Integron cassettes commonly integrate into bacterial genomes via widespread non-classical attG sites

Céline Loot, Gael A Millot, Egill Richard, Eloi Littner, Claire Vit, Frédéric Lemoine, Bertrand Néron, Jean Cury, Baptiste Darracq, Théophile Niault, et al.

Integron cassettes commonly integrate into bacterial genomes via widespread non-classical *attG* sites

Céline Loot[1#], Gael A Millot[2], Egill Richard[1,3], Eloi Littner[3,4,5], Claire Vit[1,3], Frédéric Lemoine[2], Bertrand Néron[2], Jean Cury[6], Baptiste Darracq[1,3], Théophile Niault[1,3], Delphine Lapaillerie[7,8], Vincent Parissi[7,8], Eduardo PC Rocha[5] and Didier Mazel[1]

[1]Institut Pasteur, Université Paris Cité, CNRS UMR 3525, Unité Plasticité du Génome Bactérien, F-75015 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]DGA CBRN Defence, 91710 Vert-le-Petit, France.
[5]Institut Pasteur, Université Paris Cité, CNRS UMR 3525, Microbial Evolutionary Genomics, F-75015 Paris, France.
[6]Université Paris-Saclay, Inria, Laboratoire de Recherche en Informatique, CNRS UMR 8623, Orsay, France.
[7]Université de Bordeaux, Fundamental Microbiology and Pathogenicity Laboratory, CNRS UMR5234, Département de Sciences Biologiques et Médicales, Bordeaux, France.
[8]Viral DNA Integration and Chromatin Dynamics Network (DyNAVir), France.

[#]Correspondence: celine.loot@pasteur.fr; Tel: +33 1 4061 3287

Abstract

Integrons are genetic elements involved in bacterial adaptation which capture, shuffle and express genes encoding adaptive functions embedded in cassettes. These events are governed by the integron integrase through site-specific recombination between *attC* and *attI* integron sites. Using computational and molecular genetic approaches, here we demonstrate that the integrase also catalyzes cassette integration into bacterial genomes outside of its known *att* sites. Once integrated, these cassettes can be expressed if located near bacterial promoters and can be excised at the integration point or outside, inducing chromosomal modifications in the latter case. Analysis of more than 5×10^5 independent integration events revealed a very large genomic integration landscape. We identified consensus recombination sequences, named *attG* sites, which differ greatly in sequence and structure from classical *att* sites. These results unveil an alternative route for dissemination of adaptive functions in bacteria and expand the role of integrons in bacterial evolution.

Main

Bacteria can exchange and recombine DNA, enabling acquisition of new genes and adaptation to changing environments. An active player of Gram-negative bacteria adaptation is the integron system [1]. Integrons are natural genetic "toolboxes" able to stockpile, shuffle, and express adaptive functions encoded by arrays of coding sequences (CDS) known as 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 of integrons, such as their association with transposons and conjugative plasmids. These so-called Mobile Integrons (MIs) have now disseminated among bacteria and constitute an important means of spreading antibiotic resistance. Importantly, MIs are only the tip of the iceberg since much larger Sedentary and Chromosomal Integrons (SCIs) have been discovered in ~17% of the available genomes in databases [3,4]. Cassettes in SCIs constitute a large reservoir of functions for MIs and some have been found to be involved in mobility, metabolism, biofilm formation, bacteriophage resistance or host surface polysaccharide modification [4,5]. Both MIs and SCIs share the same general organization: a stable platform and a variable cassette array (Fig. 1a). The platform is composed of three elements: the *intI* gene coding for a site-specific recombinase (the integron integrase), the *attI* recombination site in which promoterless cassettes are integrated and a P_C promoter oriented to direct transcription of proximal cassettes [2]. More distal cassettes provide a low-cost memory of valuable functions for the cell which can potentially be expressed through the reordering of the cassette array. By controlling the expression of IntI, bacteria can reshuffle the integron cassettes "on demand" in moments of stress [6]. All these features make the integron a unique recombination system [6-8].

A cassette is a mobilizable element that generally contains a CDS ended by an *attC* site. Cassette reordering is ensured by cassette excision events via recombination between two

consecutive *attC* sites, followed by integration events of the excised cassettes in the *attI* site (*attC* × *attI* recombination, Fig. 1a). *attC* sites share little sequence conservation and are instead recognized by the integrase through the single-stranded structures formed by their bottom strands (Fig. 1b) [9]. 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.

Most studies on integrons focused on the recombination properties of cassettes in the proper integron *att* sites. There have been a couple of examples of resistance cassettes being integrated into plasmids and genomes at what was called secondary sites [10-16]. These rare integration events have been overlooked for decades and considered anecdotal in the integron functioning. Here, we focused on the ability of bacteria to access these adaptive functions and to build a repertoire of cassettes pertinent for their lifestyle. Our genomic analysis revealed numerous occurrences of isolated cassettes in a wide range of sequenced bacterial genomes available in databases. In line with the nomenclature proposed previously [3], we named these cassettes, SALIN (for Single *attC* site lacking integron-integrase) and revealed the mechanism of their formation. We propose that they may result from cassette integration into genomes. By performing a series of *in vivo* experiments, we confirmed the high propensity of integron cassettes to disseminate into bacterial genomes. We demonstrated that these integrated cassettes can be expressed, providing the potential to alter bacterial function, but also excised in such a way that can induce genome modifications. Deep sequencing of cassette libraries integrated *de novo* into the *Escherichia coli* genome (1) showed that cassettes can target a very large number of unique sites and (2) enabled the characterization of these integration sites. Surprisingly, they differ greatly from both classical *attC* and *attI* recombination sites in terms of sequence and structure. We named these sites, "*attG"* for attachment site of the genome.

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

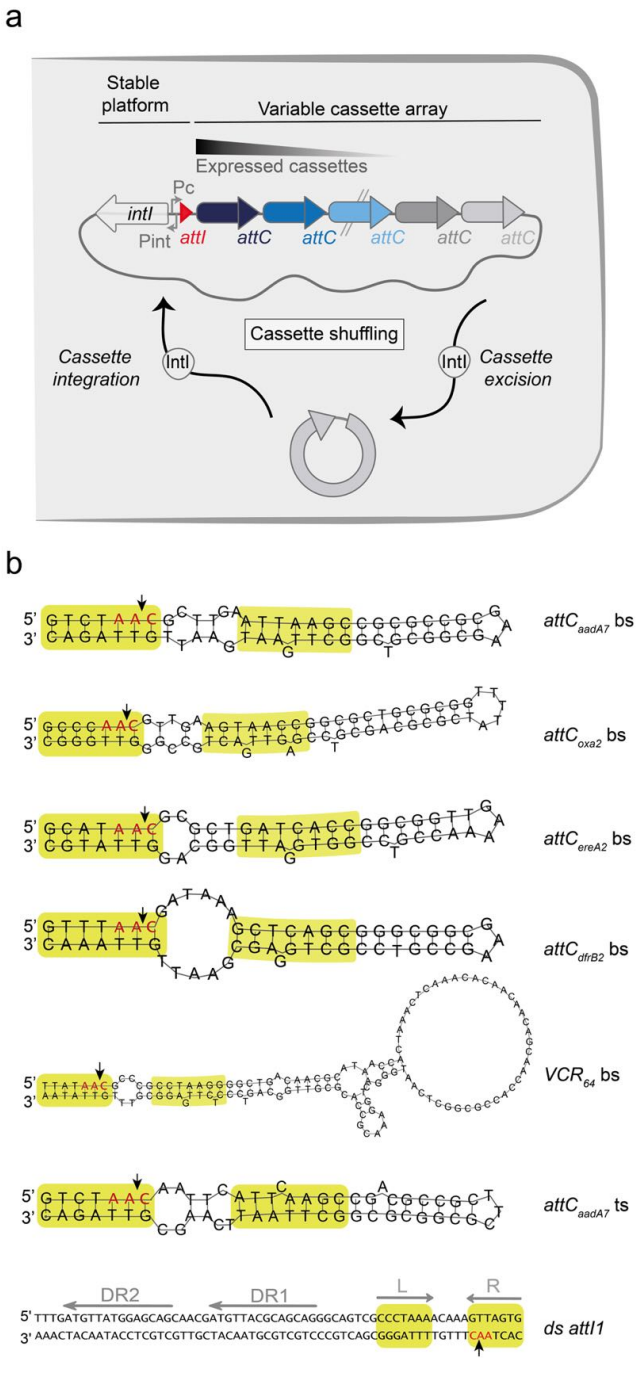


Figure 1: The integron system and the *att* sites

a) 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 the variable

cassette array. The variable cassette array contains some cassettes represented by small coloured 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 integration of the excised cassettes in the first position in the array (*attI* × *attC*).

b) The *att* sites
The sequences of the single-stranded bottom and top strands (bs and ts) of the *attC* sites and the double-stranded *attI1* (ds) site used in this study are represented. Green boxes indicate the right (R) and left (L) integrase binding sites. The 5'-AAC-3' triplet, where the cleavage takes place, is highlighted in red and the precise cleavage point is indicated by a black arrow. The Direct Repeats (DR1 and DR2) of the *attI1* site are shown by grey arrows.

RESULTS

In silico isolated cassette detection in bacterial genomes

We took advantage of the recent release of IntegronFinder 2.0 [4] to search for isolated cassettes in the 21,105 complete bacterial genomes retrieved from NCBI RefSeq database. We defined isolated cassettes as CDS associated with an *attC* site with no other detectable integron features in its vicinity (that is, 4kb). We found 2,469 genomes containing isolated cassettes (Fig. 2a). By analogy with the previously described CALINs (Clusters of *attC* sites lacking integron-integrases) [3], we called these isolated cassettes SALINs. Among the 2,469 genomes containing SALINs, 1,847 contain only one SALIN and the remaining 622 more than one SALIN (Fig. 2b). Interestingly, SALINs are more represented than CALINs (3,400 SALINs versus 2,961 CALINs, Fig. 2c). CALINs 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 CALINs (95%) are not close to recognizable *intI* pseudogenes [3]. Furthermore, we observe that many CALINs are not surrounded by transposable elements and tend to be small: among the 2,961 CALINs, 1,362 harbor only 2 cassettes (Fig.2c). Interestingly, we found certain phyla such as the Tenericutes in which we only observed SALINs (Fig. 2d). This raises the possibility that,

rather than being remnants of integrons, some of small CALINs and SALINs could result from the integration of integron cassettes in bacterial genomes.

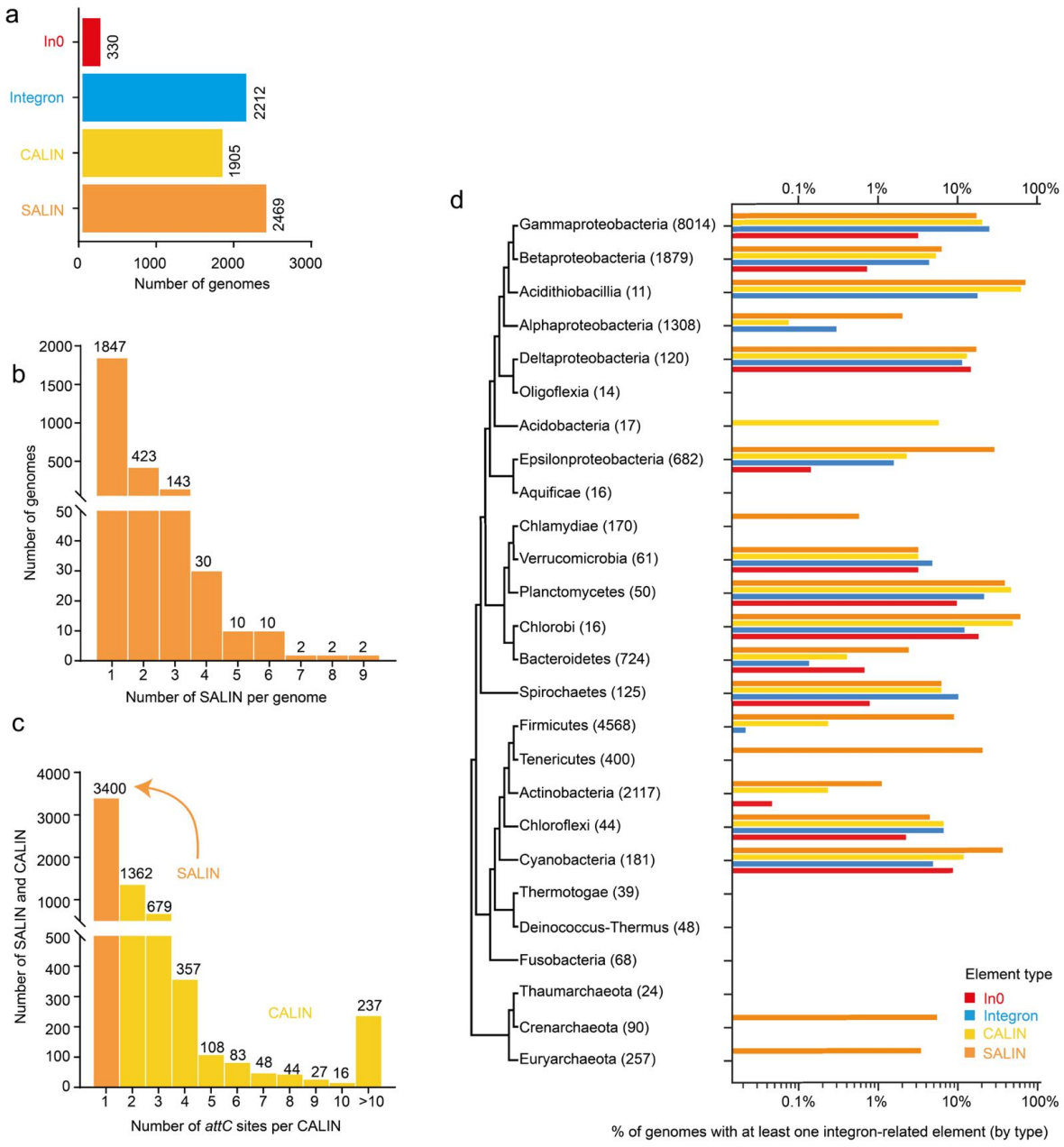


Figure 2: Distribution of integrons across bacteria using the RefSeq NCBI database

a) Number of bacterial genomes containing either a complete integron (integron), an integrase gene (In0), a CALIN or a SALIN
b) Number of SALINs found per genome
c) Number of *attC* sites per CALIN
d) Taxonomic distribution of SALINs, CALINs, In0 and integrons across major bacteria phyla

Integron cassettes can disseminate in bacterial genomes

To validate our hypothesis, we tested the integration capability of integron cassettes in bacterial genomes. Using our suicide conjugation assay, we delivered *attC*-containing plasmids (mimicking cassettes) in a single-stranded form to recipient strains containing a vector expressing (or not) the integrase IntI1 (Fig. 3a) [9,17,18]. We used the R32_H39 variant of IntI1 ($IntI1_{R32\_H39}$; see Methods) which corresponds to the one we classically use because it is the most represented and the most effective variant among MIs of class 1 in terms of recombination [19]. We tested the integration properties of different *attC* sites from both MIs and SCIs ($attC_{aadA7}$, $attC_{oxa2}$, $attC_{ereA2}$, $attC_{dfrB2}$ or VCR sites) chosen for their high recombinogenic properties (Fig. 1b) [20]. All these sites showed a rate of recombination comprised between 10^{-2} and 10^{-4}, far above what was obtained without integrase expression (Fig. 3a). Using random PCR and sequencing approaches (Extended Data Fig. 1 and Methods), we confirmed that cassettes were integrated at several locations into the *E. coli* genome. Furthermore, the rate of integration 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 substantial extent, whereas *attC* sites can recombine in genomes in the same way that they recombine with the *attI* and *attC* sites, that is, as a single-stranded structured form made by the bottom strand but not by the top one [9].

Replacing the integrase IntI1 by IntI2 or IntI3, the respective integrases of the class 2 and class 3 MIs, still provided a high rate of integration of the $attC_{aadA7}$ cassette, either in the *E. coli* genome (Fig. 3a) or in a recipient plasmid carrying the *attI2*/*attI3* or $attC_{ereA2}$ canonical sites (Extended Data Fig. 2), indicating that the property to integrate integron cassettes into the *E. coli* genome is not restricted to IntI1. We also demonstrated that IntI1 can mediate cassette integrations at high frequency (almost 10^{-3}) in the genome of another Gram-negative bacteria, a pathogenic *V. cholerae* strain (Fig. 3a), despite the presence of the *V. cholerae* endogenous chromosomal integron (Methods) [21].

We also performed a conjugation assay using an *E. coli* recipient strain containing a synthetic integron (*attI1* and $attC_{aadA1}$ sites) carried by a heat-sensitive plasmid (Extended Data Fig. 3a and Methods). First, the recombination step was performed in the presence of all the target sites (on plasmid and genome), that is, at the permissive 30°C temperature. Second, selection was carried out in parallel at 30°C and at 42°C to ensure maintenance or loss of the recombined heat-sensitive plasmid, respectively. At 42°C, since recombination events in the genome are more easily quantified by PCR (Extended Data Fig. 3b), the results are very robust. We obtained a high rate of cassette recombination in the *E. coli* genome (close to 10^{-2}, Extended Data Fig. 3c) demonstrating that integron cassettes can be disseminated in host genomes at high frequency even in the presence of *attC* and *attI* sites carried by resident integrons.

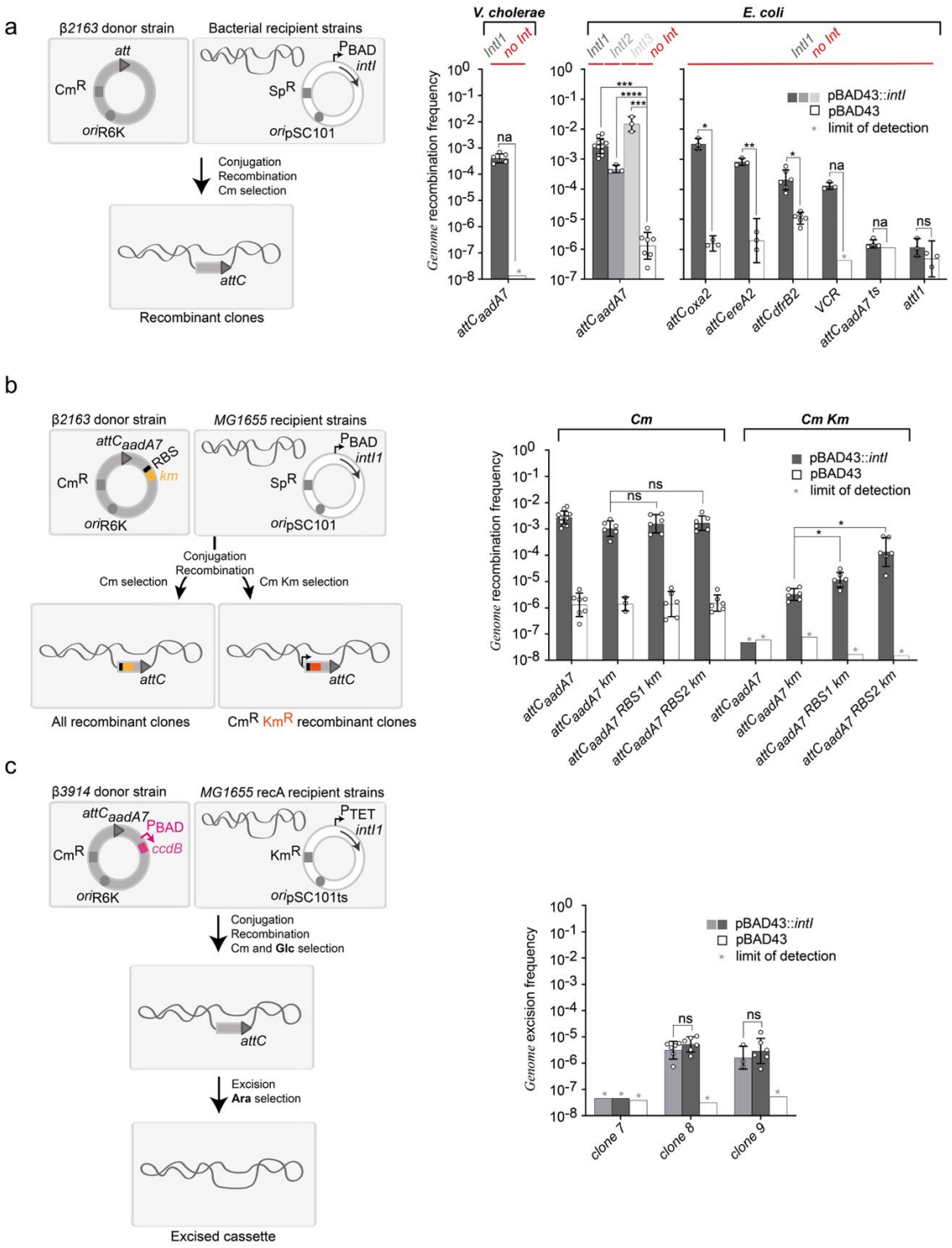


Figure 3: Cassette integration and excision in genome sites

a) Cassette integration in bacterial genome sites
The experimental setup of the cassette integration assay is described (left panel). pSW23T plasmids carrying *att* sites (grey triangle) are delivered by conjugation from the *E. coli* donor strains to recipient strains containing a plasmid carrying (or not) the integrase gene (*intI*).

Phenotypic resistances are represented by grey rectangles and origin of replication by grey circles. Recombinant clones are selected on chloramphenicol (Cm)-containing plates. The graphs represent the recombination frequencies (right panel). The recipient strains (*E. coli* or *V. cholerae*) and the expressed integrases (IntI1, IntI2 or IntI3) are indicated.

b) Expression of genome-integrated cassettes
The setup is the same as in **a**, except that the donor plasmids contain an $attC_{aadA7}$ site and a promoterless kanamycin gene (*km*, orange rectangle) associated (or not) with an RBS (RBS1 or RBS2) (left panel). Recombinant clones are selected on Cm-, and Cm- and Km-containing plates. The graph represents the recombination frequencies (right panel).

c) Excision of genome-integrated cassettes
The setup is the same as in **a**, except that (1) the donor plasmid contains an $attC_{aadA7}$ site (grey triangle) and the *ccdB* toxic gene (pink rectangle) under the control of the P_{BAD} promoter (pink arrow), induced by Ara and repressed by Glc, and (2) the MG1655 *recA E. coli* recipient strain contains a thermosensitive (ts) plasmid expressing (or not) the IntI1 integrase under the control of the P_{TET} promoter (left panel). Recombinants clones are selected on Ara-containing plates. The graph represents the excision frequencies for clones 7, 8 and 9 at the integration point (dark grey bars) and outside (light grey bars) (right panel).

For **a**, **b** and **c**, bar charts show the mean ± s.d. of at least three independent experiments ($n \geq 3$, individual plots). The precise "n" values are indicated in Source data Fig. 3. Student's *t*-test all two-sided: na, not applicable, ns, not significant; $****P < 0.0001$, $***P < 0.001$, $**P < 0.01$ and $*P < 0.05$; grey asterisk (*) indicates the recombination frequency was below detection level.

Genome-integrated cassettes can be expressed

To determine whether the promoterless integron cassettes can be expressed when integrated into the genome, we added a kanamycin resistance gene devoid of promoter, preceded (or not) by a ribosome binding site (RBS), into the donor plasmid containing the $attC_{aadA7}$ site (Fig. 3b). This matches previous observations where some CDS in cassettes are preceded by a suitably spaced RBS [2]. We tested two different RBS sites, RBS1 and RBS2, naturally found in some integron cassettes. For all tested cassettes, as expected, a high frequency ($>10^{-3}$) of genomic integrations was observed in Cm-selective medium (Fig. 3b). Using selective medium containing Cm and Km, the integration and expression rate was high, as soon as the donor plasmid presents the *km* resistance gene and especially when it is preceded by an RBS motif (up to 10^{-4}). Comparing the rates obtained in presence of Cm (2×10^{-3}) or Cm and Km (2×10^{-4}) indicated that up to 10% of integrated cassettes can be expressed when carrying RBS2.

Performing random PCR and sequencing on over 40 randomly chosen Km^R clones confirmed that the expressed cassettes were integrated near a resident promoter. These results demonstrate that a large proportion of genome-integrated cassettes can be expressed if they are located in the vicinity of a promoter, thus conferring a new phenotype on the bacteria.

To further evaluate the potential impact of cassette integration on bacterial evolution, we estimated the proportion of SALINs that might be expressed. According to the gene prediction program Prodigal [22] that also identifies translation initiation sites, ~78% of SALINs contain an RBS, which is in line with the general prevalence of RBS in bacterial genes (~77%) [23]. Noticeably, the prevalence of RBS is higher in CALIN-located (~84%) and complete integron-located (~88%) genes. By comparison, an RBS is detected in ~90% of *E. coli* reference strain K-12 MG1655 genes, in ~91% of the genes of *V. cholerae* N16961 secondary chromosome (which harbours an SCI), and in ~88% of *Klebsiella pneumoniae* reference strain HS11286 genes (the species with the highest number of SALINs detected). Thus, RBS predictions are in agreement with the hypothesis that the majority of SALIN-located genes are expressed even if a slightly larger fraction of these genes lacks an RBS compared with integron- and CALIN-located genes.

Genome-integrated cassettes can be excised

To test whether genome-integrated cassettes can be excised, we added into the $attC_{aadA7}$-containing donor plasmid, a *ccdB* gene encoding a bacterial toxin under control of the P_{BAD} promoter (Fig. 3c) [24,25]. First, we performed our conjugation assay and selected the cassette integration events, adding glucose to repress the *ccdB* gene. Second, we randomly chose 24 recombinant clones and performed an excision assay (Methods). We selected excised clones on arabinose containing plates. In these conditions, the CcdB toxin is expressed and only clones that have lost the *ccdB* gene due to a cassette excision event are selected, while the others die.

We detected excision events above the limit of detection for clones 8 and 9, but not for the others such as clone 7 (Fig. 3c). PCR and sequencing analysis revealed that excision events can occur at the precise integration site or at other nearby genomic sites, inducing genome modifications for the latter case (Fig. 3c and Extended Data Fig. 4).

Cassette integration occurs into many genomic locations

Deciphering where and how integron cassettes are integrated into genomes is fundamental to understand their potential cost and impact on host evolution. We therefore performed a genome-wide next-generation sequencing (NGS) mapping of integration sites using a library of ~50,000 recombinant clones obtained after integration of the $attC_{aadA7}$ cassettes in the *E. coli* genome, catalyzed by the IntI1 integrase (Extended Data Fig. 5). Genomic sequences flanking the integrations were extracted from each sequencing read, aligned to the *MG1655 E. coli* reference genome and used to call precise integration sites. Genomic integration into the *E. coli* genome occurred at many positions (22,271 unique integration sites; Extended Data Fig. 6), with a huge variation in site usage (Fig. 4a, b). To better analyze these data, we first removed the duplicated reads from the 3,205,043 reads, leading to 361,464 reads (Fig. 4c-h and Extended Data Fig. 6). A window size of 200,000 bps sliding every 100 bps along the genome uncovered preferential integrations near the origin of replication (Fig. 4c). Performing multiplex digital PCR, we demonstrated that during the assay, the cells contain twice as many *oriC* than *terC* copies, explaining at least in part why integrations are favoured near the origin of replication (Extended Data Fig. 7). Alignment of all the DNA sequences flanking the integration cutting sites revealed a short 5'GWT3' consensus sequence (Fig. 4d). No obvious bias of integration was detected regardless of the forward or reverse strand of the genome (Fig. 4e, left panel), and regardless of the leading or lagging strand template during replication (Fig. 4e, right panel). As expected, a decrease of integrations in the essential genes of the *E. coli* genes (7% (295/4,213)[26]), when

compared with non-essential genes or when using random integrations (Fig. 4f,g), was detected but with no obvious effect of integration in the same or opposite direction of transcription (Fig. 4g). Finally, integration around transcription start sites (TSS) appeared shifted upstream of essential genes (-105 bps, Fig. 4h). Altogether, we conclude that cassette integration can occur all throughout the *E. coli* genome with no notable effect of genomic features, except the 5'GWT3' DNA sequence motif, and with a visible counterselection of integrations in essential genes. Interestingly, the same results were observed using IntI2 and IntI3 (Extended Data Fig. 6). The only remarkable difference was a 1-base-larger 5'TGWT3' consensus integration site for IntI2 that may explain the lower frequency of cassette integration into genomes mediated by this integrase (Fig.3a).

We took advantage of the high number of integrations obtained experimentally with IntI1 to quantify the ability of IntegronFinder 2.0 to detect SALINs in bacterial genomes, in a *post hoc* analysis. We simulated 22,271 *E. coli* MG1655 genomes, each corresponding to one $attC_{aadA7}$ cassette integration event, and ran IntegronFinder 2.0 (Methods). We retrieved 74.5% of the integrated $attC_{aadA7}$ cassettes and obtained only 4 false positives, thereby confirming that IntegronFinder 2.0 faithfully detects SALINs at a level close to the general sensitivity of the software for classical *attC* sites (~90%) [3].

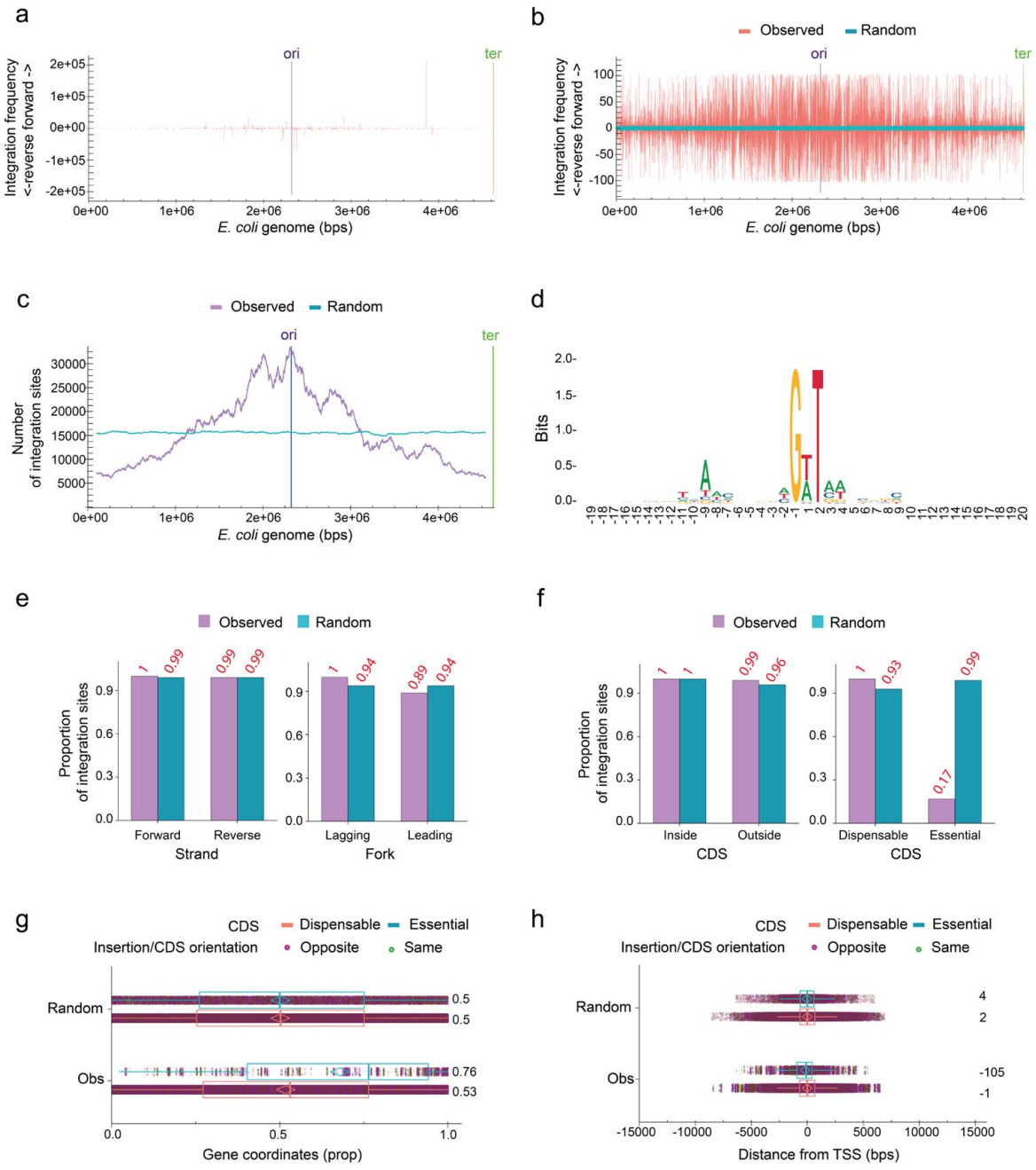


Figure 4: Computational analysis of Deep sequencing data

a) Integration site usage all throughout the *E. coli* genome without read-duplicate removal (n = 3,205,043)
Integration in the forward or reverse strand determines the orientation of the integrated cassettes (see Methods for details). bp, base pair. Ori and ter, region of origin and termination of replication, respectively.

b) Integration site usage all throughout the *E. coli* genome with read-duplicate removal (n = 361,464)

All other panels in this figure derive from read-duplicate removal. Observed, observed integrations; Random, random integrations using the GWT consensus motif of integration (see **d**).

c) Number of integration sites all throughout the *E. coli* genome using a sliding window of 200 kb sliding every 100 bases

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

e) Proportion of integration sites according to the strand polarity (Forward or Reverse) and according to replication orientation (Leading or Lagging)
Leading and lagging mean that the read corresponds to the neo-synthetized leading and lagging strand during replication, respectively. The proportions are relative to the maximal proportion set to 1.

f) Proportion of integration sites according to the CDS in the genome, either inside/outside the 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.

g) Relative position of integration sites inside dispensable (salmon pink) 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 the interquartile range and mean, respectively. Numbers on the right side correspond to median values. Each dot represents a single integration site in purple or green, depending on the opposite or same cassette orientation vs the CDS orientation, respectively.

h) As in **g** but for the distance of integration site from the closest TSS (in base pairs (bps))

Cassette integration occurs in hotspots in the *E. coli* genome

Several integration hotspots were detected when considering all the reads (that is, including the duplicate ones), regardless of the tested integrase (IntI1, IntI2 or IntI3) (Fig.4a and Extended Data Fig. 6). Note that for IntI1 and IntI3, the strongest hotspot was the same, located in the *ybhO* gene (Extended Data Fig. 6, red boxes). However, duplicate reads can be artefacts coming from the used experimental procedure (Methods). Thus, to experimentally validate these integration sites as hotspots, the six strongest hotspots resulting from the IntI1 experiment (corresponding to integrations in *ybhO*, *alsB*, *ilvD*, *pyrE*, *metC* and *yjhH* genes) were cloned in a recipient plasmid (Fig. 5a). As a control, we used a genomic site used only once by IntI1 for

cassette integration, called US-*ygcE* (US for Unique Spot), and another site used 15 times, that is, very close to the median of integration site usage, called MS-*abgA* (MS for Median Spot). In each case, the cloned segment encompassed the 300 and 200 bps flanking the 5'GWT3' integration point. The six hotspots showed a high rate of recombination (~10^{-3}, close to the classical *attI* and *attC* sites), while no recombination event was detected for the US and MS control sites (Fig. 5a). Notably, the highest integration efficiency was obtained for the *ybhO* hotspot, the one showing the highest integration usage (209,387). These results confirm that these hotspots are regions attracting cassette integrations and not the consequences of experimental bias.

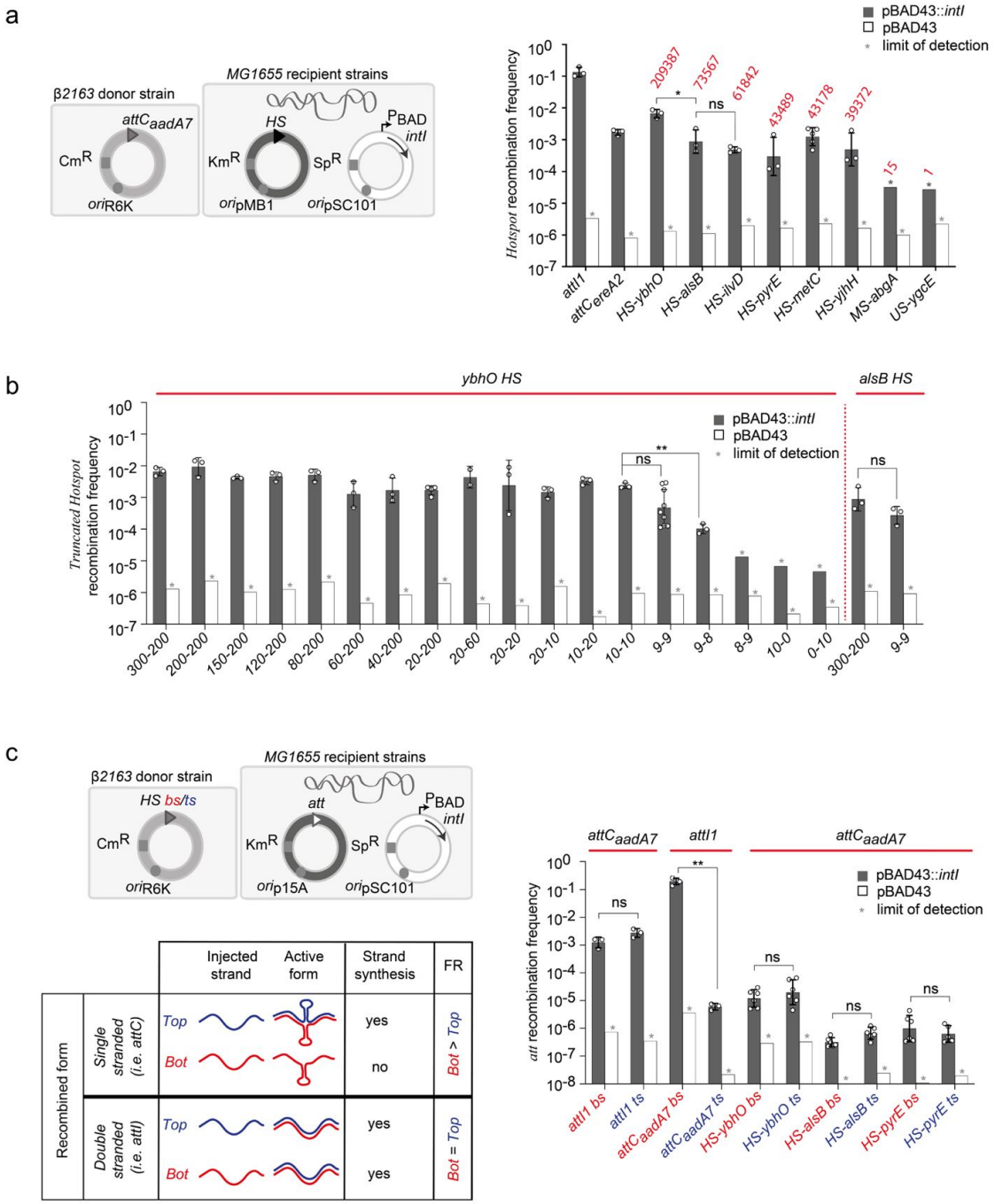


Figure 5: *Hotspots* as donor and receptor sites during conjugation assay.

a) Testing *hotspots* as receptor sites

The setup is the same as in Fig. 3a, except that the donor plasmid contains an $attC_{aadA7}$ site and the *E. coli* recipient strains contain recipient plasmids carrying *attI1*, $attC_{ereA2}$, hotspot (HS), median spot (MS) and unique spot (US) sites (black triangle) (left panel). The graph represents the recombination frequencies (right panel). The receptor sites are indicated in the *x*-axis

legend. The number of integrations that we previously obtained in each hotspot site (Fig.4a) is indicated in red at the top of the bars.

b) Testing the truncated *hotspots* as receptor sites
The graph represents the recombination frequencies. The receptor sites are indicated in the *x*-axis legend. The 2 numbers represent the number of base pairs kept on each 5' and 3' side of the cutting position in the *ybhO* and *alsB* hotspot sites (see Fig. 4d).

c) Testing *hotspots* as donor sites
The setup is the same as in Fig. 3a, except that the donor plasmids contain *att* and *hotspot* sites (grey triangle) delivering either the bottom (bs) or the top strand (ts), and the *E. coli* recipient strains contain recipient plasmids carrying *attI1* or $attC_{aadA7}$ sites (white triangle, left panel, top). Scheme shows the expected frequencies of recombination (FR) injecting the bottom (bs, red line) or the top strands (ts, blue line) as a function of the recombined form. The graph represents all the recombination frequencies (right panel). The donor sites are indicated in the *x*-axis legend and the receptor sites at the top of the bars.

For **a**, **b** and **c**, bar charts show the mean ± s.d. of at least three independent experiments ($n \geq 3$, individual plots). The precise "n" values are indicated in Source data Fig. 5. Student's *t*-test all two-sided: ns, not significant; $^{**}P < 0.01$ and $^{*}P < 0.05$.

Cassette integration sites differ from *att* recombination sites

To further characterize the properties of genome sites, we used the *ybhO* hotspot site as a proxy for integration 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 (Fig. 5b). Reducing the lengths to 9 nucleotides on each side maintained the recombination frequency above 10^{-3} (9-9, Fig. 5b), defining a functional integration site with a minimal length of 18 nucleotides, while smaller lengths hampering the recombination frequencies (Fig. 5b). These functional 9-9 lengths were confirmed for the *alsB* hotspot (Fig. 5b). Alignment of the 10 base pairs on each 5' and 3' side of the cutting position of the six hotspots from Fig. 5a uncovered a consensus sequence, keeping the highly conserved 5'GWT3' residues in positions -1 to 2 but revealing two supplementary motifs on either side of the cleavage site (Extended Data Fig. 8a). We validated the importance of the 5'CRGM3' right motif in positions 6 to 9. Indeed, replacing this motif by the 5'CCAG3' and 5'TCAG3' motifs in the *ybhO* hotspot decreased integration frequencies by more than 2 logs compared with the wild-type *ybhO*

5'CAGC3' or the consensus 5'CAGA3' sequences (Extended Data Fig. 8b). Interestingly, performing the same position replacements in the right part of the *attI1* site did not reduce the recombination frequency (Extended Data Fig. 8c), indicating structural differences between the *attI* and genome sites. From these, we conclude that genome sites differ broadly from both classical *attC* and *attI* recombination sites in terms of sequence and structure. We named these integron sites "*attG*", where G stands for "genome".

attG sites recombine as double-stranded forms

To determine the double-stranded or the single-stranded nature of the *attG* sites, we cloned three hotspots in the donor plasmid 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 an $attC_{aadA7}$ receptor site cloned in the recipient plasmid (Fig. 5c). If *attG* sites are recombined as a strand-specific single-stranded form, second-strand synthesis is only required when the non-recombined strand is injected. Therefore, a difference in the recombination frequency would be expected while injecting the top or the bottom strand as shown for *attC* sites [9] (Fig. 5c). In contrast, if *attG* sites are recombined as a double-stranded form, second-strand synthesis is required whatever the injected strand. In this case, no difference in the recombination frequency is expected as shown for the *attI* site [9] (Fig. 5c). Obtained integration frequencies varied between the tested hotspots but were similar regardless of their orientation (Fig. 5c). These results confirm that *attG* sites are recombined as double-stranded form, similar to the *attI* sites. PCR and sequencing of recombined products demonstrated a sequence homogeneity around the integration site, thus confirming that recombination occurs by a single cleavage of the doubled-stranded matrix (Extended Data Fig. 9a). Indeed, a double cleavage would lead to a heterogeneity of sequences at the integration point as expected when the core sequences of *att* sites are different (Extended Data Fig. 9a). We also confirmed by sequence

analysis and determination of the cassette integration orientation that recombination takes place between the bottom strands of both the hotspots and *attC* receptor sites and at the expected recombination points (Extended Data Fig. 9b). 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), illustrating the robustness of our results. We therefore conclude that during cassette integration into the genome, the *attG* sites recombine as double-stranded forms and that only their bottom strands are cleaved.

DISCUSSION

A few previous studies revealed the ability of cassettes to insert at a very low rate (around 10^{-6}) in secondary sites located in plasmids and genomes [10-16]. To determine the extent of this cassette dissemination, out of the integron platform, we decided to perform an extensive bioinformatics analysis of available sequenced bacterial genomes. This analysis revealed that many isolated cassettes are integrated in bacterial genomes. We called these cassettes SALIN and 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 at a frequency close to that obtained using canonical *attI* and *attC* sites. We unveiled that integrase integration sites have a very small 5'GWT3' consensus sequence, meaning that the cassette integration 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 demonstrated that cassette integrations can induce genome modifications by disrupting genes and that excisions of these same cassettes, when they take place outside the integration site, can lead to the deletion of pieces of the genome. However, for most of the cassettes integrated into

the genome (22 of the 24 tested), no excision events could be detected. This does not mean that excision events cannot occur, but rather that if they do occur, it is at a very low level since the limit of detection in this experiment is below 10^{-7}. The rate of excision could depend on whether the integration of the genomic cassette reconstitutes an effective genomic site. Nevertheless, a cassette, once integrated into the genome, is quite stable. We have shown that these cassettes can even represent a gain in function for the bacterium under the conditions in which they are expressed. These cassettes could also be domesticated, as folded single-stranded *attC* sites can be easily eliminated by replication slippage events [27,28]. In this way, the integron system could contribute to the turnover of bacterial genetic content, at least for genes involved in adaptation to changing environments.

Interestingly, another site-specific recombinase, the lambda (λ) integrase, which ensure λ phage integration at the *attB* specific site in the host chromosome, was also described as being able to catalyze phage integration into genome sites. However, the observed consensus integration sequence is much larger than that of integron integrases, restraining the secondary integration sites to a small number of locations in bacterial chromosome (that is, 304 unique sites [29,30]). Consequently, the mutational landscape generated by phage λ integration in secondary genomic sites should be much more restricted than that generated by integron cassette integration. Another difference is that the secondary sites of the λ integrase are actually very similar to the *attB* recombination site, probably resulting from an off-target activity of the λ integrase. In contrast, we have shown that the genomic sites of the integrase are structurally different from both *attC* and *attI* sites, that is, with a cleavage point located in their central part, thus bringing them closer to classical double-stranded core recombination sites such as *dif* sites [31]. We therefore called these sites, *attG* sites. As these *attG* sites are recombined as a double-stranded form, we suggest that *attI* sites could have been derived from the sequence of one of these hotspots through co-evolution with an integrase.

In conclusion, we have demonstrated the existence of an alternative cassette recombination route that greatly expands the role of integrons in the dissemination of adaptive functions such as antibiotic resistance (Fig. 6). This route could allow bacteria to "safeguard" cassettes in their genomes. This could be particularly advantageous in conditions where MIs are carried by conjugative plasmids that cannot be maintained in the cell. Beyond that, these results show that the integron system could represent a general mechanism for genomic diversification driving bacterial evolution.

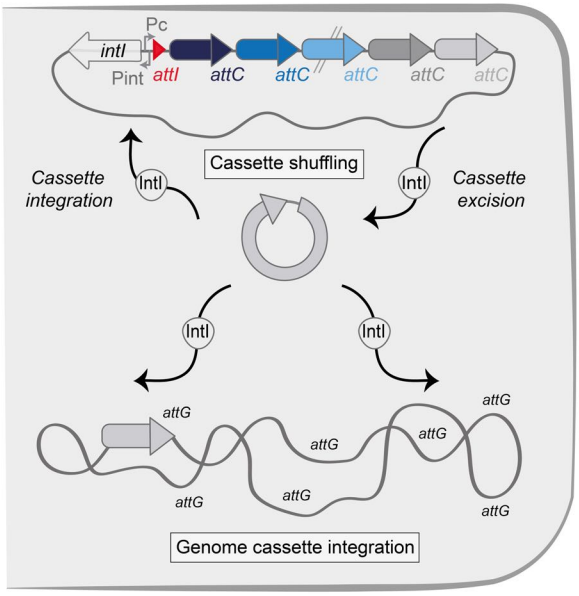


Figure 6: The alternative route of the integron cassette dissemination

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

METHODS

Bacterial strains, plasmids and primers

All plasmids, strains and primers used or constructed in the present study are described in Supplementary Table 1.

Media

E. coli and *V. cholerae* were grown in Luria Bertani broth (LB) at 37°C. *E. coli* strains containing a plasmid with a thermosensitive 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 100 ng/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).

Genomics analysis of isolated integrated cassettes

Data

The genome dataset analyzed in the current study is the same as in ref [4], consisting of 21,105 complete genomes downloaded from the NCBI RefSeq database of high-quality non-redundant prokaryotic genomes (accessed on 30 March 2021, https://ftp.ncbi.nlm.nih.gov/genomes/refseq/). Using IntegronFinder 2.0 [4], we screened the dataset to automatically and accurately identify integrons, CALINs and isolated integrases (In0). Data analysis and visualization were conducted using the Python packages jupyter (v.1.0.0) [32], pandas (v.1.4.0) [33], biopython (v.1.79) [34] and seaborn (v.0.11.2) [35].

SALIN analysis in bacterial genomes

SALIN elements are single *attC* sites lacking integrase. They are a subtype of CALIN where the cluster comprises solely an *attC*. SALINs were detected with IntegronFinder 2.0 with the option "--calin-threshold 1" to prevent filtering out single *attC* sites.

Distribution of SALINs across bacterial phyla

Each genome of the dataset was assigned to a bacterial phylum according to its taxonomic identifier retrieved from the NCBI Taxonomy database. Phyla were organized in a cladogram adapted from the tree published in ref [36]. Only phyla comprising at least one genome from our dataset were considered to build the cladogram.

RBS detection in SALINs

To detect RBS in SALINs, we re-predicted CDS using Prodigal (which includes RBS prediction) [22]. When integrated, the SALIN's *attC* site is located at its 3' end and thus delimits its 3' border. However, detecting the 5' border is challenging, as it would require the precise identification of the short integration site. Since most of integron cassettes contain a single gene, we considered for each SALIN the first CDS beginning before the detected *attC* site and checked whether it harbored an RBS. The same procedure was applied to *attC* sites detected in complete integrons and CALINs to measure the prevalence of RBS.

IntegronFinder 2.0 ability to detect SALINs

To quantify the ability of IntegronFinder 2.0 to detect SALINs in bacterial genomes, we simulated 22,271 *E. coli* MG1655 genomes, each corresponding to one $attC_{aadA7}$ cassette integration event. The integration sites were obtained after the genome-wide NGS mapping of the $attC_{aadA7}$ cassettes catalyzed by the IntI1 integrase (Fig. 4 and Extended Data Fig. 6). We ran IntegronFinder 2.0 on this set of genome sequences and quantified the numbers of true-positives ($attC_{aadA7}$ retrieved at the expected locations), false-positives (*attC* sites found at unexpected locations) and false-negatives ($attC_{aadA7}$ not retrieved). All the scripts used to

generate these simulations, as well as the IntegronFinder 2.0 analysis, are available at https://gitlab.pasteur.fr/hub/salin.

Suicide conjugation assays

The conjugation assays were based on refs [9,18,37]. This assay mimics the natural conditions in which cassettes are delivered through horizontal gene transfer. Conjugation ensures the delivery of one of the recombination substrates (*att* sites) carried by a pSW23T suicide plasmid into a recipient cell containing a plasmid (pBAD43 or pBAD18) with or without integrase 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 [38], 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 as 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. To be maintained, the pSW vector has to recombine with *att* sites contained in the recipient strain or to integrate into the bacterial genome. Briefly, the donor strains were grown overnight in LB media supplemented with Cm (resistance of the pSW plasmid), 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 (to repress the integrase gene when P_{BAD} promoter is used). Both donor and recipient overnight cultures were diluted 1:100 in LB with DAP or arabinose, respectively, and incubated until optical density (OD)=0.7-0.8. Of each culture, 1 ml was then mixed and centrifuged at 3,500 *g* for 6 min. The pellet was suspended in 100 µl LB, spread on a conjugation membrane (mixed cellulose ester membrane from Millipore, 47 mm diameter and 0.45 µm pore size) over an LB+agarose+DAP+Ara Petri dish and incubated for 3 h for

conjugation and recombination to take place. The membrane with the cells was then resuspended in 5 ml LB, after which serial 1:10 dilutions were made and plate on LB+agarose media supplemented with appropriate antibiotics. Note that we adapted this protocol when using *V. cholerae* as recipient strain in which plasmids are more easily lost in the absence of antibiotic selection than in *E. coli* [18,39]. The recombination frequency was calculated as the ratio of recombinant colony-forming units (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). When we did not detect any recombinant CFUs, we considered that we obtained only one recombinant CFU and we calculated that we called, the limit of detection (*). It corresponds to "1 recombinant CFU / the total number of recipient CFUs obtained for all the *n* replicates". When we detected a few recombinants CFUs (α clones) but not in all *n* replicates, we calculated a unique recombination frequency for all replicates. This corresponds to the ratio "α recombinants CFUs/the total number of recipient CFUs obtained for all the *n* replicates". Note that in the last two cases, we cannot calculate the standard deviation, so the bar chart does not show an error bar and the calculation of the P value is not applicable (na).

It should be noted that integrase-independent microhomology recombination can also occur between donor plasmids and bacterial genomes (or recipient plasmids), albeit at a low level. For this reason, we systematically performed integrase-free controls. In some experiments, we detected a few recombination events without integrase, while in others we detected none. This variation could be due to the difference between the sequences of the *attC* sites carried by the donor plasmids, the difference between the sequences of *E. coli* and *V. cholerae* genomes, or simply the fact that these frequencies are established with a very low number of clones, making the results less robust. Note that recombination by microhomologies is probably enhanced by the delivery of donor plasmids in single-stranded form (that is, by conjugation).

Conjugation assay in presence of integron

To determine the genome cassette integration 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 (Sp^R, pM335). Once constructed, this vector was transformed into the MG1655 recipient strain containing the carbecillin-resistant ($Carb^R$) pBAD18 plasmid with or without integrase (p3938 and p979 respectively [40,41]). These donor and recipient strains were conjugated for 3 h but at 30°C. Thus, the conjugation and recombination reactions were performed in presence of the synthetic integron, that is, under a condition in which the plasmid can replicate (at 30°C). The membrane with the cells was then resuspended in 2 ml LB+Carb+Glc, divided in two parts and incubated 24 h at 30°C and 42°C. These incubation temperatures respectively favour and disfavour the pM335 plasmid maintenance. The recombination frequency of cassettes was calculated as the ratio of recombinant CFUs (obtained on Cm- (to select cassette integration), 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 incubated *in parallel* at 30°C and at 42°C. At 30°C, all cassette integration events (in *attI1*, $attC_{aadA1}$ and in genome sites) are selected. At 42°C, integration events corresponding to integration in the recipient plasmid are counter selected and those corresponding to integration in the chromosome are therefore enriched and more easily detectable.

Cassette expression assay

To test whether *attC* cassettes can be expressed once integrated into the 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 resistance gene (*km*) without promoter

just downstream the *attC* site (pN695-pN697; Cm^R). We also added two different RBSs just upstream the *km* gene (RBS1=GGAGG, pN708-pN709 and RBS2=AGGAG, pN705-pN707, Cm^R). Once constructed, these plasmids were transformed into the β2163 donor strain. These donor strains and the recipient MG1655 strain containing the pBAD43 plasmid with or without integrase (pL294 and pL290, respectively; Sp^R) were conjugated. The recombination frequency of cassettes was calculated as the ratio of recombinant CFUs (obtained on Cm-, Sp- (resistance of the pBAD43 plasmid) and Glc-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).

Cassette excision assay

To test whether *attC* cassettes can be excised once integrated into the genome, we performed a conjugation assay proceeding exactly as described above. We constructed a 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 *ccdB* gene was previously used as a potent counterselection marker in several commonly used applications [24,25,42]. This plasmid was transformed into the β3914 donor strain to perform conjugation (a *pir*+ CcdB-resistant *E. coli* strain [24]). We also constructed a pBAD43 temperature-sensitive replicating vector expressing the integrase under the control of a P_{TET} promoter (pN435-pN440, Km^R). This plasmid was transformed into the MG1655*ΔrecA* recipient strain. Both donor and recipient strains were conjugated. Conjugation was performed at 30°C by adding aTc to express the integrase gene. We also used Glc to repress the *ccdB* toxin gene. Cassette integration events in genome were selected by plating cells on Cm- and Glc-containing plates. Then, 24 recombinants clones were randomly picked and cultivated at 42°C to ensure the loss of the thermosensitive pBAD43 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 8 h in presence of aTc, 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 clones that have lost the *ccdB* gene due to a cassette excision event are selected, while the others die. We checked the Cm sensitivity of a large number of recombinant clones.

Testing the hotspots as receptor and donor sites

To test whether the hotspots (HS), 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 HS, 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 HS, MS and US. Once constructed, each vector was transformed into the MG1655 recipient strain containing the pBAD43 with or without integrase (pL294 and pL290 respectively; Sp^R). The donor strain and these recipient MG1655 strains were conjugated. The recombination frequency of cassettes was calculated as the ratio of recombinant CFUs (obtained on Cm-, Sp- and Glc-containing plates) to the total number of recipient CFUs (obtained on Sp- and Glc-containing plates).

To test the HS as donor sites, we constructed suicide plasmid vectors containing the *ybhO*, *alsB* and *pyrE* hotspot sites delivering the bottom strand (pO323-pO324, pO749-pO750 or pO752, respectively) and the top ones (pO321-pO322, pO751 or pO753-pO755, respectively). Once constructed, these plasmids were transformed into the β2163 donor strain. These donor strains and the recipient MG1655 *recA* strain containing the pBAD43 plasmid with or without

integrase (pL294 and pL290, respectively; Sp^R) and the pSU38Δ::$attC_{aadA7}$ plasmid (pO371-pO372) were conjugated. The recombination frequency of cassettes was calculated as the ratio of recombinant CFUs (obtained on Cm-, Sp-, Km- (resistance of the pSU38 plasmid) and Glc-containing plates) to the total number of recipient CFUs (obtained on Sp-, Km- and Glc-containing plates).

Analysis of recombination events and point localization

For each experiment, clones were randomly picked and isolated on antibiotic-containing plates. Recombination events were checked by PCR using the Dream*Taq* 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 integration point (Eurofins).

Integration events in att sites carried on pSU38 vector

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) [43]. Recombination points were precisely determined by sequencing PCR products using MRV or MFD primers.

Integration events in HS, MS and US sites carried on pTOPO vector

For analysis of co-integrates formation, we performed PCR reactions on randomly chosen clones per experiment using SWbeg/MFD primers to confirm the *ybhO*, *alsB*, *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.

Integration events in att sites carried on pSC101ts vector

For analysis of co-integrates formation, we performed PCR reactions on randomly chosen clones per 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.

Integration events into bacterial genomes

For analysis of integrations of *attC*-containing pSW23T plasmids into *E. coli* genome, we performed random PCR (Extended Data Fig. 1). 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 integration. For this purpose, we used o1865 and o1388 primers. These primers respectively hybridize to the fixed part of the 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.

We also used the same procedure to detect integrations of *attC*-containing pSW23T plasmids into the *V. cholerae* genome [21]. Performing random PCR on 48 randomly chosen clones and sequencing 15 of them, we did not detect any integration events in the *attC* sites carried by the resident SCI.

Excision events on bacterial genome

For analysis of excision events from integrations of *attC*-containing pSW23T plasmids into the *E. coli* genome, we performed PCR. Note that to perform PCR, we designed appropriate primers on the basis of the knowledge on cassette recombination points during integration. We

performed PCR on clones 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 64 PCR products corresponding to clones 8 and 9.

Principles of integration profiling by deep sequencing

Library preparation

Clones obtained from three independent conjugation assays ($n = 3$) were collected and compiled before performing genomic extraction using the DNeasy Tissue Kit (Qiagen). DNA was mechanically fragmented using the Covaris method (DNA shearing with sonication). Adaptors were ligated to the fragmented DNA pieces. Nested PCR, including 30 rounds of amplification, was performed to amplify low-abundance junctions in the DNA population by first using o1366 and o6036 to amplify the cassette genome junction, and then o6035 and o6036 to reconstitute adaptors. PCR-enriched junctions were deep-sequenced using NGS technologies (Illumina MiSeq v.3, single-end, 150 cycles) (Extended Data Fig. 5).

Bioinformatics analysis

The Nextflow pipeline used to analyze the raw fastq files and generate the Fig. 4 and Extended Data Fig. 6 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 in their 5' parts, the *attC* sequence up to the G cleavage point (that is, CAATTCATTCAAGCCGACGCCGCTTCGCGGCGCGGCTTAATTCAAGC**G**) and expected by the PCR-enriched junctions described above, were selected. The *attC* sequence was trimmed such that the 5' end of reads corresponds now to the integration 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 [44]. Q20 mapped reads were selected and checked for absence of soft clipping in each extremity. In our experimental design, reads showing the same plasmid

integration site result from three non-exclusive processes: (1) same plasmid integration 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 integration was defined by the 5' extremity of forward and 3' extremity of reverse read alignments. Sequence around integration sites were extracted using bedtools (https://github.com/arq5x/bedtools2) and the nucleotide consensus of these *n* sequences were visualized with the R package ggseqlogo [45]. Random integrations were determined by deducing a sequence motif from the integration consensus and by randomly select with replacement *n* positions among all the motif positions present in the reference genome. For instance, for the IntI1 library, the obtained GWT consensus motif has been found present 338,348 times in the *E. coli* reference genome: 169,209 and 169,139 times in the forward and reverse strands, respectively. From here, we randomly selected with replacement 361,464 sites (number of integration sites after read-duplicate removal) among the 338,348 5'GWT3' sites present in the genome to use them as a "random control" of cassette integration (Fig. 4 and Extended Data Fig. 6).

Digital PCR

The absolute quantification of *ori* and *ter* was performed in multiplex by digital PCR (Stilla Technologies) and used to generate the *oriC/terC* ratio. Digital PCR was based on refs [46,47]. Genomic DNA was purified at 0, 30, 60 and 180 min from the beginning of the conjugation and after an overnight time of incubation. Note that 60 min correspond to the time frame during which most conjugation events occur.

Statistics and reproducibility

Bar charts show the mean and geometric mean (for logarithmic values) of at least three independent experiments ($n \geq 3$, individual plots). No data were excluded from the analysis. Error bars show the standard deviation or the geometric standard deviation (for logarithmic values). Means and error bars were calculated using GraphPad Prism. For comparisons of recombination assay results (with equal variance), Student's *t*-test was used. We used GraphPad Prism to determine the statistical differences between groups. Statistical comparisons are all two-sided: na, not applicable; ns, not significant; $****P < 0.0001$, $***P < 0.001$, $**P < 0.01$ and $*P < 0.05$. All exact P values for figures and extended data figures are indicated in the corresponding Source Data files.

DATA AVAILABILITY

Fastq sequences of genomic cassette integration are publicly available in NCBI SRA (accession no. SRR23447848, SRR23447849 and SRR23447850). Bacterial genomes were directly downloaded from the NCBI RefSeq database (https://ftp.ncbi.nlm.nih.gov/genomes/refseq/). Source data are provided with the paper.

CODE AVAILABILITY

All the scripts used to quantify the ability of IntegronFinder 2.0 to detect SALINs are available at https://gitlab.pasteur.fr/hub/salin. The Nextflow pipeline used to analyze the raw fastq files and generate the Fig. 4 and Extended Data Fig. 6 is available at https://gitlab.pasteur.fr/gmillot/14985_loot. The MarkDuplicates-Picard tool of GATK is available at https://github.com/broadinstitute/gatk. Bedtools used to extract sequence around integration sites is available at https://github.com/arq5x/bedtools2. IntegronFinder v.2.0.2 is available at https://github.com/gem-pasteur/Integron_Finder.

ACKNOWLEDGEMENTS

We thank G. Macaux for its experimental help; all the lab members for helpful discussion; M. Monot and L. Ma from the Biomics platform, C2RT, Institut Pasteur, Paris, France, supported by France Génomique (ANR-10-INBS-09) and IBISA. 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; D.M.), ANR Chromintevol (ANR-21-CE12-0002-01; C.L.), and by the French Government's Investissement d'Avenir program Laboratoire d'Excellence 'Integrative Biology of Emerging Infectious Diseases' (ANR-10-LABX-62-IBEID; D.M.).

AUTHOR CONTRIBUTIONS

C.L. and D.M. designed the research. C.L., E.R., C.V., B.D., D. L., V.P., F.L. and T.N. performed the experiments. G.A.M. and F. L. performed the computational analysis of Deep sequencing data. E.L., J.C., B.N. and E.P.C.R. performed the bioinformatics genomics analysis. C.L. and G.A.M. wrote the draft of the manuscript. All authors read, amended the manuscript, and approved the final version of the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

REFERENCES

1 Stokes, H. W. & Hall, R. M. A novel family of potentially mobile DNA elements encoding site-specific gene-integration functions: integrons. *Molecular Microbiology* **3**, 1669-1683 (1989).

2 Escudero, J. A., Loot, C., Nivina, A. & Mazel, D. The Integron: Adaptation On Demand. *Microbiology spectrum* **3**, MDNA3-0019-2014, doi:10.1128/microbiolspec.MDNA3-0019-2014 (2015).

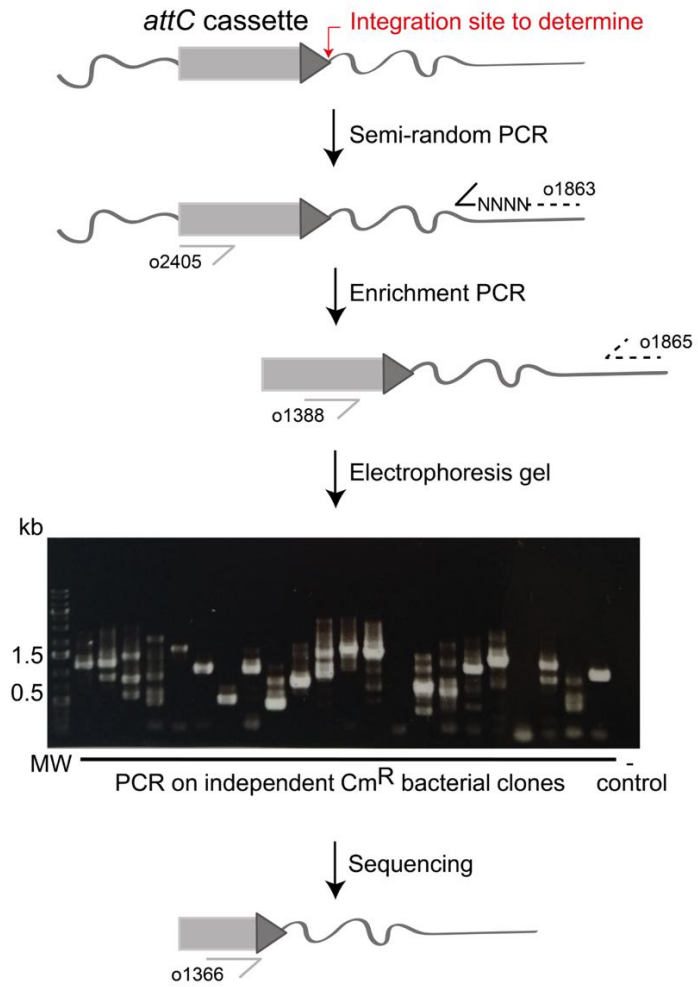
3 Cury, J., Jove, T., Touchon, M., Neron, B. & Rocha, E. P. Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucleic Acids Res* **44**, 4539-4550, doi:10.1093/nar/gkw319 (2016).

4 Neron, B. *et al.* IntegronFinder 2.0: Identification and Analysis of Integrons across Bacteria, with a Focus on Antibiotic Resistance in Klebsiella. *Microorganisms* **10**, doi:10.3390/microorganisms10040700 (2022).

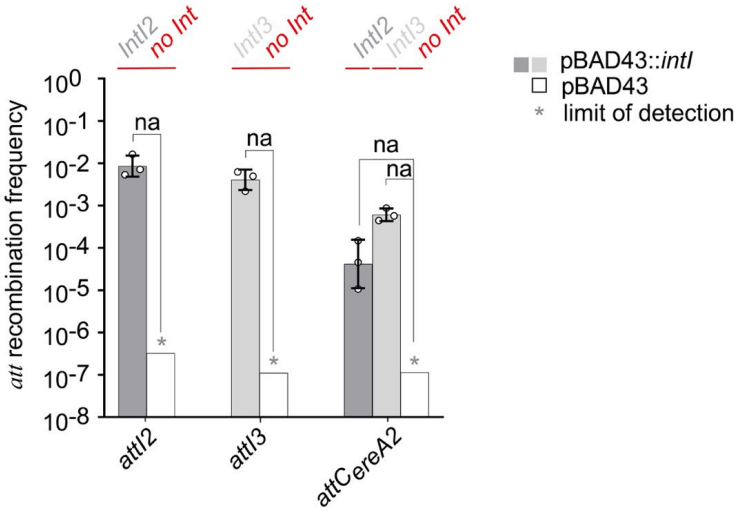
5 Buongermino Pereira, M. *et al.* A comprehensive survey of integron-associated genes present in metagenomes. *BMC Genomics* **21**, 495, doi:10.1186/s12864-020-06830-5 (2020).
6 Guerin, E. *et al.* The SOS response controls integron recombination. *Science* **324**, 1034, doi:324/5930/1034 [pii]
10.1126/science.1172914 (2009).
7 Baharoglu, Z. & Mazel, D. Vibrio cholerae triggers SOS and mutagenesis in response to a wide range of antibiotics: a route towards multiresistance. *Antimicrob Agents Chemother* **55**, 2438-2441, doi:10.1128/AAC.01549-10 (2011).
8 Richard, E., Darracq, B., Loot, C. & Mazel, D. Unbridled Integrons: A Matter of Host Factors. *Cells* **11**, doi:10.3390/cells11060925 (2022).
9 Bouvier, M., Demarre, G. & Mazel, D. Integron cassette insertion: a recombination process involving a folded single strand substrate. *Embo J* **24**, 4356-4367 (2005).
10 Francia, M. V., de la Cruz, F. & Garcia Lobo, J. M. Secondary-sites for integration mediated by the Tn21 integrase. *Molecular Microbiology* **10**, 823-828 (1993).
11 Recchia, G. D., Stokes, H. W. & Hall, R. M. Characterisation of specific and secondary recombination sites recognised by the integron DNA integrase. *Nucleic Acids Res* **22**, 2071-2078. (1994).
12 Recchia, G. D. & Hall, R. M. 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. (1995).
13 Francia, M. V. & Garcia Lobo, J. M. Gene integration in the Escherichia coli chromosome mediated by Tn21 integrase (Int21). *J Bacteriol* **178**, 894-898. (1996).
14 Segal, H. & Elisha, B. G. Identification and characterization of an aadB gene cassette at a secondary site in a plasmid from Acinetobacter. *FEMS Microbiol Lett* **153**, 321-326 (1997).
15 Segal, H., Francia, M. V., Lobo, J. M. & Elisha, G. Reconstruction of an active integron recombination site after integration of a gene cassette at a secondary site. *Antimicrob Agents Chemother* **43**, 2538-2541. (1999).
16 Souque, C., Escudero, J. A. & MacLean, R. C. Off-Target Integron Activity Leads to Rapid Plasmid Compensatory Evolution in Response to Antibiotic Selection Pressure. *mBio* **14**, e0253722, doi:10.1128/mbio.02537-22 (2023).
17 Loot, C., Bikard, D., Rachlin, A. & Mazel, D. Cellular pathways controlling integron cassette site folding. *EMBO J* **29**, 2623-2634, doi:10.1038/emboj.2010.151 (2010).
18 Vit, C. *et al.* Cassette recruitment in the chromosomal Integron of Vibrio cholerae. *Nucleic Acids Res* **49**, 5654-5670, doi:10.1093/nar/gkab412 (2021).
19 Jove, T., Da Re, S., Denis, F., Mazel, D. & Ploy, M. C. Inverse correlation between promoter strength and excision activity in class 1 integrons. *PLoS Genet* **6**, e1000793, doi:10.1371/journal.pgen.1000793 (2010).
20 Nivina, A., Escudero, J. A., Vit, C., Mazel, D. & Loot, C. 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, doi:10.1093/nar/gkw646 (2016).
21 Mazel, D., Dychinco, B., Webb, V. A. & Davies, J. A distinctive class of integron in the *Vibrio cholerae* genome. *Science* **280**, 605-608 (1998).
22 Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119, doi:10.1186/1471-2105-11-119 (2010).
23 Omotajo, D., Tate, T., Cho, H. & Choudhary, M. Distribution and diversity of ribosome binding sites in prokaryotic genomes. *BMC Genomics* **16**, 604, doi:10.1186/s12864-015-1808-6 (2015).

24 Le Roux, F., Binesse, J., Saulnier, D. & Mazel, D. 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 (2007).
25 Betton, J. M. Cloning vectors for expression-PCR products. *Biotechniques* **37**, 346-347 (2004).
26 Rousset, F. *et al.* Genome-wide CRISPR-dCas9 screens in E. coli identify essential genes and phage host factors. *PLoS Genet* **14**, e1007749, doi:10.1371/journal.pgen.1007749 (2018).
27 Loot, C. *et al.* 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, doi:10.1128/JB.01109-13 (2014).
28 Loot, C. *et al.* Differences in Integron Cassette Excision Dynamics Shape a Trade-Off between Evolvability and Genetic Capacitance. *mBio* **8**, doi:10.1128/mBio.02296-16 (2017).
29 Rutkai, E., Dorgai, L., Sirot, R., Yagil, E. & Weisberg, R. A. Analysis of insertion into secondary attachment sites by phage lambda and by int mutants with altered recombination specificity. *J Mol Biol* **329**, 983-996, doi:10.1016/s0022-2836(03)00442-x (2003).
30 Tanouchi, Y. & Covert, M. W. Combining Comprehensive Analysis of Off-Site Lambda Phage Integration with a CRISPR-Based Means of Characterizing Downstream Physiology. *mBio* **8**, doi:10.1128/mBio.01038-17 (2017).
31 Crozat, E., Fournes, F., Cornet, F., Hallet, B. & Rousseau, P. Resolution of Multimeric Forms of Circular Plasmids and Chromosomes. *Microbiology spectrum* **2**, doi:10.1128/microbiolspec.PLAS-0025-2014 (2014).
32 Kluyver, T. *et al.* Jupyter Notebooks -- a publishing format for reproducible computational workflows. *In F. Loizides & B. Schmidt (Eds.), Positioning and Power in Academic Publishing: Players, Agents and Agendas.*, 87 - 90 (2016).
33 McKinney, W. Data Structures for Statistical Computing in Python. *In Proceedings of the 9th Python in Science Conference* **445**, 51-56 (2010).
34 Cock, P. J. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422-1423, doi:10.1093/bioinformatics/btp163 (2009).
35 Waskom, M. L. seaborn: statistical data visualization. *Journal of Open Source Software* **6**, 3021, doi:10.21105/joss.03021 (2021).
36 Hug, L. A. *et al.* A new view of the tree of life. *Nat Microbiol* **1**, 16048, doi:10.1038/nmicrobiol.2016.48 (2016).
37 Biskri, L., Bouvier, M., Guerout, A. M., Boisnard, S. & Mazel, D. Comparative Study of Class 1 Integron and *Vibrio cholerae* Superintegron Integrase Activities. *J Bacteriol* **187**, 1740-1750 (2005).
38 Demarre, G., Frumerie, C., Gopaul, D. N. & Mazel, D. Identification of key structural determinants of the IntI1 integron integrase that influence attC x attI1 recombination efficiency. *Nucleic Acids Res* **35**, 6475-6489 (2007).
39 Jaskolska, M., Adams, D. W. & Blokesch, M. Two defence systems eliminate plasmids from seventh pandemic Vibrio cholerae. *Nature* **604**, 323-329, doi:10.1038/s41586-022-04546-y (2022).
40 Demarre, G. *et al.* A new family of mobilizable suicide plasmids based on the broad host range R388 plasmid (IncW) or RP4 plasmid (IncPα) conjugative machineries and their cognate *E.coli* host strains. *Research in Microbiology* **156**, 245-255 (2005).

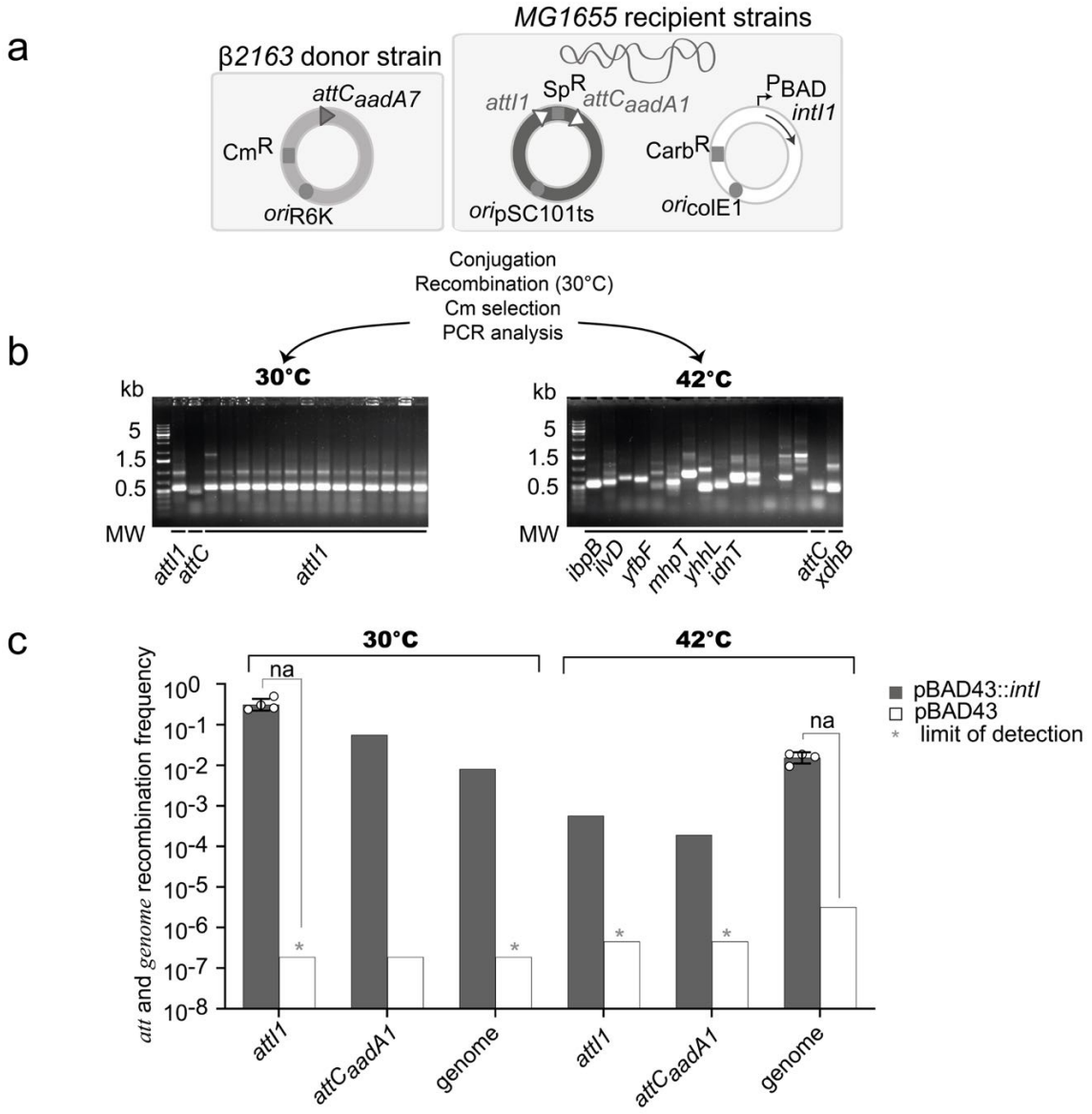
41 Guzman, L. M., Belin, D., Carson, M. J. & Beckwith, J. Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. *J Bacteriol* **177**, 4121-4130. (1995).
42 Val, M. E., Skovgaard, O., Ducos-Galand, M., Bland, M. J. & Mazel, D. Genome engineering in Vibrio cholerae: a feasible approach to address biological issues. *PLoS Genet* **8**, e1002472, doi:10.1371/journal.pgen.1002472 (2012).
43 Bouvier, M., Ducos-Galand, M., Loot, C., Bikard, D. & Mazel, D. Structural features of single-stranded integron cassette attC sites and their role in strand selection. *PLoS Genet* **5**, e1000632, doi:10.1371/journal.pgen.1000632 (2009).
44 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).
45 Wagih, O. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* **33**, 3645-3647, doi:10.1093/bioinformatics/btx469 (2017).
46 de Lemos Martins, F., Fournes, F., Mazzuoli, M. V., Mazel, D. & Val, M. E. Vibrio cholerae chromosome 2 copy number is controlled by the methylation-independent binding of its monomeric initiator to the chromosome 1 crtS site. *Nucleic Acids Res* **46**, 10145-10156, doi:10.1093/nar/gky790 (2018).
47 Madic, J. *et al.* Three-color crystal digital PCR. *Biomol Detect Quantif* **10**, 34-46, doi:10.1016/j.bdq.2016.10.002 (2016).



Extended Data Fig. 1: Random PCR approach used to determine the genome integration sites
The genome integrated *attC* cassette is represented by a light grey arrow (coding sequence) followed by a dark grey triangle (*attC*). The red arrow indicates the integration site to determine. Primers used for the PCR amplification and sequencing are shown. Electrophoresis analysis of PCR products is shown. MW: Molecular Weight Marker, kb: kilobases.



Extended Data Fig. 2: Cassette recombination events mediated by IntI2 and IntI3 integrases
The graph represents the recombination frequencies of $attC_{aadA7}$ donor sites into *attI* and $attC_{ereA2}$ sites mediated by IntI2 and IntI3 integrases. The expressed integrases (IntI2 or IntI3) are indicated above the graph and the receptor sites in the axis-x legends. Bar charts show the mean of three independent experiments (n = 3, individual plots) and error bars show the standard deviation. Statistical comparisons (Student's t test) are as follow: na. (not applicable);
Grey asterisk (*) indicates the recombination frequency was below detection level.



Extended Data Fig. 3: Cassette integration events in genome in presence of an integron

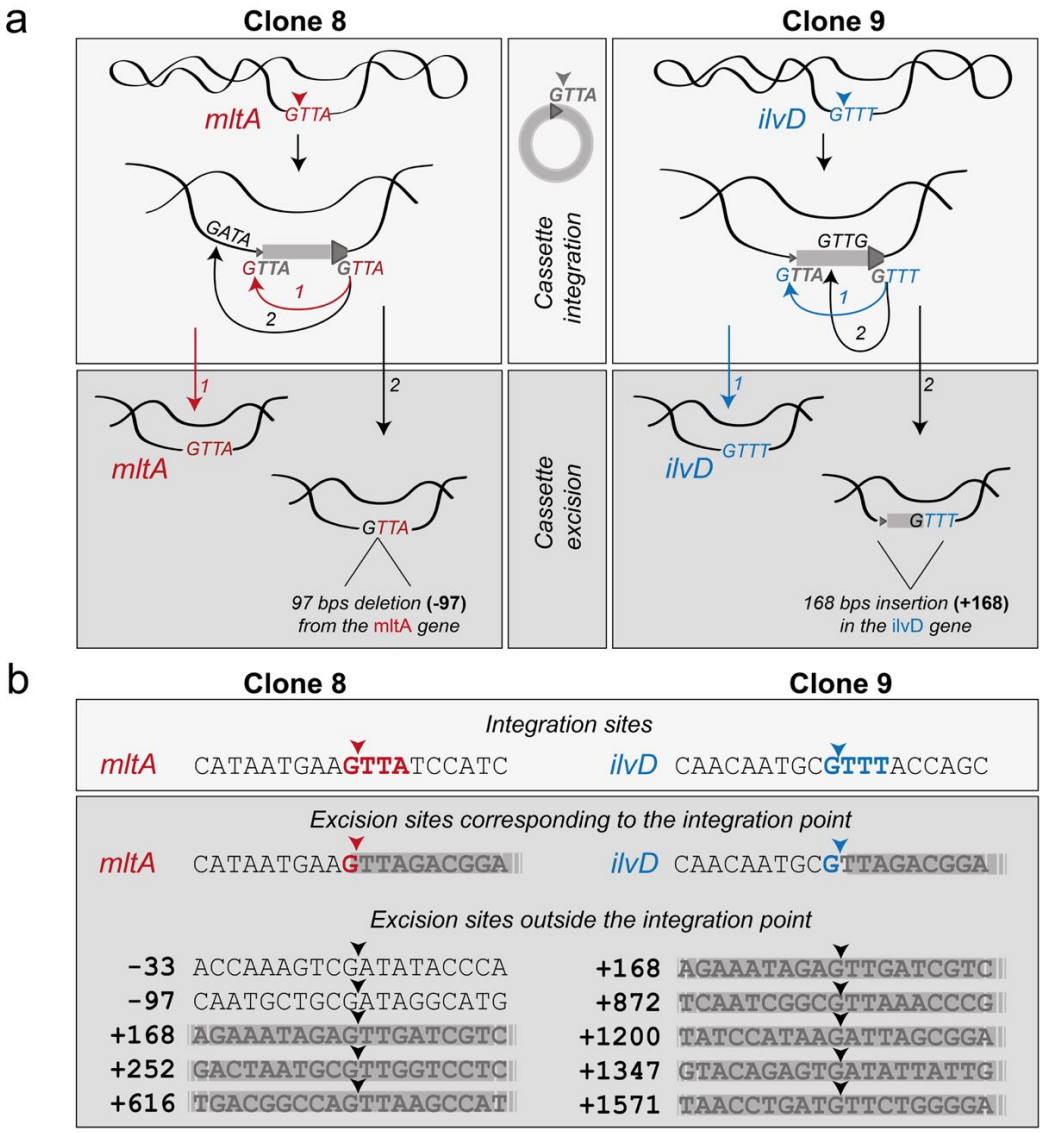
a) Experimental setup of the cassette integration assay (as in Fig. 3a)

b) Electrophoresis migration of PCR products

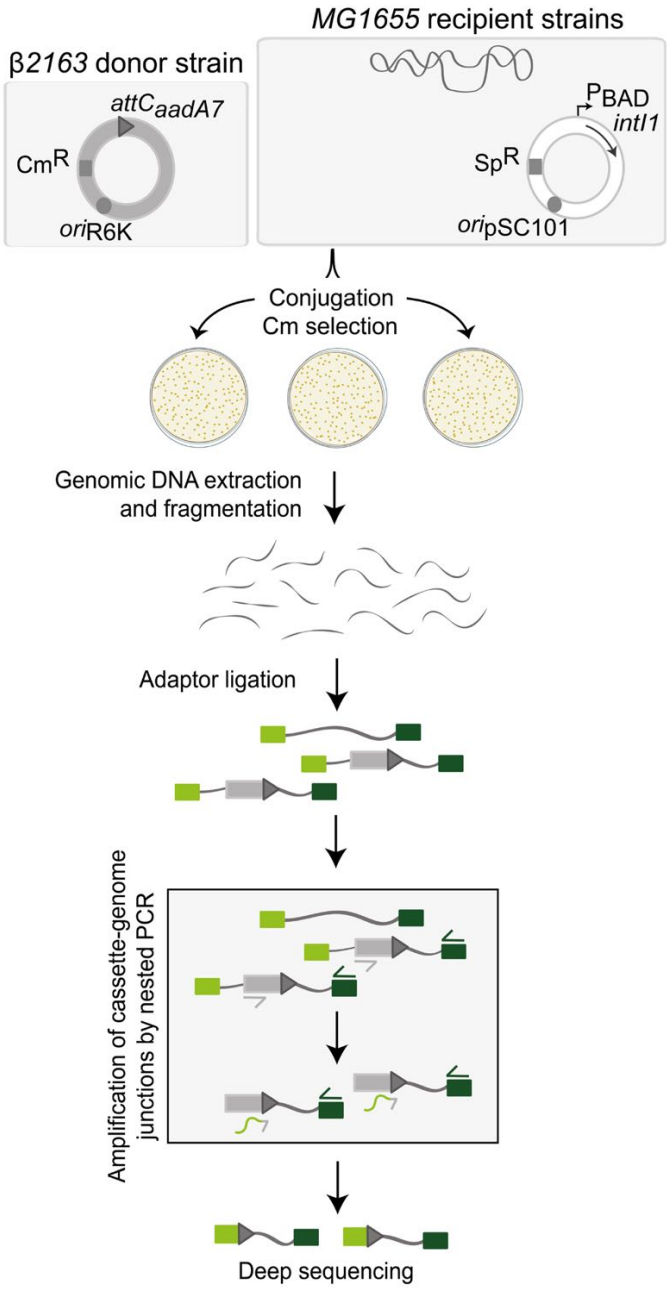
Several PCR products are sequenced and the integration sites are indicated (*att* sites and gene location).

c) Recombination frequencies of *attCaadA7* donor sites into the *att* sites and genome

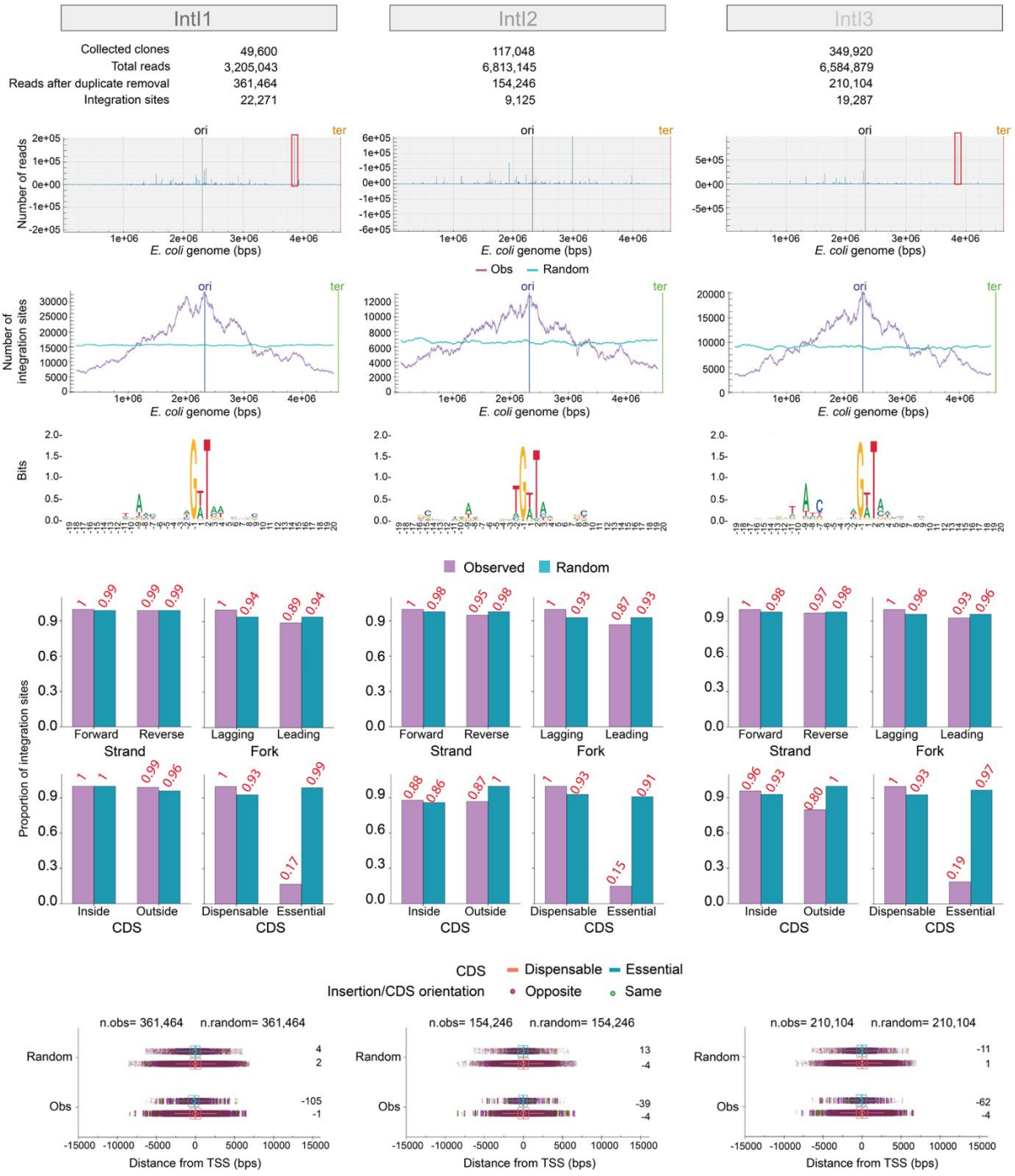
Integration sites are indicated in the axis-x legends. Bar charts show the mean of three independent experiments (n=3, individual plots) and error bars show the standard deviation. Statistical comparisons (Student's t test) are as follow: na. (not applicable); Grey asterisk (*) indicates the recombination frequency was below detection level. ts: thermosensitive; MW: Molecular Weight Marker, kb: kilobases.



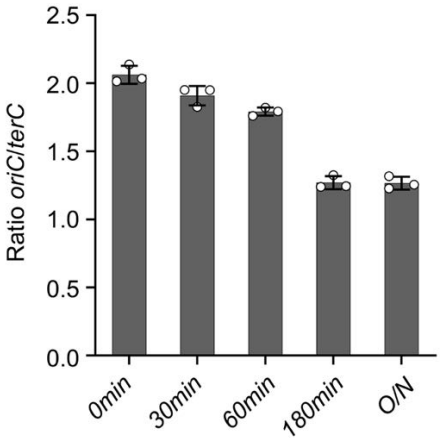
Extended Data Fig. 4: Excision events observed for two genome-integrated cassettes
a) Excisions occurring at the integration point (1) and reconstituting the *mltA* and *ilvD* genes are shown. Excisions occurring outside the integration point (2) inducing either genome deletion (excision point located in the genome) or insertion (excision point located in the integrated cassette) are shown.
b) Sequences of the *mltA* and *ilvD* integration sites and of excision sites are shown. All the sequenced excision events (*i.e.* 64) are shown. Sizes in bps of each induced genome deletion (-) and insertion (+) are indicated. Cassette sequences are highlighted in grey and cleavage points are indicated by head of arrows. bps: base pairs.



Extended Data Fig. 5: Library construction for Deep sequencing
All the steps of the library construction are indicated. Adaptors are represented by dark and light green rectangles. *attC* cassettes are represented by light grey rectangles (coding sequence) followed by dark grey triangles (*attC*).

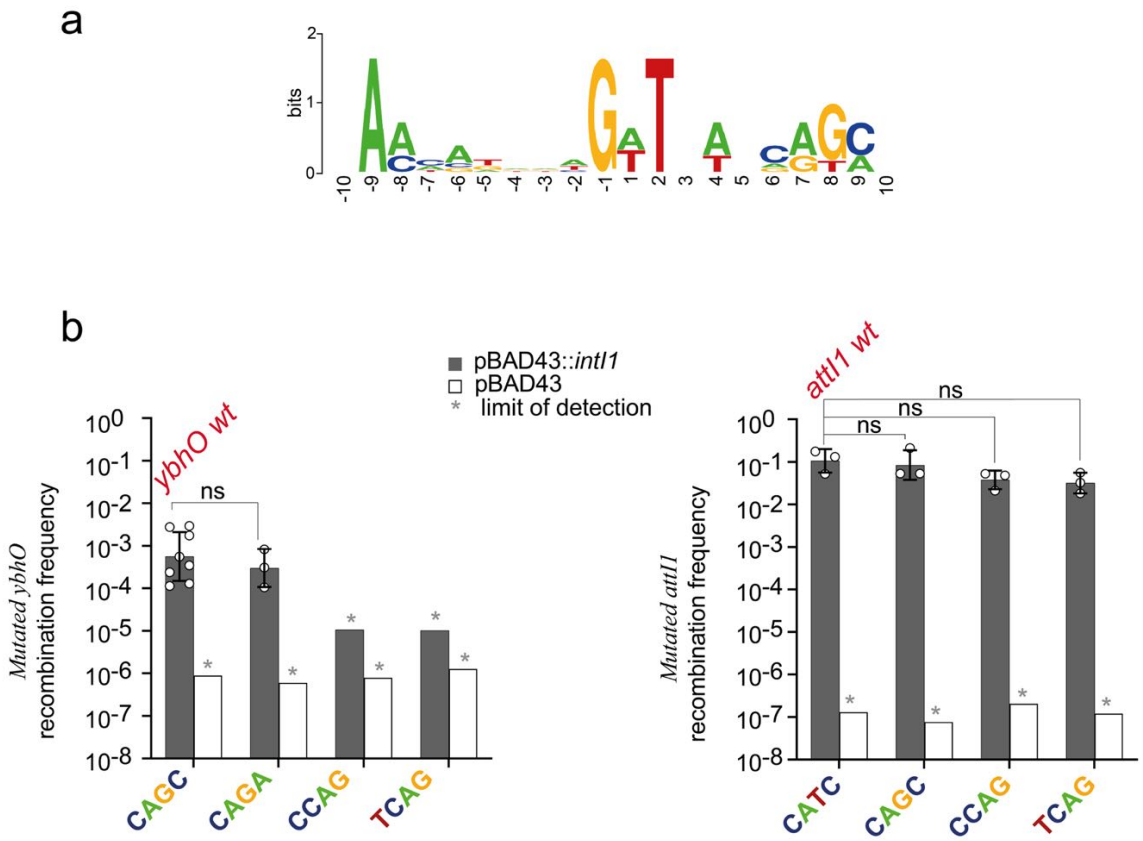


Extended Data Fig. 6: Computational analysis of Deep sequencing data
The figure shows the previous results obtained with IntI1 grouped with the new results obtained with IntI2 and IntI3. The legend is the same as in Fig. 4. The red boxes show the *ybhO* hotspot sites.



Extended Data Fig. 7: Digital PCR analysis of the *oriC* copy number relative to *terC* in *E. coli* recepient strain during a conjugation mimicking assay

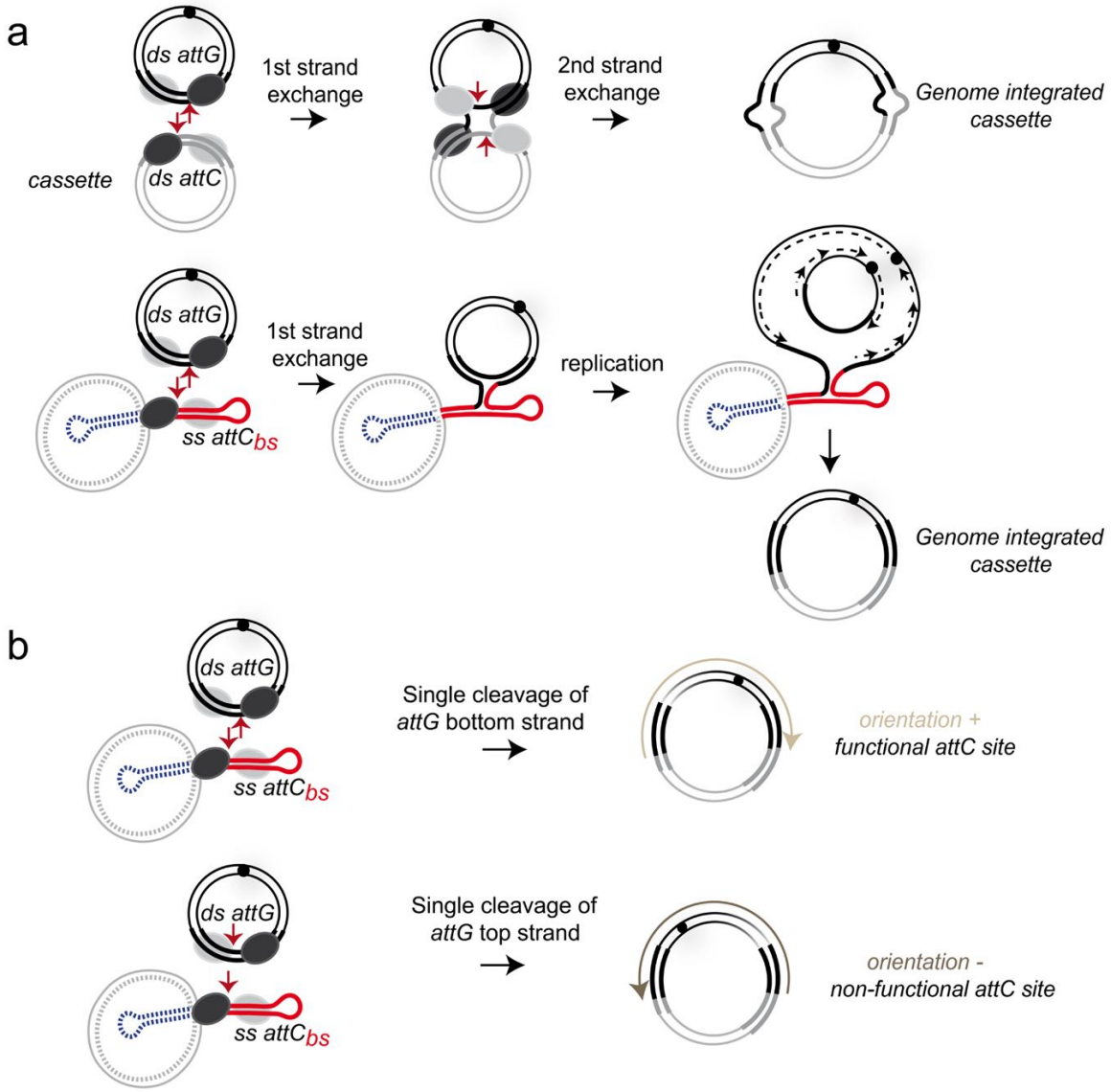
DNA extraction and PCR were performed at 0, 30, 60, 180 minutes (min) and at an overnight (O/N) time of incubation after the beginning of the conjugation. Bar charts show the mean of three independent experiments (n=3, individual plots) and error bars show the standard deviation. Statistical comparisons (Student's t test) are as follow: na. (not applicable); Grey asterisk (*) indicates the recombination frequency was below detection level.



Extended Data Fig. 8: Analysis of the *attG* integration site motifs

a) Consensus sequence of the six highest hotspot integration sites.
A total of 20 bases around the cleavage point was inputted into the WebLogo program (http://weblogo.berkeley.edu/) to generate the motif. The cleavage occurs between the -1 and 1 bases. Bits refers to the information content.

b) Recombination frequencies of *attCaadA7* donor sites into the mutated *ybhO* hotspot and *attI1* sites. The 4 nucleotide sequences indicated in the axis-x legends correspond to the position 6 to 9 given that position -1 corresponds to the G of the cleaved triplet site. The *wt ybhO* and *attI1* sites are indicated in red at the top of the bars. Bar charts show the mean of at least three independent experiments (n≥3, individual plots) and error bars show the standard deviation. Statistical comparisons (Student's t test) are as follow: ns. (not significant); all two-sided. Grey asterisk (*) indicates the recombination frequency was below detection level.



Extended Data Fig. 9: Determination of the *attG* recombination nature
a) Single or double cleavage of *attG* sites?
Double cleavage: we expect heterogeneity of products due to the repair process.
Single cleavage: we expect homogeneity of products due to the replication process.
b) Bottom or top strand cleavage of *attG* sites?
Bottom strand cleavage: we expect an orientation + (light brown arrow) and a functional *attC* site.
Top strand cleavage: we expect an orientation - (dark brown arrow) and a non-functional *attC* site.
attG containing replicons are represented by dark lines and *attC* cassettes by grey ones. The precise cleavage point is indicated by a red arrow. ss and ds: single and double-stranded; bs: bottom strand.