase, which ensure λ phage genome integration at the *attB* specific site in the host chromosome, was also described as being able to catalyze phage genome integration into secondary sites. The found consensus insertion sequence is much larger than that of integron

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

integrases and actually very similar to the *attB* recombination site restraining the secondary insertion sites to a small number of locations in bacterial chromosome (32,33). Therefore, genome insertions mediated by the λ integrase would impact very poorly the genome evolution compared to the integron integrase which, with its extensive insertion landscape, could have a driving role in evolution. First, by disrupting genes, random cassette insertions could cause phenotypic changes and possibly evolve relevant traits in bacteria. Second, once inserted, certain cassettes can be imprecisely excised thus inducing genome modifications. Third, cassette insertion can represent a gain of function for the bacteria since genome inserted cassettes can be expressed if inserted in a vicinity of a bacterial promoter. Knowing furthermore that a promoter can be easily created from a random sequence with the acquisition of only 1 or 2 mutations (34), this would increase the probability of the promoterless cassettes to be expressed once inserted in bacterial genomes. These cassettes could also be domesticated if mutations inactivating their *attC* sites occur or even if their *attC* sites are deleted. Indeed, we have previously demonstrated that *attC* sites, due to their high propensity to fold into secondary structures, can be easily lost by replication slippage events (35,36).

In conclusion, we demonstrated here the existence of a new route of cassette recombination largely expanding the role of integrons in dissemination of adaptive functions such as antibiotic resistance. Under conditions where incoming conjugative plasmids carrying integrons (MIs) would be unstable, this new way could enable bacteria to “save” cassettes in their genomes before the loss of the plasmid. Beyond of that, these results show that the integron system could could represent a general mechanism for the genomic diversification driving bacterial evolution

SUPPLEMENTAL INFORMATION

Table S1: Bacterial strains used in this study

Table S2: Plasmids used in this study

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table S3: Primers used in this study

Figure S1: Sequences and structures of *attC* and *attI* sites used in this study

Figure S2: Random PCR approach used to determine the genome insertion sites

Figure S3: Cassette insertion events in genome in presence of an integron

Figure S4: Cassette recombination events mediated by IntI2 and IntI3 integrases

Figure S5: Representation of the imprecise excisions of genome-inserted cassettes

Figure S6: Description of the library construction for Deep sequencing

Figure S7: Computational analysis of Deep sequencing data

Figure S8: Digital PCR analysis of the *oriC* copy number relative to *terC* in *E. coli* receptor strain during a conjugation mimicking assay

Figure S9: Analysis of the *attG* insertion site motifs

Figure S10: Determination of the *attG* recombination nature

FUNDING

This work was supported by the Institut Pasteur, the Centre National de la Recherche Scientifique (CNRS-UMR 3525), the Fondation pour la Recherche Médicale (FRM Grant No. EQU202103012569), ANR Chromintevol (ANR-21-CE12-0002-01), and by the French Government's Investissement d'Avenir program Laboratoire d'Excellence 'Integrative Biology of Emerging Infectious Diseases' [ANR-10-LABX-62-IBEID].

ACKNOWLEDGEMENTS

We would like to thank Gaspard Macaux for its experimental help. We also thank Jason Bland, a native English speaker, for helpful reading of the manuscript and all the lab members for helpful discussion.

AUTHOR CONTRIBUTIONS

C.L and D.M designed the research. C.L, E.R, C.V, B.D, V.P, F.L and T.N performed the experiments. G.M and F. L performed the computational analysis of Deep sequencing data. J.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

C and E.P.C.R performed the bioinformatics genomics analysis. C.L and G. M wrote the draft of the manuscript. All authors read, amended the manuscript, and approved its final version.

CONFLICT OF INTEREST

None declared

REFERENCES

1. Stokes, H.W. and Hall, R.M. (1989) A novel family of potentially mobile DNA elements encoding site-specific gene-integration functions: integrons. *Molecular Microbiology*, **3**, 1669-1683.
2. Escudero, J.A., Loot, C., Nivina, A. and Mazel, D. (2015) The Integron: Adaptation On Demand. *Microbiology spectrum*, **3**, MDNA3-0019-2014.
3. Cury, J., Jove, T., Touchon, M., Neron, B. and Rocha, E.P. (2016) Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucleic Acids Res*, **44**, 4539-4550.
4. Neron, B., Littner, E., Haudiquet, M., Perrin, A., Cury, J. and Rocha, E.P.C. (2022) IntegronFinder 2.0: Identification and Analysis of Integrons across Bacteria, with a Focus on Antibiotic Resistance in Klebsiella. *Microorganisms*, **10**.
5. Bouvier, M., Demarre, G. and Mazel, D. (2005) Integron cassette insertion: a recombination process involving a folded single strand substrate. *Embo J*, **24**, 4356-4367.
6. Guerin, E., Cambray, G., Sanchez-Alberola, N., Campoy, S., Erill, I., Da Re, S., Gonzalez-Zorn, B., Barbe, J., Ploy, M.C. and Mazel, D. (2009) The SOS response controls integron recombination. *Science*, **324**, 1034.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

7. Baharoglu, Z. and Mazel, D. (2011) Vibrio cholerae triggers SOS and mutagenesis in response to a wide range of antibiotics: a route towards multiresistance. *Antimicrob Agents Chemother*, **55**, 2438-2441.

8. Richard, E., Darracq, B., Loot, C. and Mazel, D. (2022) Unbridled Integrons: A Matter of Host Factors. *Cells*, **11**.

9. Buongermino Pereira, M., Osterlund, T., Eriksson, K.M., Backhaus, T., Axelson-Fisk, M. and Kristiansson, E. (2020) A comprehensive survey of integron-associated genes present in metagenomes. *BMC Genomics*, **21**, 495.

10. Francia, M.V., de la Cruz, F. and Garcia Lobo, J.M. (1993) Secondary-sites for integration mediated by the Tn21 integrase. *Molecular Microbiology*, **10**, 823-828.

11. Francia, M.V. and Garcia Lobo, J.M. (1996) Gene integration in the Escherichia coli chromosome mediated by Tn21 integrase (Int21). *J Bacteriol*, **178**, 894-898.

12. Recchia, G.D., Stokes, H.W. and Hall, R.M. (1994) Characterisation of specific and secondary recombination sites recognised by the integron DNA integrase. *Nucleic Acids Res*, **22**, 2071-2078.

13. Vit, C., Richard, E., Fournes, F., Whiteway, C., Eyer, X., Lapaillerie, D., Parissi, V., Mazel, D. and Loot, C. (2021) Cassette recruitment in the chromosomal Integron of Vibrio cholerae. *Nucleic Acids Res*, **49**, 5654-5670.

14. Le Roux, F., Binesse, J., Saulnier, D. and Mazel, D. (2007) Construction of a Vibrio splendidus mutant lacking the metalloprotease gene vsm by use of a novel counterselectable suicide vector. *Appl Environ Microbiol*, **73**, 777-784.

15. Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods*, **9**, 357-359.

16. Wagih, O. (2017) ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics*, **33**, 3645-3647.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

17. Ghaly, T.M., Tetu, S.G. and Gillings, M.R. (2021) Predicting the taxonomic and environmental sources of integron gene cassettes using structural and sequence homology of attC sites. *Commun Biol*, **4**, 946.

18. Loot, C., Bikard, D., Rachlin, A. and Mazel, D. (2010) Cellular pathways controlling integron cassette site folding. *EMBO J*, **29**, 2623-2634.

19. Nivina, A., Escudero, J.A., Vit, C., Mazel, D. and Loot, C. (2016) Efficiency of integron cassette insertion in correct orientation is ensured by the interplay of the three unpaired features of attC recombination sites. *Nucleic Acids Res*, **44**, 7792-7803.

20. Mazel, D., Dychinco, B., Webb, V.A. and Davies, J. (1998) A distinctive class of integron in the *Vibrio cholerae* genome. *Science*, **280**, 605-608.

21. Betton, J.M. (2004) Cloning vectors for expression-PCR products. *Biotechniques*, **37**, 346-347.

22. Rousset, F., Cui, L., Siouve, E., Becavin, C., Depardieu, F. and Bikard, D. (2018) Genome-wide CRISPR-dCas9 screens in E. coli identify essential genes and phage host factors. *PLoS Genet*, **14**, e1007749.

23. Lorenz, R., Bernhart, S.H., Honer Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F. and Hofacker, I.L. (2011) ViennaRNA Package 2.0. *Algorithms for molecular biology : AMB*, **6**, 26.

24. Loot, C., Ducos-Galand, M., Escudero, J.A., Bouvier, M. and Mazel, D. (2012) Replicative resolution of integron cassette insertion. *Nucleic Acids Res*, **40**, 8361-8370.

25. Hansson, K., Skold, O. and Sundstrom, L. (1997) Non-palindromic attl sites of integrons are capable of site-specific recombination with one another and with secondary targets. *Molecular Microbiology*, **26**, 441-453.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

26. Célia Souque, J.A.E., R.Craig MacLean. (2022) Off-target integron activity leads to rapid plasmid compensatory evolution in response to antibiotic selection pressure. *BioRxiv*.

27. Recchia, G.D. and Hall, R.M. (1995) Plasmid evolution by acquisition of mobile gene cassettes: plasmid pIE723 contains the aadB gene cassette precisely inserted at a secondary site in the incQ plasmid RSF1010. *Mol Microbiol*, **15**, 179-187.

28. Segal, H. and Elisha, B.G. (1997) Identification and characterization of an aadB gene cassette at a secondary site in a plasmid from Acinetobacter. *FEMS Microbiol Lett*, **153**, 321-326.

29. Segal, H., Francia, M.V., Lobo, J.M. and Elisha, G. (1999) Reconstruction of an active integron recombination site after integration of a gene cassette at a secondary site. *Antimicrob Agents Chemother*, **43**, 2538-2541.

30. Crozat, E., Fournes, F., Cornet, F., Hallet, B. and Rousseau, P. (2014) Resolution of Multimeric Forms of Circular Plasmids and Chromosomes. *Microbiology spectrum*, **2**.

31. Escudero, J.A., Loot, C., Parissi, V., Nivina, A., Bouchier, C. and Mazel, D. (2016) Unmasking the ancestral activity of integron integrases reveals a smooth evolutionary transition during functional innovation. *Nature communications*, **7**, 10937.

32. Rutkai, E., Dorgai, L., Sirot, R., Yagil, E. and Weisberg, R.A. (2003) Analysis of insertion into secondary attachment sites by phage lambda and by int mutants with altered recombination specificity. *J Mol Biol*, **329**, 983-996.

33. Tanouchi, Y. and Covert, M.W. (2017) Combining Comprehensive Analysis of Off-Site Lambda Phage Integration with a CRISPR-Based Means of Characterizing Downstream Physiology. *mBio*, **8**.

34. Yona, A.H., Alm, E.J. and Gore, J. (2018) Random sequences rapidly evolve into de novo promoters. *Nature communications*, **9**, 1530.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

35. Loot, C., Parissi, V., Escudero, J.A., Amarir-Bouhram, J., Bikard, D. and Mazel, D. (2014) The integron integrase efficiently prevents the melting effect of Escherichia coli single-stranded DNA-binding protein on folded attC sites. *J Bacteriol*, **196**, 762-771.

36. Loot, C., Nivina, A., Cury, J., Escudero, J.A., Ducos-Galand, M., Bikard, D., Rocha, E.P. and Mazel, D. (2017) Differences in Integron Cassette Excision Dynamics Shape a Trade-Off between Evolvability and Genetic Capacitance. *mBio*, **8**.

FIGURE CAPTIONS

Figure 1: The integron system.

The integron system is composed of the integron platform (the integrase expressing gene, *intI*, the two promoters, P_C and P_{int} and the *attI* recombination site (red triangle)) and of the variable cassette array. The variable cassette array contains some cassettes represented by small colored arrows. Only the first cassettes of the array are expressed, and the subsequent ones can be seen as a low-cost cassette reservoir. Upon expression of the integrase, cassette shuffling can occur through cassette excision (*attC* × *attC*) and insertion of the excised cassettes in the first position in the array (*attI* × *attC*).

Figure 2: Distribution of integrons across bacteria using the RefSeq NCBI database.

(A) Number of bacterial genomes containing either a complete integron (integron), an integrase gene (In0), a Cluster of *attC* sites lacking integron-integrase (CALIN) or a Single *attC* site lacking integron-integrase (SALIN).

(B) Number of SALIN found per genome.

(C) Number of *attC* sites per CALIN.

Figure 3: Cassette insertion in *att* and *Escherichia coli* genome sites during conjugation assay.

(A) Experimental setup of the cassette insertion assay.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The pSW23T::*attC*$_{aadA7}$ suicide vector is delivered from the β2163 *E. coli* donor strain to the MG1655 *E. coli* recipient strain. The recipient strain contains an *att*-carrying plasmid and a vector expressing, or not, the IntI1 integrase. The *attC*$_{aadA7}$ site carried by the suicide vector is represented by a grey triangle and the *attI1* and *attC* receptor sites by respectively red and green triangles. Phenotypic resistances to chloramphenicol (Cm^R), kanamycin (Km^R) and spectinomycin (Sp^R) are represented by grey rectangles and origin of replication by grey circles. The inducible P$_{bad}$ promoter is represented by a back arrow. The delivery of the suicide vector occurs by conjugation. As pSW23T cannot replicate in the *E. coli* recipient strain, recombinant clones can be selected on appropriate Cm containing plates to evaluate the recombination frequency (see Materials and Methods).

(B) Frequency of cassette insertion into the *att* or *Escherichia coli* genome sites

Total recombination frequency (left panel) corresponds to the frequency of all Cm resistant recombinant clones. *att* (middle panel) and *genome* (right panel) recombination frequencies were obtained by adjusting the total recombination frequencies of the left panel to the proportion obtained by PCR reactions (see Materials and Methods). pBAD43::IntI1 and pBAD43 means that recipient strains contain the pBAD43 integrase expressing vector or the empty pBAD43 vector, respectively. Asterisk (*) indicates the recombination frequency was below detection level, indicated by the bar height (limit of detection). The recipient plasmids are indicated in the axis-x legends (see Table S2 for the description). Values represent the mean of at least three independent experiments and error bars represent mean absolute error.

Figure 4: Cassette insertion in *Escherichia coli* and *Vibrio cholerae* genome sites.

(A) Experimental setup of the cassette insertion assay.

As in Figure 3A, except that the donor plasmids carry different *att* sites (as indicated in B) and that the receptor strains (MG1655 *E. coli* or N16961 *Vibrio cholerae*) lack any recipient plasmid and express, or not, the IntI1, IntI2 and IntI3 integrases (as indicated in B).

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(B) Frequency of cassette insertion into the *Escherichia coli* and *Vibrio cholerae* genome sites. The donor plasmids are indicated in the axis-x legends (see table S2). ts, means that top strand is injected during conjugation. The recipient strains and the expressed integrases are indicated at the top of the bars. For more details on the calculation of the recombination frequency see the legend of the figure 3B.

Figure 5: Cassette expression of inserted cassettes in *Escherichia coli* genome sites.

(A) Experimental setup of the cassette insertion and expression assay.

As in Figure 3A, except that the donor plasmids contain a promoterless *kamamycin* gene (orange rectangle) associated or not with a Ribosome Binding Site (RBS1 or RBS2) and that the receptor strain (MG1655 *E. coli*) lacks any recipient plasmid and express, or not, the IntI1 integrase (as indicated in B). All recombinants are selected on Cm containing plates and recombinants clones expressing the *kanamycin* gene (orange rectangle) are selected on Cm and Km containing plates. A bacterial promoter is shown by a black arrow.

(B) Frequency of cassette insertion and expression into the *Escherichia coli* genome.

The donor plasmids are indicated in the axis-x legends (see table S2). For more details on the calculation of the recombination frequency see the legend of the figure 3B.

(C) Schemes of two inserted and expressed cassettes

Cassettes are shown inserted between the promoter (black arrow) and the CDS of the *ompC* and *rseA E. coli* genes.

Figure 6: Cassette excision of inserted cassette in *Escherichia coli* genome sites during conjugation assay.

(A) Experimental setup of the cassette excision assay

As in Figure 3A, except that the donor plasmid contains the *ccdB* toxic gene (pink rectangle) under the control of the P_{bad} promoter (pink arrow), induced by arabinose (Ara) and repressed by glucose (Glc), and that the receptor strain (MG1655 *recA E. coli*) lacks any recipient plasmid

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and contain a thermosensitive (ts) plasmid expressing, or not, the IntI1 integrase and encoding for the kanamycin resistance (Km^R). The integrase gene is under the control of the P_{tet} promoter (green arrow). 24 recombinants clones were randomly selected and placed at 42°C to remove the thermosensitive plasmid and re-transformed with plasmid expressing or not the integrase. Recombinants clones corresponding to excised cassettes are selected on arabinose (Ara) containing plates.

(B) Frequency of cassette excision from the *Escherichia coli* genome sites

The excision frequencies (obtained for clone 7, 8 and 9) are shown and correspond to the frequency of recombinant clones growing in presence of arabinose. Precise (light grey bars) and imprecise (dark grey bars) excision frequencies are indicated for each clone.

Figure 7: Computational analysis of Deep sequencing data.

A) Insertion site usage all along the *E. coli* genome without read duplicate removal (n = 3,205,043) (see material and methods for details). Insertion in the forward or reverse strand determines the orientation of the inserted cassettes. bp, base pair. Ori and ter, region of origin of replication and of termination respectively.

B) Insertion site usage all along the *E. coli* genome with read duplicate removal (n = 361,464). All the other panels from this figure derive from read duplicate removal. Obs, observed insertions; Random, random insertions using the G[AT]T consensus motif of insertion (see D);

C) Number of insertion sites all along the *E. coli* genome using a sliding window of 200 kb sliding every 100 bases.

D) Consensus sequence of insertion sites. A total of 20 bases around the cleavage point of each read was aligned to generate the motif. The cleavage occurs between the -1 and 1 bases. Bits refers to the information content.

E) Proportion of insertion sites according to the strand polarity (Forward or Reverse) and according to replication orientation. Leading and lagging mean that the read corresponds to the

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

neo-synthetized leading and lagging strand during replication, respectively. The proportions are relative to the maximal proportion set to 1.

F) Proportion of insertion sites all along the *E. coli* genome according to replication orientation using a sliding window of 200 kb sliding every 100 bases.

G) Proportion of insertion sites according to Coding sequence (CDS) in the genome, either inside/outside CDS or dispensable/essential CDS.

The proportions are relative to the proportion of CDS regions in the genome (8%) and relative to the maximal proportion set to 1.

H) Relative position of insertion sites inside dispensable (red) or essential (blue) CDS, between 0 (start codon) and 1 (stop codon). Box, inside vertical bar, whisker and diamond indicate quartiles, median, 1.5 x Inter Quartile Range and mean, respectively. Numbers on the right side correspond to median values. Each dot represents a single insertion site in purple or gold, depending on the opposite or same plasmid orientation versus CDS orientation respectively.

I) As in H but for the distance of insertion site from the closest Transcription Start Site (TSS) in base pairs (bp).

Figure 8: Cassette insertion in *hotspot* sites during conjugation assay.

(A) Experimental setup of the cassette insertion assay.

As in Figure 3A except that the *att* sites of the recipient plasmids are replaced by several hotspot (HS) sites (black triangle) and that the recipient plasmid is a high copy number *ori*pMB1 plasmid.

(B) Frequency of cassette insertion into the *ybhO*, *aslB*, *ilvD*, *pyrE*, *metC*, *yjhH* hotspot sites, into the *abgA* median site and into the *ygcE* unique site.

The receptor plasmids are indicated in the axis-x legends (see table S2). HS, MS and US respectively mean Hotspot, Median and Unique Sites. The number of insertions that we previously obtained in each hotspot site (Figure 7A) are indicated in red at the top of the bars.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

For more details on the calculation of the recombination frequency see the legend of the figure 3B.

(C) Frequency of cassette insertion into truncated *ybhO* hotspot sites.

The receptor plasmids are indicated in the axis-x legends (see table S2). The 2 numbers represent the number of base pairs kept on each 5' and 3' side of the cutting position in the *ybhO* hotspot site (see Figure 7D). For more details on the calculation of the recombination frequency see the legend of the figure 3B.

(D) Frequency of cassette insertion into truncated *aslB* hotspot sites.

The legend is the same as in C except that it refers to the *aslB* site.

Figure 9: Testing the *hotspot* sites as donor sites.

(A) Experimental setup and frequency of the cassette insertion events.

As in Figure 3A, except that the donor plasmids carry different *hotspot* sites (black triangle) as indicated in the graph (right panel) and that the $attC_{aadA7}$ site (green triangle) is used in the recipient plasmid. Depending on the orientation of the *hotspot* site, either the bottom (bs) or the top strand (ts) is delivered.

(B) Scheme of expected frequencies injecting the bottom or the top strands in function of the active recombination forms.

Bottom strand (bs, green line) or top strand (ts, red line) are injected. If the recombination occurs by a single-stranded bottom form, a strand synthesis (dotted line) is necessary only when the top strand is injected and not the bottom strand. The frequency of recombination is therefore expected higher injecting the bottom strand than the top one (bs > ts). If recombination occurs by a double-stranded form, a strand synthesis (dotted line) is necessary injecting either of the two strands. The frequency of recombination is expected to be the same injecting the bottom strand or the top one (bs = ts).

(C) Experimental setup and frequency of the cassette insertion events.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.11.507474; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

As in Figure 3A except that the donor plasmids contain *attI1* or $attC_{aadA7}$ sites and that the recipient strain contains the $attC_{aadA7}$ site when the donor used site is *attI1*, or the *attI1* site when the donor used site is $attC_{aadA7}$. Depending on the orientation of the *att* sites, either the bottom or the top strands are delivered.

Figure 10: The new route of the integron cassette dissemination

Upon expression of the integrase, cassette shuffling inside the integron can occur through cassette excision (*attC* × *attC*) and insertion of the excised cassettes in the first position in the array (*attI* × *attC*). Cassettes can also be inserted in bacterial genomes through *attC* × *attG* recombination.

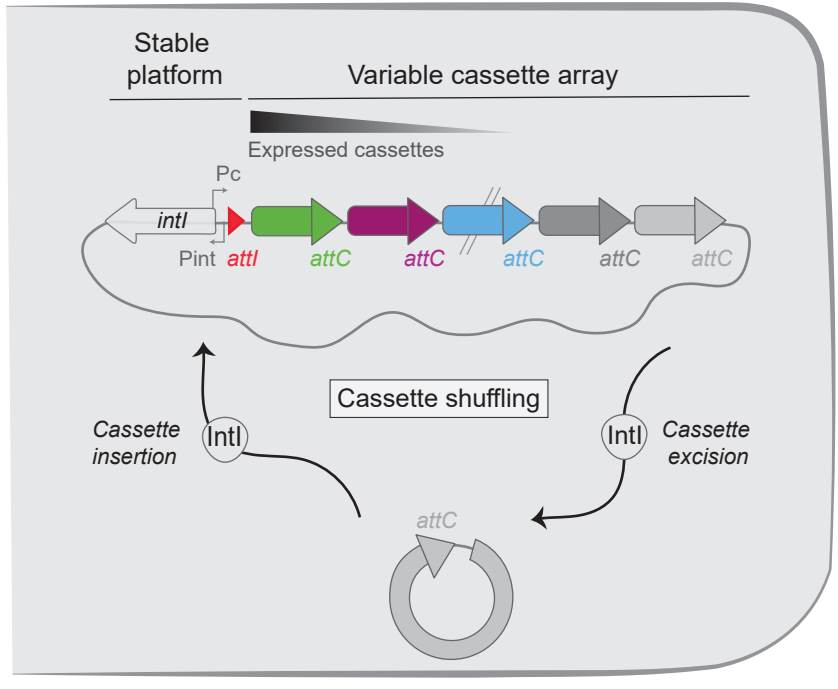

Stable platform
Variable cassette array
Expressed cassettes
Pc
intI
Pint
attI
attC
attC
attC
attC
attC
Cassette shuffling
Cassette insertion
IntI
IntI
Cassette excision
attC

Figure 1

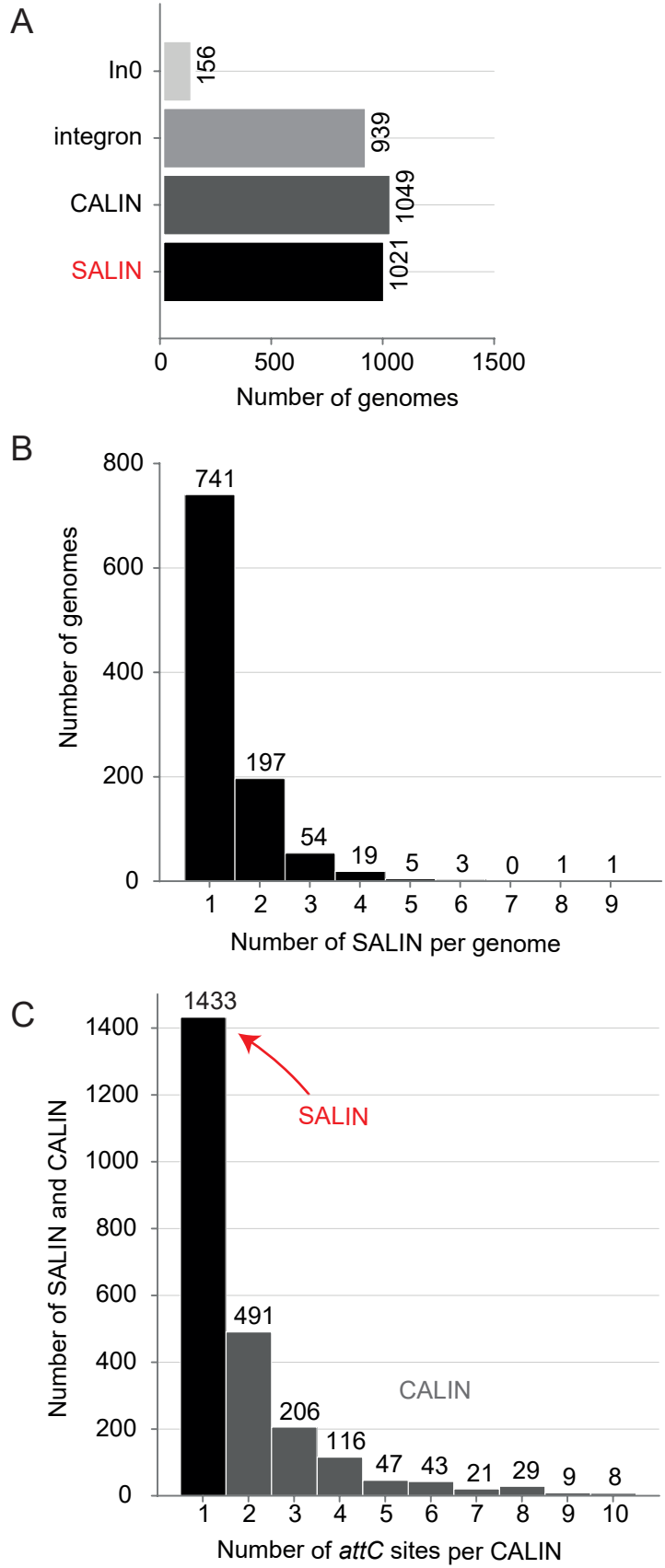

A
In0
integron
CALIN
SALIN
156
939
1049
1021
0
500
1000
1500
Number of genomes
B
800
600
400
200
0
741
197
54
19
5
3
0
1
1
1
2
3
4
5
6
7
8
9
Number of genomes
Number of SALIN per genome
C
1433
1400
1200
1000
800
600
400
200
0
SALIN
491
206
116
47
43
21
29
9
8
CALIN
1
2
3
4
5
6
7
8
9
10
Number of SALIN and CALIN
Number of attC sites per CALIN

Figure 2

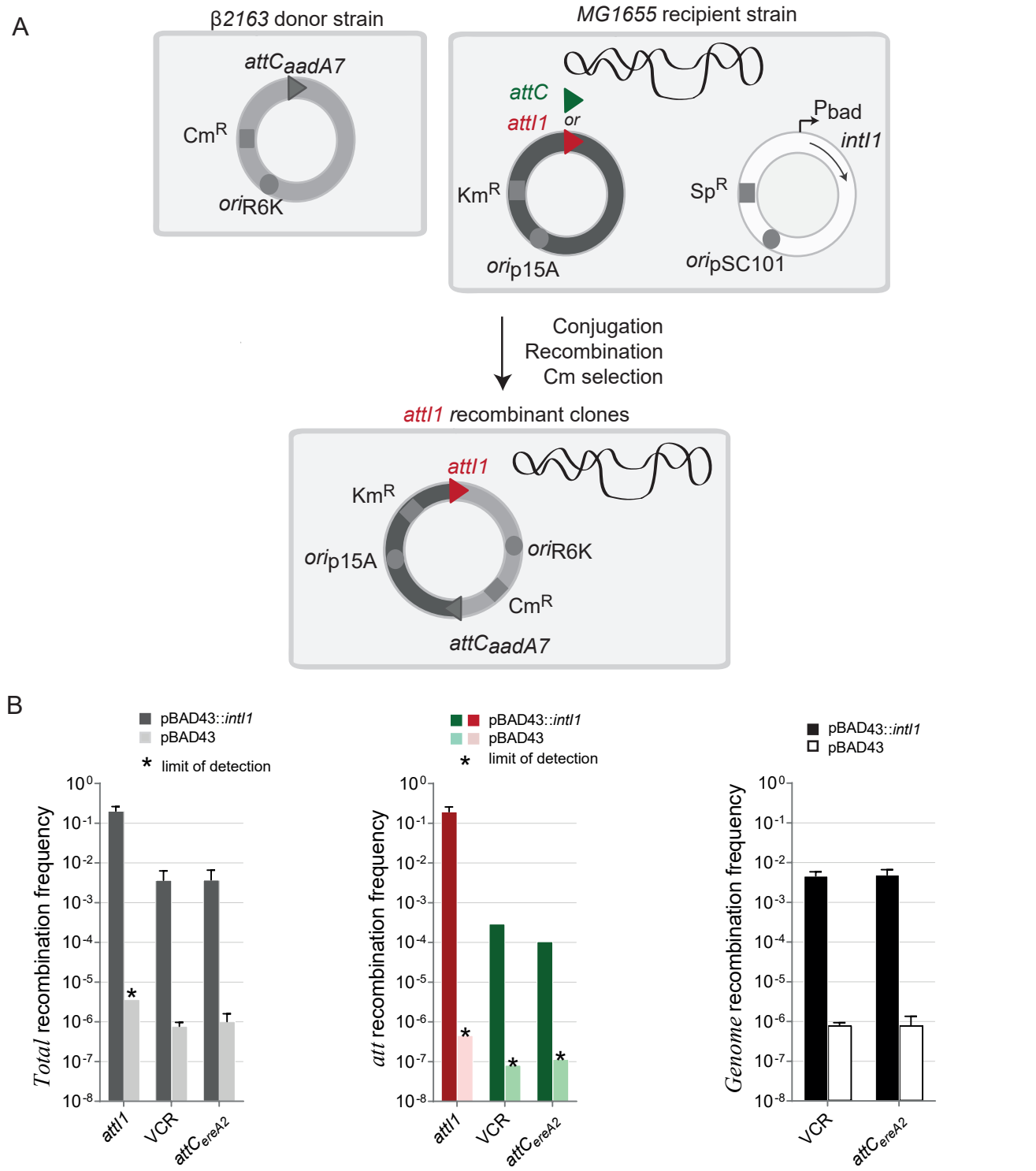
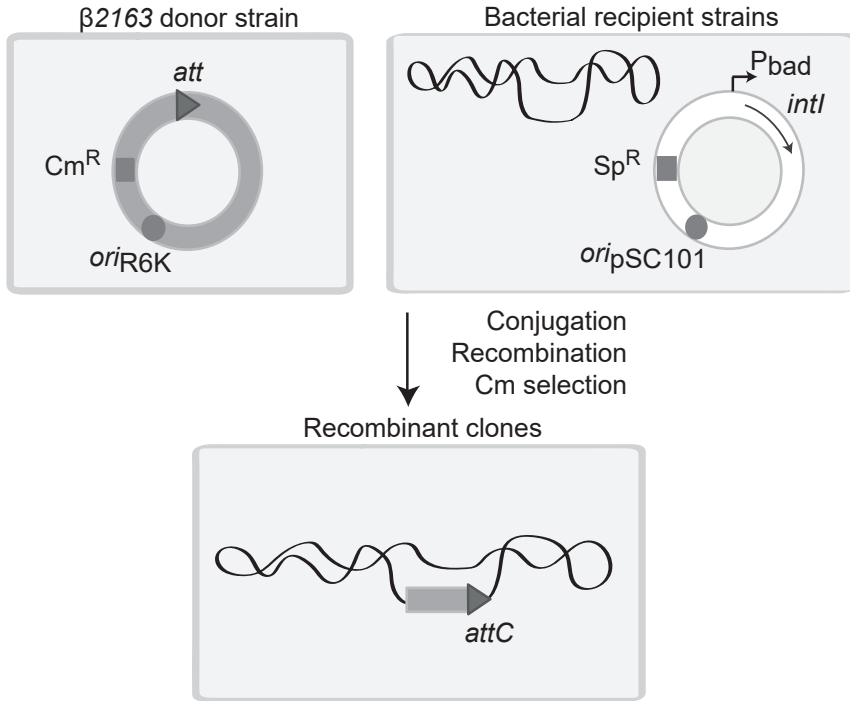

A
β2163 donor strain
attC aadA7
Cm R
ori R6K
MG1655 recipient strain
attC
attI1
or
Km R
ori p15A
Pbad
intI1
Sp R
ori pSC101
Conjugation
Recombination
Cm selection
attI1 recombinant clones
attI1
Km R
ori p15A
ori R6K
Cm R
attC aadA7
B
pBAD43::intI1
pBAD43
* limit of detection
Total recombination frequency
att recombination frequency
Genome recombination frequency
attI1
VCR
attC ereA2

Figure 3

A


β2163 donor strain
att
Cm^R
ori R6K
Bacterial recipient strains
Pbad
intI
Sp^R
ori pSC101
Conjugation
Recombination
Cm selection
Recombinant clones
attC

B

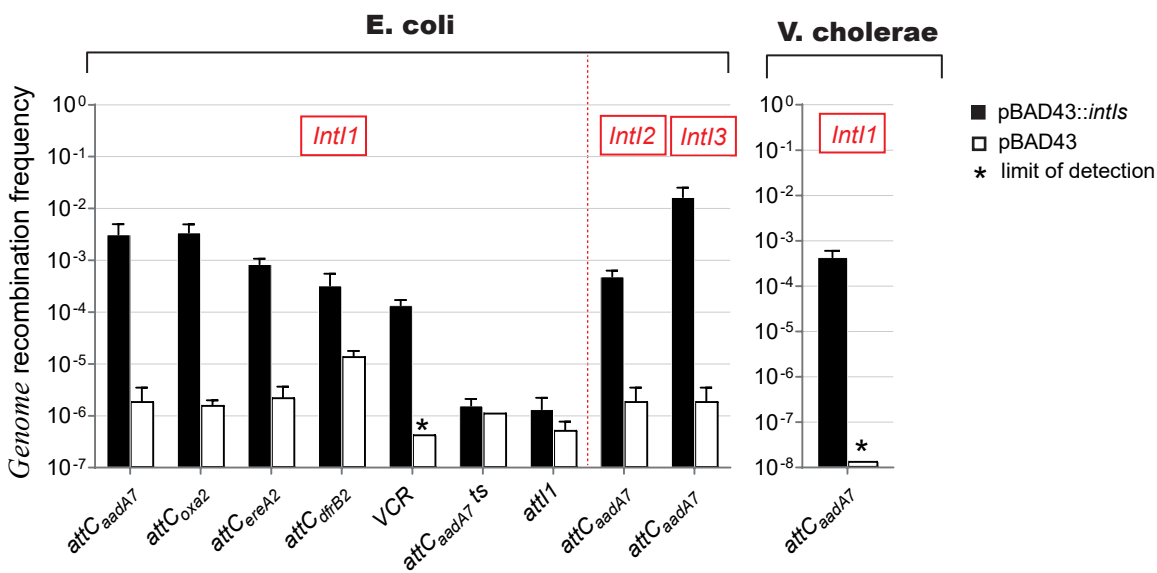

E. coli
V. cholerae
Genome recombination frequency
IntI1
IntI2
IntI3
IntI1
pBAD43::intIs
pBAD43
* limit of detection
attC aadA7
attC oxa2
attC ereA2
attC dfrB2
VCR
attC aadA7 ts
attI1
attC aadA7
attC aadA7
attC aadA7

Figure 4

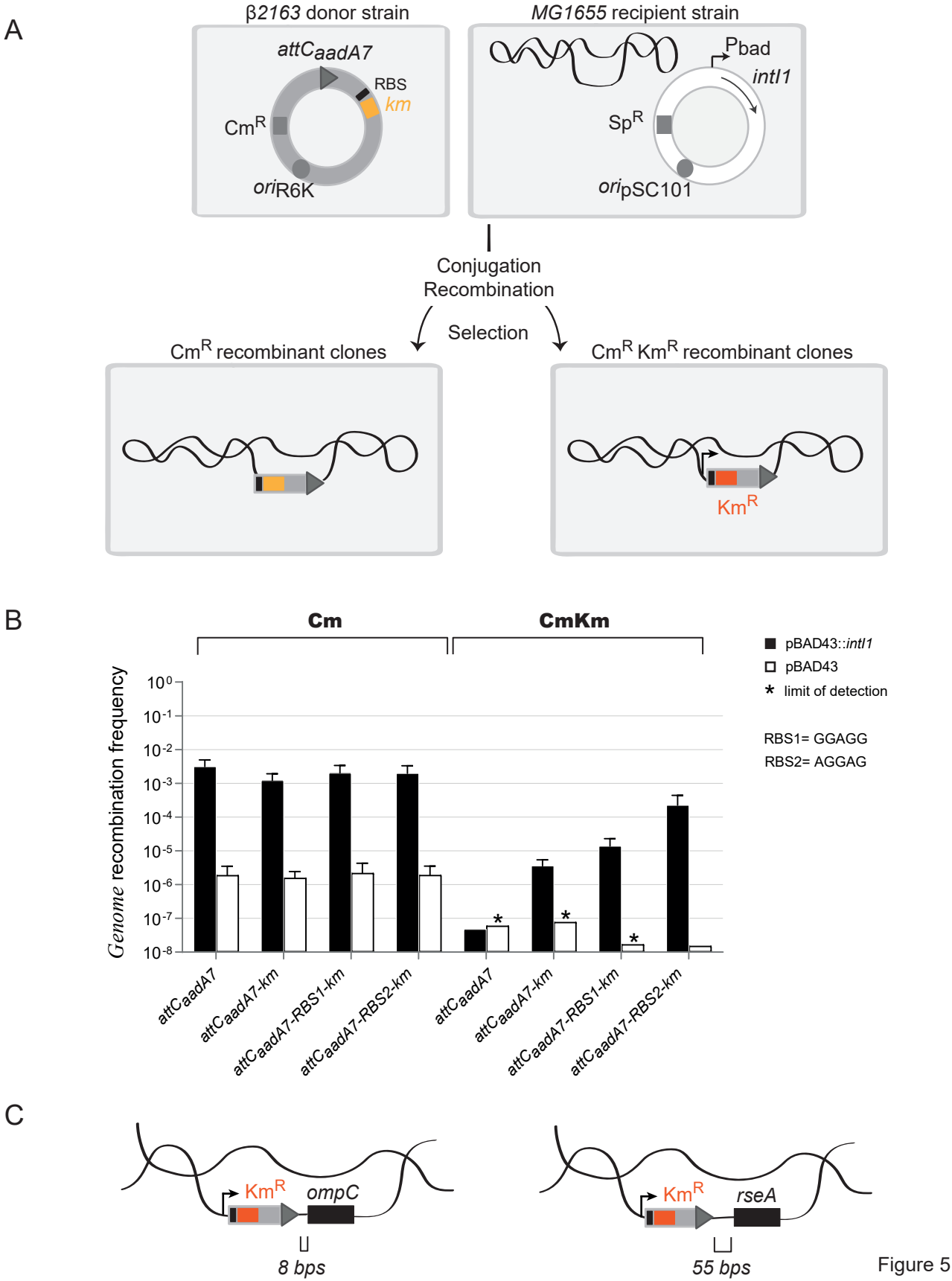
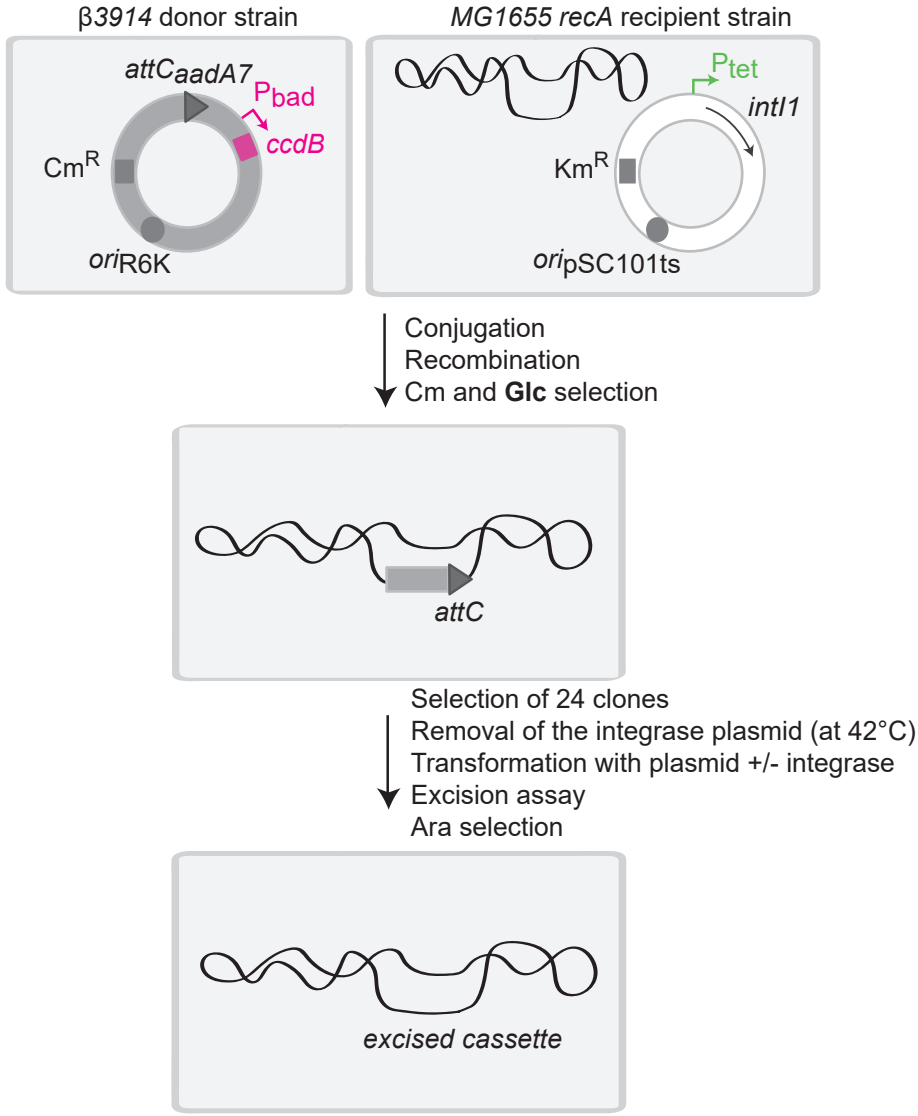

A
β2163 donor strain
attCaadA7
RBS
km
CmR
oriR6K
MG1655 recipient strain
Pbad
intI1
SpR
oripSC101
Conjugation
Recombination
Selection
CmR recombinant clones
CmR KmR recombinant clones
KmR
B
Cm
CmKm
pBAD43::intI1
pBAD43
* limit of detection
RBS1= GGAGG
RBS2= AGGAG
Genome recombination frequency
10^0
10^-1
10^-2
10^-3
10^-4
10^-5
10^-6
10^-7
10^-8
attCaadA7
attCaadA7-km
attCaadA7-RBS1-km
attCaadA7-RBS2-km
attCaadA7
attCaadA7-km
attCaadA7-RBS1-km
attCaadA7-RBS2-km
C
KmR
ompC
8 bps
KmR
rseA
55 bps

Figure 5

A


β3914 donor strain
MG1655 recA recipient strain
attC aadA7
Pbad
ccdB
CmR
oriR6K
Ptet
intI1
KmR
oripSC101ts
Conjugation
Recombination
Cm and Glc selection
attC
Selection of 24 clones
Removal of the integrase plasmid (at 42°C)
Transformation with plasmid +/- integrase
Excision assay
Ara selection
excised cassette

B

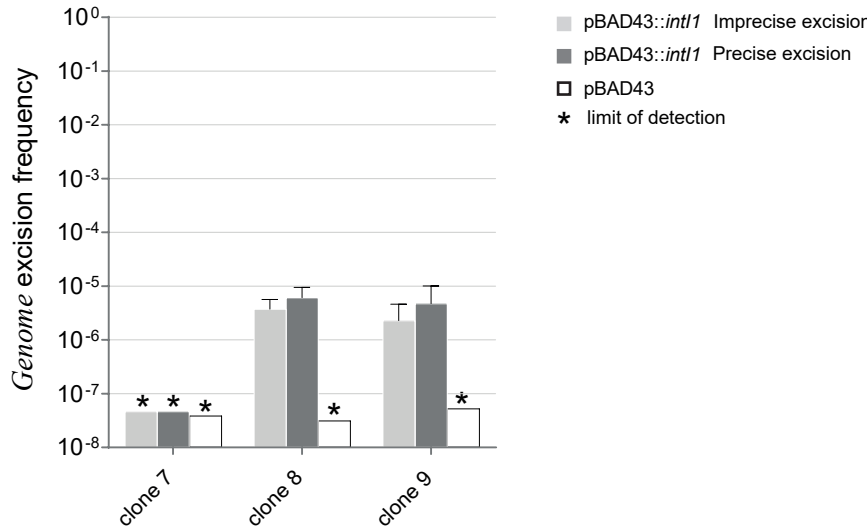

Genome excision frequency
10^0
10^-1
10^-2
10^-3
10^-4
10^-5
10^-6
10^-7
10^-8
clone 7
clone 8
clone 9
pBAD43::intI1 Imprecise excision
pBAD43::intI1 Precise excision
pBAD43
* limit of detection

Figure 6

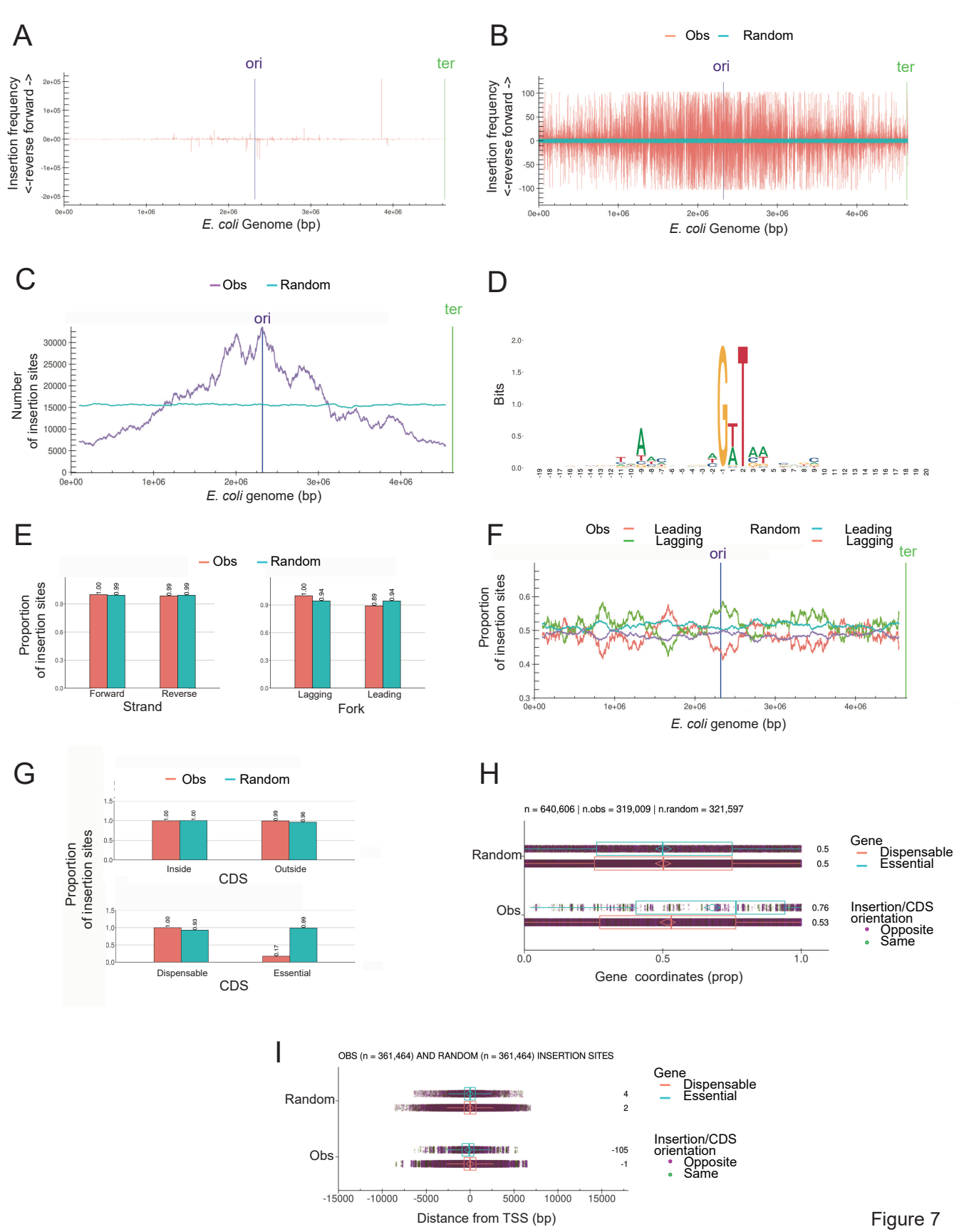

A
ori
ter
Insertion frequency
<-reverse forward ->
E. coli Genome (bp)
B
Obs
Random
ori
ter
Insertion frequency
<-reverse forward ->
E. coli Genome (bp)
C
Obs
Random
ori
ter
Number
of insertion sites
E. coli genome (bp)
D
Bits
E
Obs
Random
Proportion
of insertion sites
Forward
Reverse
Strand
Lagging
Leading
Fork
F
Obs
Leading
Lagging
Random
Leading
Lagging
ori
ter
Proportion
of insertion sites
E. coli genome (bp)
G
Obs
Random
Proportion
of insertion sites
Inside
Outside
CDS
Dispensable
Essential
CDS
H
n = 640,606 | n.obs = 319,009 | n.random = 321,597
Random
Obs
Gene
Dispensable
Essential
Insertion/CDS
orientation
Opposite
Same
Gene coordinates (prop)
I
OBS (n = 361,464) AND RANDOM (n = 361,464) INSERTION SITES
Random
Obs
Gene
Dispensable
Essential
Insertion/CDS
orientation
Opposite
Same
Distance from TSS (bp)

Figure 7

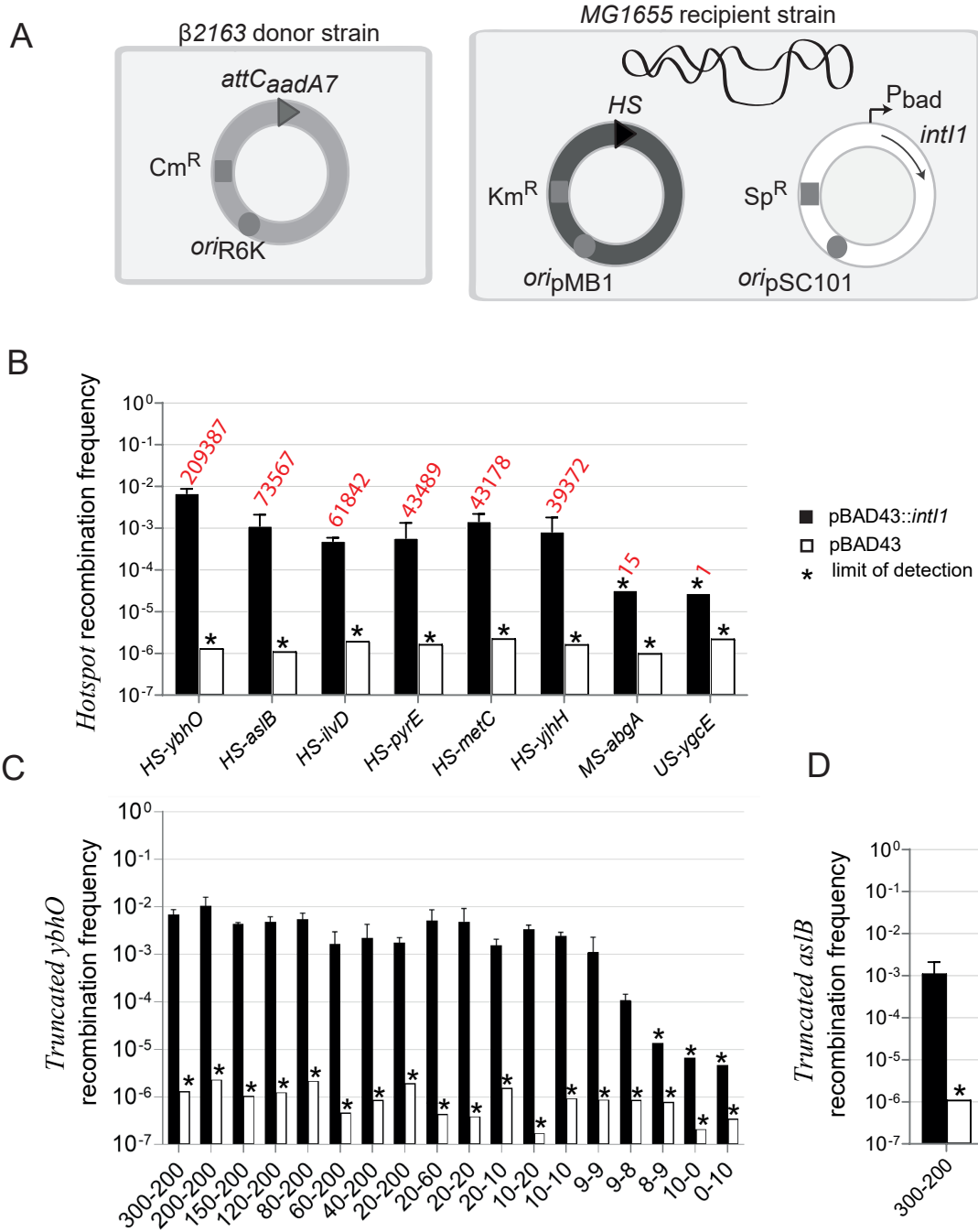

A
β2163 donor strain
attC aadA7
Cm R
ori R6K
MG1655 recipient strain
HS
Km R
ori pMB1
Pbad
intI1
Sp R
ori pSC101
B
Hotspot recombination frequency
209387
73567
61842
43489
43178
39372
15
1
HS-ybhO
HS-aslB
HS-ilvD
HS-pyrE
HS-metC
HS-yjhH
MS-abgA
US-ygcE
pBAD43::intI1
pBAD43
limit of detection
C
Truncated ybhO recombination frequency
300-200
200-200
150-200
120-200
80-200
60-200
40-200
20-200
20-60
20-20
20-10
10-20
10-10
9-9
9-8
8-9
10-0
0-10
D
Truncated aslB recombination frequency
300-200
9-9

Figure 8

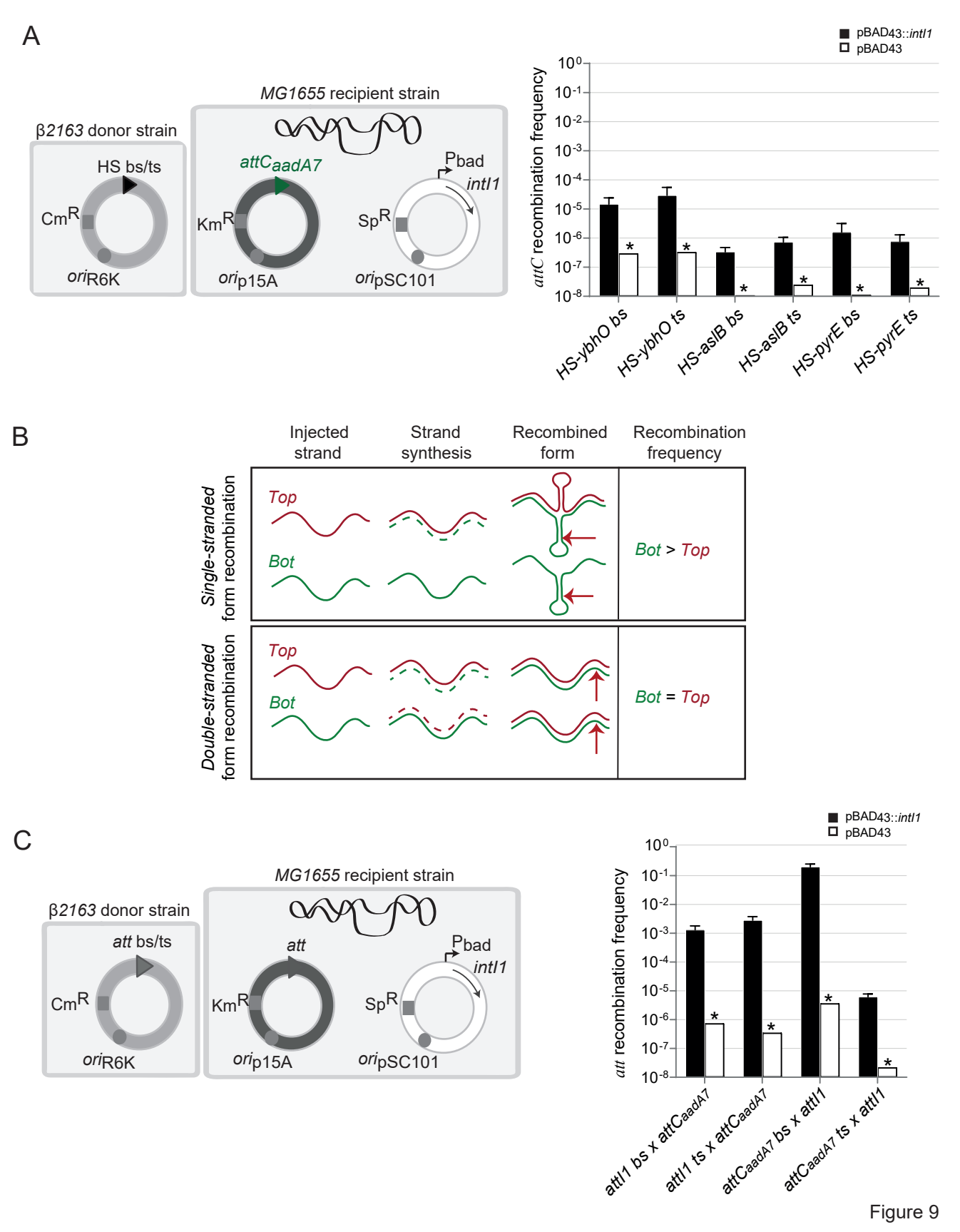

A
β2163 donor strain
HS bs/ts
CmR
oriR6K
MG1655 recipient strain
attCaadA7
KmR
orip15A
Pbad
intI1
SpR
oripSC101
pBAD43::intI1
pBAD43
attC recombination frequency
HS-ybhO bs
HS-ybhO ts
HS-aslB bs
HS-aslB ts
HS-pyrE bs
HS-pyrE ts
B
Injected strand
Strand synthesis
Recombined form
Recombination frequency
Single-stranded form recombination
Top
Bot
Bot > Top
Double-stranded form recombination
Bot = Top
C
β2163 donor strain
att bs/ts
MG1655 recipient strain
att
att recombination frequency
intI1 bs x attCaadA7
intI1 ts x attCaadA7
attCaadA7 bs x intI1
attCaadA7 ts x intI1

Figure 9

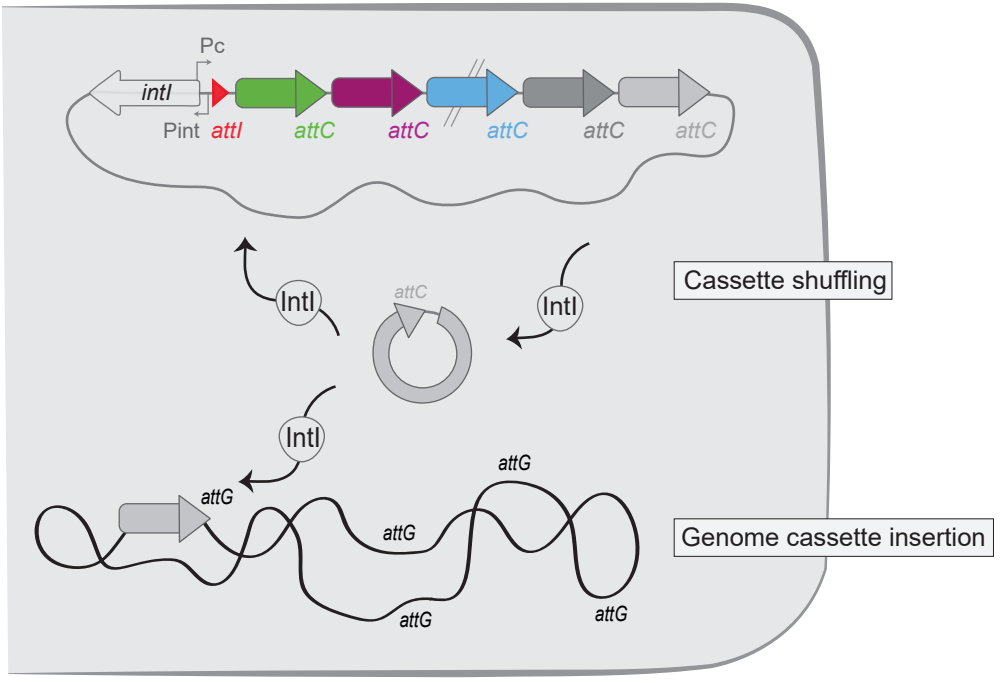

Pc
intI
Pint
attI
attC
attC
attC
attC
attC
IntI
attC
IntI
Cassette shuffling
IntI
attG
attG
attG
attG
attG
Genome cassette insertion

Figure 10