Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/6/30/eaay2922/DC1

Supplementary Materials for

Structure-specific DNA recombination sites: Design, validation, and machine learning–based refinement

Aleksandra Nivina, Maj Svea Grieb, Céline Loot, David Bikard, Jean Cury,
Laila Shehata, Juliana Bernardes*, Didier Mazel*

*Corresponding author. Email: juliana.silva_bernardes@upmc.fr (J.B.); didier.mazel@pasteur.fr (D.M.)

Published 24 July 2020, *Sci. Adv.* **6**, eaay2922 (2020)
DOI: 10.1126/sciadv.aay2922

This PDF file includes:

Supplementary Materials

Strains, media and antibiotic concentrations used in this study

Bacterial strains used in this study are *Escherichia coli* DH5α, MG1656 (*47*), β2163 (*15*) and DHP1(*18*). They were grown in Luria-Bertani (LB) broth at 37°C. Antibiotics were used at the following concentrations: chloramphenicol (Cm), 25 µg/ml; kanamycin (Kan), 25 µg/ml; carbenicillin (Carb), 100 µg/ml. Diaminopimelic acid (DAP) was supplemented when necessary to a final concentration of 0.3 mM. To induce the pBAD promoter, 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. Isopropyl-β-D-thiogalactopyranoside (IPTG) and 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal) were supplemented at 1 mM and 40 µg/ml, respectively.

Protocol of *attC* site recombination assay

This protocol was adapted from (*6*) and (*8*). This conjugative assay consists of delivering a single strand of a pSW23T plasmid from a donor *E.coli* ß2163 strain into a recipient *E. coli* DH5α strain that expresses IntI1 and contains a plasmid carrying an *attI1* site. The pSW23T plasmid cannot be maintained in the recipient cell due to the absence of the π protein, but the delivered strand carries a resistance marker, as well as the bottom strand of the *attC* site, which can be recombined with an *attI* site. This assay is designed to measure the frequency of *attI1*×*attC* recombination 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 recombination frequency is thus dependent on the reactivity of the *attC* site.

For testing recombination frequency of synthetic *attC* sites embedded into *lacZ*, a variant of this protocol was used, where the *attI* site is delivered from the donor strain through conjugation and the *attC* sites are carried by the recipient strain.

The donor strains were grown overnight in LB media supplemented with Cm, Kan and DAP; the recipient strain was grown overnight in LB media supplemented with Carb, Kan and Glc. Both overnight cultures were diluted 1/100 in LB with Kan+DAP or Kan+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 resuspended in 50µl 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 overnight 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. 100µl of non-diluted sample as well as 10^{-1}, 10^{-2} and 10^{-3} dilutions were plated on plates with LB+agarose media supplemented with Cm. 100µl of 10^{-3} , 10^{-4} and 10^{-5} dilutions were plated on plates with LB+agarose media supplemented with Carb, Kan and Glc. Plates were incubated at 37°C for approximately 40 hours. For each strain, 3 independent recombination tests were performed, and at least 8 colonies per test were subject to PCR with primers SWbeg and MFD to measure the true positive rate (Table S2). The recombination frequency was calculated as the ratio of recombinant CFUs [CmR] to the total number of recipient CFUs [CarbR KanR], multiplied by the true positive rate. For each strain, the overall recombination frequency is a mean of 3 independent experiments, and error bars represent the mean deviation.

For negative controls where no recombinant CFUs were detected, the level of detection was calculated as one divided by the number of recipient CFUs. The overall level of detection is a mean of all independent experiments, and an asterisk indicates that the value represents the maximum level of detection, not the actual recombination frequency.

Approach for generating synthetic *attC* sites encoding peptide linkers

To generate synthetic *attC* sites encoding peptide linkers, we used an *in silico* directed evolution approach (Fig. 3A). Briefly, a set of 1000 synthetic *attC* sites was generated using the first version of the algorithm as described above (Fig.2A) and submitted to 100 rounds of mutation and selection. At each round, mutations were introduced into random positions of each *attC* site, except in "unmutable" positions which included the R box and the EHBs. In case a mutation was introduced into a position required to be paired in the *attC* site structure (any "mutable" position except within the UCS or VTS), a complementary mutation was introduced into the sequence in order to reconstitute pairing. Then, each site was translated in all three possible reading frames and aligned to the linker database (http://www.ibi.vu.nl/programs/linkerdbwww/) (*17*) using the Smith-Waterman algorithm. Out of three alignments, the best score was used to eliminate 50 lower-scoring sites, after which the remaining 50 sites were duplicated and propagated to the next round of directed evolution. This algorithm was run three times and each time the synthetic *attC* site with best similarity score to any linker within the database was chosen for further experiments, resulting in $attC_{L1}$, $attC_{L2}$ and $attC_{L3}$.

Protocol of the ELISA-based linker assay

This protocol was adapted from (*18*) and (*19*). Briefly, a cAMP-biotinylated-BSA conjugate was coated on ELISA plates, and nonspecific protein-binding sites were blocked with BSA. Boiled bacterial cultures were then added, followed by diluted rabbit anti-cAMP antiserum in 50 mM Hepes, pH 7.5, 150 mM NaCl, 0.1% Tween 20 (HBST buffer) containing 10 mg/ml BSA. After overnight incubation at 4°C, the plates were washed extensively with HBST, then goat anti-rabbit IgG coupled to alkaline phosphatase (AP) was added and incubated for 1 hr at 30°C. After washing, the AP activity was revealed by 5-para-nitrophenyl phosphate. cAMP concentrations were calculated from a standard curve established with known concentrations of cAMP diluted in LB. For each strain, the overall cAMP concentration is a mean of 3 independent experiments, and error bars represent the standard deviation.

Approach for generating synthetic *attC* sites embedded into a protein of choice

To generate synthetic *attC* sites embedded into a protein of choice, we used an *in silico* directed evolution approach (Fig. 3A) and chose β-galactosidase as target protein. Three 72nt locations within *lacZ* gene were chosen such that recombination would occur after a Valine (encoded by GTT, but this additional constraint is not required). In each case, a set of 1000 synthetic *attC* sites was generated using the first version of the algorithm as described above (Fig.2A), with a variable VTS size allowed. They were submitted to 100 rounds of mutation and selection. At each round, mutations were introduced into random positions of each *attC* site, except in "unmutable" positions which included the R box and the EHBs. In case a mutation was introduced into a position required to be paired in the *attC* site structure (any "mutable" position except within the UCS or VTS), a complementary mutation was introduced into the sequence in order to reconstitute pairing. Then, both arms of the site (separated by the VTS) were translated in all three possible reading frames using the Smith-Waterman algorithm and aligned against the desired protein. The right arm of the site was aligned against the protein region chosen to contain the recombination site, whereas the left arm was aligned against the rest of the protein. The best score for the combination of two alignments was used to eliminate 50 lower-scoring sites, after which the remaining 50 sites were duplicated and propagated to the next round of directed evolution. The test was run five times and synthetic *attC* site with the best similarity score to the chosen protein region was reported.

Protocol of the *lacZ* test

To test whether β-galactosidase with embedded synthetic *attC* sites was functional, we constructed pSU::*pLac-lacZ*:*attC* plasmids. Wild-type *lacZ* gene was used as positive control. Because we expect that significant changes to *lacZ* sequence would cause the protein to become non-functional, we replaced a region of *lacZ* with $attC_{aadA7}$ sequence as a negative control. Oligonucleotides encoding *lacZ* regions with embedded *attC* sites (Supplementary Table S2) were annealed and inserted into the pSU::*lacZ* vector, amplified with corresponding oligonucleotides (Supplementary Table S2). Plasmids were transformed into *E.coli* MG1656 strain. For each construct, one clone was confirmed through sequencing and streaked on LB+agarose+Kan+IPTG+X-gal plate. The color of the clone was appreciated by naked eye, blue color signifying that β-galactosidase was functional (like for positive control) and white color signifying that β-galactosidase was not functional (like for negative control).

Protocol of the library competition (enrichment) assay

The library in ß2163 strain was grown overnight in 100ml LB media supplemented with Chloramphenicol (Cm), Kanamycin (Kan) and DAP; the recipient strain 9669 was grown overnight in 5ml LB media supplemented with Carbenicillin (Carb), Kan and Glucose (Glc). Both overnight cultures were diluted 1/100 in 10ml LB with Kan+DAP or Kan+Arabinose (Ara) respectively, and incubated until OD=0.4-0.5. They were then mixed and centrifuged at 5000rpm for 10mins. The pellet was resuspended in 3ml LB, plated on a 100mm LB+agarose+DAP+Ara Petri dish and incubated overnight for conjugation and recombination to take place. The plate was then scraped, and the collected culture resuspended in 5ml LB. 2ml of this suspension was added to 100ml LB containing Glc and a 10-fold concentration of Cm, and grown for 5 hours to exert an initial selection on recombinants. 1ml of this culture was then diluted 1/10 in LB, plated on 10 100mm LB+agarose+Cm+Glc Petri dishes and incubated overnight for further selection. The plates were scraped, and the collected culture resuspended in 50ml LB and vortexed. 200µl of this suspension was diluted in 100ml LB+Cm+Glc and incubated overnight to produce a liquid culture of selected

recombinants. This constituted a library of *attI*×*attC* recombinants that contained a mix of *attC* sites from the library according to their recombination efficiency.

DNA was then extracted from this liquid culture using the Thermo Fisher Scientific GeneJET Plasmid Miniprep kit. *attC* sites were amplified with Gibson1 and Gibson2 primers. The pSW23T vector was amplified with Gibson3 and Gibson4 primers. The two products were then purified, joint together through Gibson assembly (*46*) and transformed into the ß2163 strain (*15*). The transformants were plated on 10 100ml LB+Cm+Kan+DAP Petri dishes, incubated overnight, scraped, resuspended in 100ml LB+DAP and vortexed. 200µl of this suspension was diluted in 100ml LB+Cm+Kan+DAP and incubated overnight to produce a liquid culture of the *attC* site library selection, which contained *attC* sites according to their recombination efficiency, and was used for NGS. It was also used as a starting culture for further cycles of selection, each of them repeating all the above-mentioned steps.

All liquid cultures were performed at 37°C with shaking; all incubation steps were performed at 37°C without shaking.

Value calculations for global features used in ML algorithm

Predictions of *attC* site folding were made using RNAfold program from ViennaRNA2.1.8 package (*24*), using the -p option to compute the partition function. All values were calculated based on the output provided by RNAfold.

The Gibbs free energy (ΔG) of the thermodynamic ensemble of folded molecules, the diversity of this ensemble, the ΔG of the Minimal Free Energy (MFE) structure and the frequency of MFE structure in the ensemble were direct outputs of the RNAfold program.

The number of Hydrogen bonds in the MFE structure, the number of non-Watson-Crick G:T pairings in the MFE structure, the difference and the ratio of the ΔG of bottom and top strand MFEs were derived from the outputs of the RNAfold program.

To calculate the probability to fold a functional structure (pfold), i.e. a structure with correctly folded integrase binding sites (*25*, *26*), we performed folding predictions using the -C option to add a constraint of pairing the R and L boxes, and calculated the pfold values as follows:

$$pfold = e^{\frac{\Delta G_u - \Delta G_c}{RT}} \qquad (1)$$

where ΔG_u is the Gibbs free energy of the unconstrained thermodynamic ensemble (kcal), ΔG_c is the Gibbs free energy of the constrained thermodynamic ensemble (kcal), R is the gas constant (kcal K^{-1} mol^{-1}) and T is the temperature (K).

Value calculations for base-specific features used in ML algorithm

To obtain the positional entropy and the pairing probability for each base, as well as the pairwise base-pairing probabilities, we used the relplot program from the ViennaRNA2.1.8 package. We also extracted pairwise base-pairing probabilities for any taken base with the base located opposite it in the hairpin, and its immediate neighbors. We did not include all pairwise base-pairing probabilities as features, since there are more than 2,000 such features, and their values were nil or almost nil for most *attC* sites. This created a sparse matrix of input values, which was detrimental for the ML algorithms, and we decided to include only a set of base-pairs, since they presented probability values significantly differed from 0 for least of a number of *attC* sites.

We also described the sequences of $attC_{r0}$ mutants, namely the nature of the base at each position (A, T, G or C), as a discrete value.

Normalization and balancing of the final non-sparse data matrix

To ensure that the input data matrix was not sparse, all features with zero variance (total of 14 features) were eliminated from the analysis. The remaining feature values were then normalized to be comprised in the interval [0, 1], either through linear or logarithmic normalization.

We assigned non-nil values to data points for which there were no reads detected in the library after recombination. These data points corresponded to *attC* sites that were completely depleted from the library after just one cycle of selection. For these data points, we calculated the upper limit of the possible enrichment value by using 1 instead of 0 as the number of reads in the library after selection.

To normalize the enrichment values, we defined the threshold between the enriched and the depleted mutants at 1. At this value, the occurrence of a mutant before and after selection did not change. All mutants with higher enrichment values were considered to be enriched; all others were considered to be depleted. We then normalized all the enrichment values on a logarithmic scale, for them to be comprised in the interval [0, 1].

The dataset was equilibrated by selecting all enriched data points (1,762) and randomly selecting an equal number among 11,117 depleted mutants, resulting in a dataset of 3,524 data points.

Performance measures used to evaluate ML algorithms.

To measure the performance of regression models, we calculated four different measures: the Pearson correlation coefficient, the mean absolute error, the root mean square error and the explained variance score. All measures take as inputs the enrichment values (as a proxy for the actual recombination frequencies), and the predicted values.

The Pearson Correlation Coefficient (PCC) is defined as:

$$PCC = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_i)(z_i - \bar{z}_i)}{\sqrt{\left[\sum_{i=1}^{n}(y_i - \bar{y}_i)^2\right]\left[\sum_{i=1}^{n}(z_i - \bar{z}_i)^2\right]}} \qquad (2)$$

where y_i and z_i are the actual and the predicted enrichment values respectively, $\bar{y}_i$ and $\bar{z}_i$ are their corresponding means, and n is the total of data points. PCC=1 indicates that the two sets of values are fully correlated, while PCC=0 indicates that they are completely uncorrelated.

The Mean Absolute Error (MAE) is defined as the average difference between the actual and the predicted enrichment values of all data points:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}(|y_i - z_i|) \qquad (3)$$

The Root Mean Square Error (RMSE) is the square root of variance of the residuals (predicted minus actual value), defined as:

$$RMSE = \frac{1}{n}\sum_{i=1}^{n}\sqrt{(y_i - z_i)^2} \qquad (4)$$

The Explained Variance Score (VarScore) is defined as:

$$varScore = \frac{1}{n}\sum_{i=1}^{n}\left[1 - \frac{Var(y_i - z_i)}{Var(y_i)}\right] \qquad (5)$$

where Var is the variance of each distribution.

Feature selection and dimensionality reduction methods used for three ML algorithms

Reducing the dimensionality of the feature space or selecting the most relevant set of features can improve the performance of ML algorithms. Here, we tried both strategies and used Principal component analysis (PCA) (*33*) to dimensional reduction and k-best-features (*34*) to select the k most relevant features.

PCA has been widely applied in data mining and pattern recognition. It transforms the existing features into a lower dimensional space, where new orthogonal variables (principal components) are obtained by maximizing the variance of the data. PCA greatly reduces the dimensionality of the space, but it does not reduce the number of the original variables, all original variables are used to generate new ones (principal components). We trained SML algorithms by using a reduced set of features that consider just the k-significant principal components obtained by PCA, k being in the interval [2, 50]. We used the PCA algorithm available in the package "decomposition" of sklearn library (*31*).

Feature selection methods do not combine variables as PCA does, they just evaluate the quality and the predictive power of each feature to select the best set. Among a number of feature selection methods available, we chose the k-best-features method that selects the k most relevant features based on univariate statistical tests. First, the statistical test is computed between each feature and the output value, then the k features with higher score values are selected. Here, we used the chi-squared test (*48*) that measures dependence between two variables, so using this test we can detect features that are the most likely to be dependent on the output, by consequence the more relevant for constructing predictive models. Next, we have trained regression models using just the k-best features, where k is in the interval [2, 50]. We used the k-best-features algorithm available in the package feature_selection of sklearn library (*31*).

We also performed a manual feature selection by organizing the features into lists according to their properties, see Lists A-D in Table 1, and trained regression models with all possible list combinations.

Analysis of features with an importance score >0.01 in Random Forest Regression

Pairing of the bases within the stem represented one set of base-specific features with an importance score >0.01 (Fig. 5D, ovals). To understand how such base-pairings influenced recombination, we decided to see whether the correlation of the base-pairing probabilities with the measured enrichment value was positive (red) or negative (blue), and what was the correlation coefficient (Fig. S4). For the purpose of easier interpretation, we visualized the base-pairing probabilities on two separate maps: the first one corresponding to expected base-pairings (Fig. S4A), and the second one corresponding to base-pairings that were not expected according to the structural prediction (Fig. S4B).

Most expected base-pairing probabilities showed a strong positive correlation with the measured enrichment value (Fig. S4A, red contours), coherent with previous results showing the importance of the stem for integrase binding and recombination (*6*, *13*). However, two base-pairing probabilities for bases located in the second region of interest (T24-A38 and T25-A37) correlated

negatively with the enrichment values (Fig. S4A, blue contours), meaning that the mutants where these base-pairings were preserved showed lower recombination propensity and got depleted.

Among the base-pairings that were not expected according to the structural prediction, most indeed showed negative correlation with the enrichment values (Fig. S4B, blue contours). However, regions around both EHBs contained base-pairs that correlated positively (Fig. S4B, red contours). This was particularly striking since, according to the design of $attC_{r0}$, these were not supposed to be paired. Our analysis showed that recombination was improved in $attC_{r0}$ mutants where the EHBs were "shifted" by one nucleotide towards the apex of the stem: EHB G45 instead of G46 (with G46 becoming paired with the base in position 18) and EHB T38 instead of T39 (with T39 becoming paired with the base in position 24). This explained the positive correlations of these base-pairings with enrichment.

Positional entropies represented another set of important features (Fig. 5D, asterisks). The positional entropy of a base reflects how unstable it is: low positional entropy means that the base is stabilized in only one major state (either paired with a particular base, or unpaired), whereas high positional entropy means the base can be found in various states within the thermodynamic ensemble of possible structures. The positional entropies identified as important features all correlated negatively with the enrichment values, suggesting that a stabilized state of the stem is more favorable for recombination (Fig. 5D and Fig. S4).

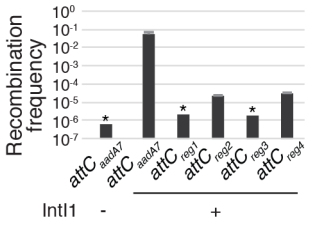

Fig. S1. Recombination frequencies of *attC*$_{aadA7}$ and synthetic *attC* sites embedded into *lacZ*.

Values represent the mean of three independent experiments; error bars represent mean absolute error. Asterisks (*) indicate that the recombination frequency was below detection level, indicated by the bar height.

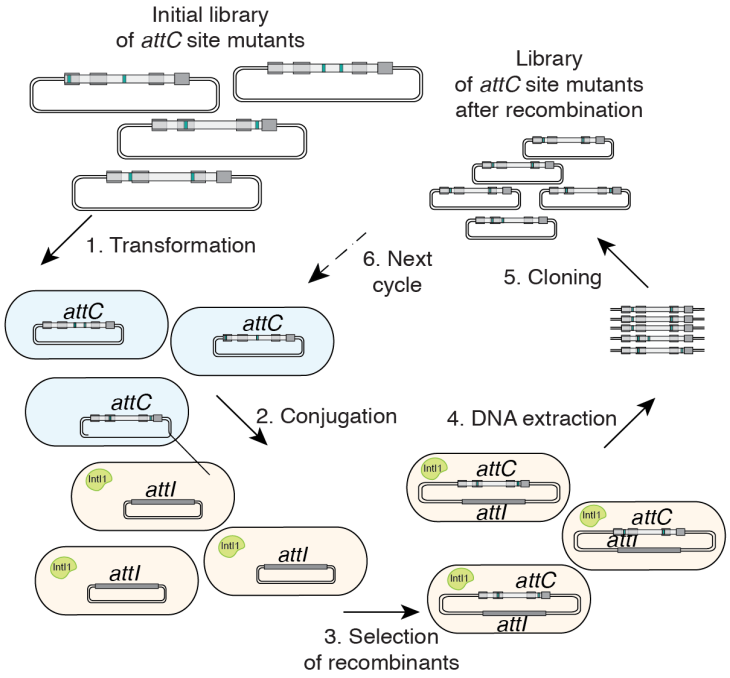

Fig. S2. Schematic of the competition assay used to assess recombination frequencies of mutants from the library.

To construct a library of single and double mutants of $attC_{r0}$, a custom oligonucleotide $attC_{r0}$_library was PCR-amplified with primers Gibson1 and Gibson2. The pSW23T vector p4383 was PCR-amplified with primers Gibson3 and Gibson4. The two products were then purified, joint together through Gibson assembly (*46*) and transformed into the ß2163 strain (*15*). The library of *attC* sites in a pSW23T vector was transformed into a ß2163 strain (Step 1) that maintains its replication through the presence of the π protein. It served as donor in a conjugation (Step 2), where the bottom strands of the vector were delivered into the recipient strain lacking the π protein, containing *attI* and expressing IntI1. The successful recombinants were selected based on the presence of the pSW23T resistance marker (Step 3), their DNA was extracted by PCR (Step 4) and cloned into a pSW23T vector through Gibson assembly (Step 5), which constituted the DNA library after selection. When transformed into a ß2163 strain (Step 6), this library could serve as donor for the next cycle of selection. The DNA used for NGS was the one extracted from cells after Step 1 for the initial library, and after Step 6 for the selected library.

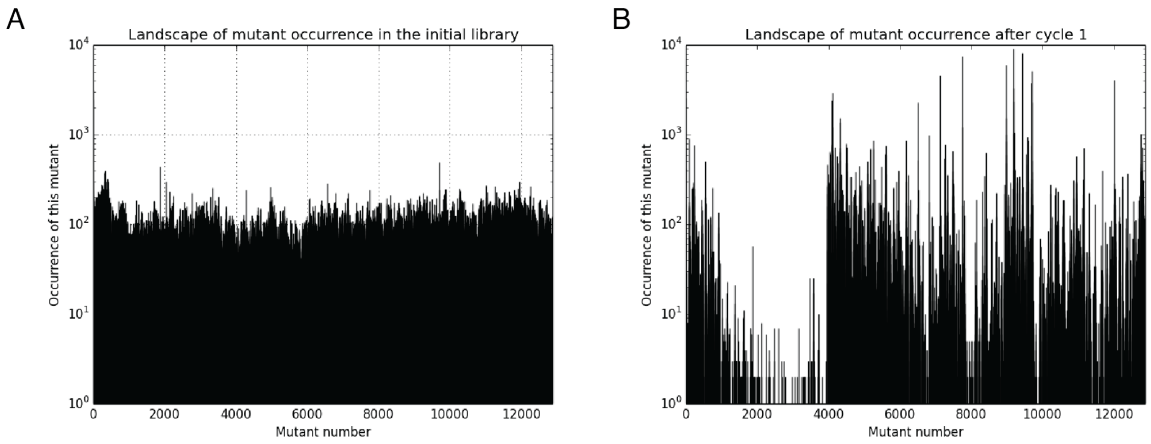

Fig. S3. Next Generation Sequencing of the library before and after enrichment.

The landscapes of mutant occurrences in the library of *attC*$_{r0}$ single and double mutants before (**A**) and after (**B**) competition assay, as measured by Next Generation Sequencing. Each sample contained approximately 10^6 colony forming units (CFUs) and their sequencing yielded between 10^5 and 10^6 reads per library, which corresponds to a high depth coverage, given that the complexity of the library was on the order of 10^4. The order of the mutants along the axis corresponds to the order of mutants along the rows of the top left diagonal in Fig.4C, with the addition of single mutants before each set of double mutants.

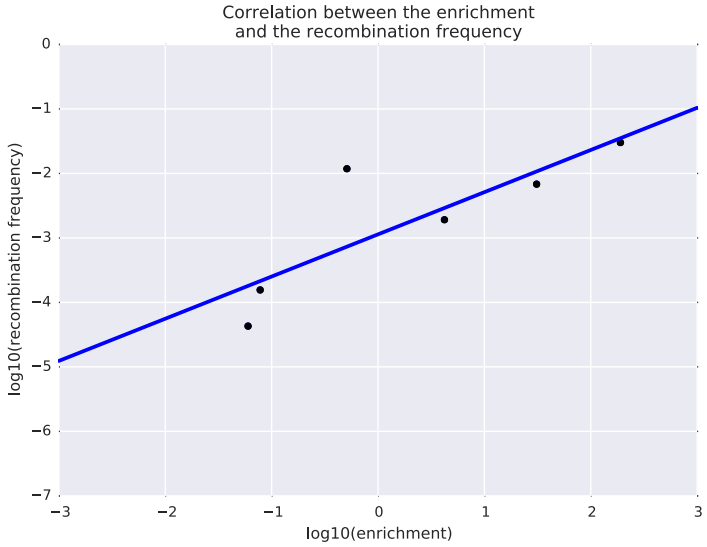

Fig. S4. Validation of enrichment value as a proxy for recombination frequency.

Six *attC*$_{r0}$ double mutants were chosen such that their measured enrichments span the entire range of observed values on a logarithmic scale. Their recombination frequencies were measured experimentally and the mean of three experiments is reported here. The enrichment value can be used as a proxy for the recombination frequency for *attC* site mutants from the library, since the two values show high correlation for this sample of mutants (Pearson R=0.92, p<0.01).

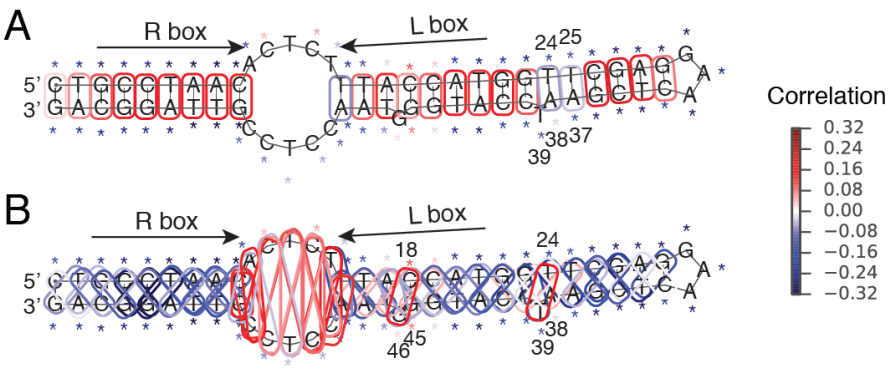

Fig. S5. Correlations of positional entropies and base pairings with enrichment.

Correlations between feature values and enrichment values mapped onto the predicted structure of *attC*$_{r0}$. (**A**) Correlations for base-pairings that were expected according to the structural prediction (ovals). (**B**) Correlations for base-pairings that were not expected according to the structural prediction (ovals). Positional entropies are shown on both schemes (asterisks).

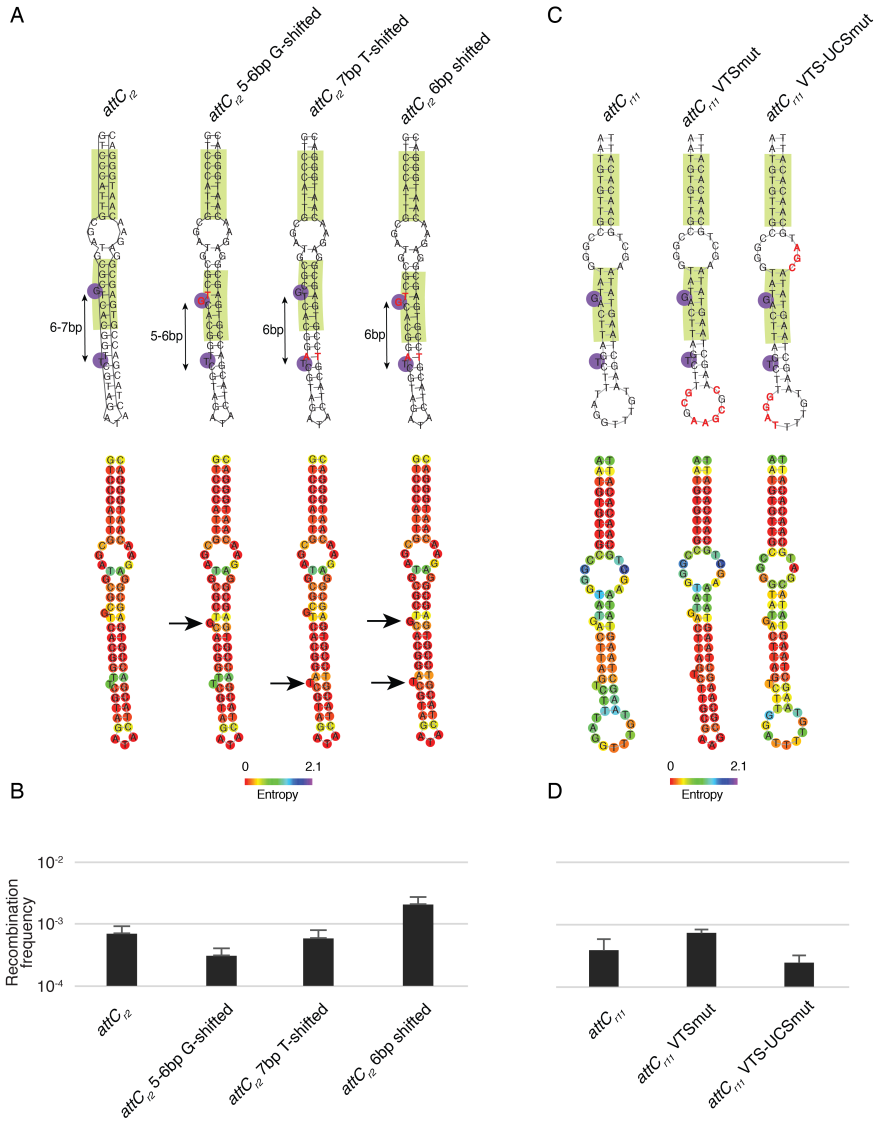

Fig. S6. Additional *attC*$_{r2}$ and *attC*$_{r11}$ mutants, and their recombination frequencies.

(**A**) Top: structural predictions of initial and mutated *attC*$_{r2}$. Mutations are shown in red. Bottom: positional entropies of bases. Arrows indicate EHBs that were "locked" in low entropy state through mutations. (**B**) Recombination frequencies of initial and mutated *attC*$_{r2}$ showing that the decrease in positional entropy and shifting of each EHB alone does not increase its recombination frequency, whereas locking and shifting them both significantly increases recombination. (**C**) Top: structural predictions of initial and mutated *attC*$_{r11}$. Mutations are shown in red. Bottom: positional entropies of bases. Arrows indicate EHBs that were "locked" in low entropy state through mutations. (**D**) Recombination frequencies of initial and mutated *attC*$_{r11}$ showing that mutations in the UCS and/or in the VTS that stabilize the overall stem without affecting the location of EHBs do not increase its recombination frequency. Recombination values represent the mean of three independent experiments; error bars represent mean absolute error.

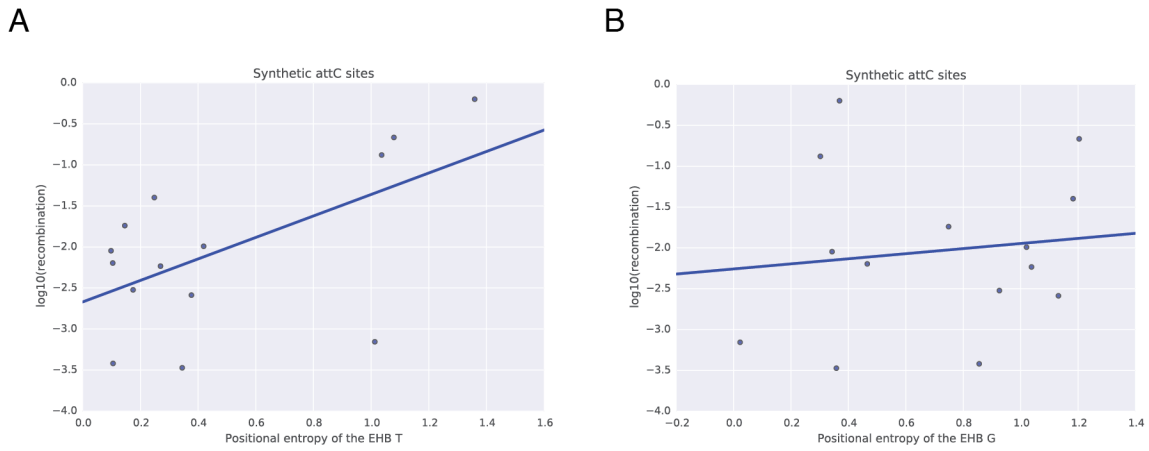

Fig. S7. Correlation of EHB entropies with recombination frequency.

(**A**) Correlation of EHB_G positional entropy of with recombination frequency of 14 synthetic *attC* sites (Pearson R=0.12, p>0.5). (**B**) Correlation of EHB_T positional entropy of with recombination frequency of 14 synthetic *attC* sites (Pearson R=0.58, p<0.05).

Table S1. Constraints used to generate synthetic *attC* sites.

Probability of incorporating each base into the R' sequence of the generated synthetic *attC* site. The probabilities for the first 4 bases are based on the probability distribution of each base in wt *attC* sites from the INTEGRALL database (*14*). The last 3 bases (AAC) are kept constant.

	A	**C**	**G**	**T**
1st position	0.13	0.08	0.69	0.10
2nd position	0.03	0.51	0.16	0.30
3rd position	0.14	0.78	0.01	0.07
4th position	0.01	0.11	0.01	0.87
5th position	1	0	0	0
6th position	1	0	0	0
7th position	0	1	0	0

Table S2. DNA oligonucleotides (A) and plasmids (B) used in this study.

A.

Primer name, orientation		Sequence
attC$_{r0}$	Fw	AATTCTGCCTAACGGAGGTTACCCATGGATTCGAGTTCCTCGAACCATGGTAAAGAGTGTTAGGCAG
	Rev	CTGCCTAACGGAGGTTACCCATGGATTCGAGTTCCTCGAACCATGGTAAAGAGTGTTAGGCAGGATC
attC$_{r1}$	Fw	AATTCCGTCTAACTCATCGCGCGTGAATAAACCTCTTGGAGGTTATTCACCGCAAAATGTTAGACGG
	Rev	GATCCCGTCTAACATTTTGCGGTGAATAACCTCCAAGAGGTTTATTCACGCGCGATGAGTTAGACGG
attC$_{r2}$	Fw	AATTCAGGGTAACGCTACGCGCAGTGCCAAGCATCTATGATGCTGGCACTCGCCTCTTGTTACCCTG
	Rev	GATCCAGGGTAACAAGAGGCGAGTGCCAGCATCATAGATGCTTGGCACTGCGCGTAGCGTTACCCTG
attC$_{r3}$	Fw	AATTCTCTCTAACCTGCTACTCTATAGTACAGTAAGGTTTACTGACTATAAGTCGGGTGTTAGAGAG
	Rev	GATCCTCTCTAACACCCGACTTATAGTCAGTAAACCTTACTGTACTATAGAGTAGCAGGTTAGAGAG
attC$_{r4}$	Fw	AATTCTTTCTAACTCCGCCAACCCGGAGAAGCGTGGCCCACGCTCTCCGGTTGATATTGTTAGAAAG
	Rev	GATCCTTTCTAACAATATCAACCGGAGAGCGTGGGCCACGCTTCTCCGGGTTGGCGGAGTTAGAAAG
attC$_{r5}$	Fw	AATTCGCCTTAACAATACAGGCTATGTTATTTGGTCGGACCAAAAACATACCTAGTAGGTTAAGGCG
	Rev	GATCCGCCTTAACCTACTAGGTATGTTTTTGGTCCGACCAAATAACATAGCCTGTATTGTTAAGGCG
attC$_{r6}$	Fw	AATTCATCTTAACTGCTTACACCCGGGCAACCTTTCCTAAAGGTGCCCGGTGTGGCTTGTTAAGATG
	Rev	GATCCATCTTAACAAGCCACACCGGGCACCTTTAGGAAAGGTTGCCCGGGTGTAAGCAGTTAAGATG
attC$_{r7}$	Fw	AATTCTCGGCTAACCGCTCAGACTATCGCACTACCTGGTTTCTAGGCGATATCTTCGGAGTTAGCCGAG
	Rev	GATCCTCGGCTAACTCCGAAGATATCGCCTAGAAACCAGGTAGTGCGATAGTCTGAGCGGTTAGCCGAG
attC$_{r8}$	Fw	AATTCTAGGCTAACAGAACGGTCAATATGAGGGAGGCTGGATCCCCATATTACCTCGGTGTTAGCCTAG
	Rev	GATCCTAGGCTAACACCGAGGTAATATGGGGATCCAGCCTCCCTCATATTGACCGTTCTGTTAGCCTAG
attC$_{r9}$	Fw	AATTCACGACGAACTCAGACGACAGATATAACCTAAAAGTTCGGTATATCTTCGGGCCGTGTTCGTCGTG
	Rev	GATCCACGACGAACACGGCCCGAAGATATACCGAACTTTTAGGTTATATCTGTCGTCTGAGTTCGTCGTG

attC$_{r10}$	Fw	AATTCCTGCCTAACGTCGTCTGCAGCGTCACACTGTACGCATGT GGACGCTCAGTCAAATGTTAGGCAGG
	Rev	GATCCCTGCCTAACATTTGACTGAGCGTCCACATGCGTACAGTG TGACGCTGCAGACGACGTTAGGCAGG
attC$_{r11}$	Fw	AATTCTTACACAACGGCCCATACTGAATCAGAAATCCAAACAT TCGATTCATATTCGACGTTGTGTAAG
	Rev	GATCCTTACACAACGTCGAATATGAATCGAATGTTTGGATTTCT GATTCAGTATGGGCCGTTGTGTAAG
attC$_{r12}$	Fw	AATTCATGGCTAACTAGTAATACTCAGGGAATCGATCACGGTG ATCCCTGATATGCTCTGTTAGCCATG
	Rev	GATCCATGGCTAACAGAGCATATCAGGGATCACCGTGATCGAT TCCCTGAGTATTACTAGTTAGCCATG
attC$_{r13}$	Fw	AATTCGCCACTAACAGTTTCTACGATTTGAATGTCGATATCGCA TCAAATCTAGCGGCTCGTTAGTGGCG
	Rev	GATCCGCCACTAACGAGCCGCTAGATTTGATGCGATATCGACA TTCAAATCGTAGAAACTGTTAGTGGCG
attC$_{r2}$_ 6bp	Fw	AATTCAGGGTAACGCTACGCGCAGTGCCATGCATCTATGATGC AGGCACTCGCCTCTTGTTACCCTG
	Rev	GATCCAGGGTAACAAGAGGCGAGTGCCTGCATCATAGATGCAT GGCACTGCGCGTAGCGTTACCCTG
attC$_{r2}$_ 6bp_ shifted	Fw	AATTCAGGGTAACGCTACGCGACGTGCCTAGCATCTATGATGC AGGCACTCGCCTCTTGTTACCCTG
	Rev	GATCCAGGGTAACAAGAGGCGAGTGCCTGCATCATAGATGCTA GGCACGTCGCGTAGCGTTACCCTG
attC$_{r6}$_ 6bp	Fw	AATTCATCTTAACTGCTTACGCACGGGCATCCTTTCCTAAAGGA GCCCGTCGTGGCTTGTTAAGATG
	Rev	GATCCATCTTAACAAGCCACGACGGGCTCCTTTAGGAAAGGAT GCCCGTGCGTAAGCAGTTAAGATG
attC$_{r6}$_ 6bp_ shifted	Fw	AATTCATCTTAACTGCTTACAACGGGGCTACCTTTCCTAAAGGA GCCCCTTGTGGCTTGTTAAGATG
	Rev	GATCCATCTTAACAAGCCACAAGGGGCTCCTTTAGGAAAGGTA GCCCCGTTGTAAGCAGTTAAGATG
attC$_{r11}$_ 6bp	Fw	AATTCTTACACAACGGCCCATGCGGAATCACCAATCCAAACAT GGGATTCCCATTCGACGTTGTGTAAG
	Rev	GATCCTTACACAACGTCGAATGGGAATCCCATGTTTGGATTGGT GATTCCGCATGGGCCGTTGTGTAAG
attC$_{r11}$_ 6bp_ shifted	Fw	AATTCTTACACAACGGCCCATGGCGAATCCACAATCCAAACAT GGGATTCCCATTCGACGTTGTGTAAG
	Rev	GATCCTTACACAACGTCGAATGGGAATCCCATGTTTGGATTGTG GATTCGCCATGGGCCGTTGTGTAAG
attC$_{r2}$_ 5-6bp_ G-shifted	Fw	AATTCAGGGTAACGCTACGCGACGTGCCAAGCATCTATGATGC TGGCACTCGCCTCTTGTTACCCTG
	Rev	GATCCAGGGTAACAAGAGGCGAGTGCCAGCATCATAGATGCTT GGCACGTCGCGTAGCGTTACCCTG
attC$_{r2}$_ 7bp_	Fw	AATTCAGGGTAACGCTACGCGCAGTGCCTAGCATCTATGATGC AGGCACTCGCCTCTTGTTACCCTG

T-shifted	Rev	GATCCAGGGTAACAAGAGGCGAGTGCCTGCATCATAGATGCTAGGCACTGCGCGTAGCGTTACCCTG
attC$_{r11}$_ VTSmut	Fw	AATTCTTACACAACGGCCCATACTGAATCAGAACGCTTCGCGTTCGATTCATATTCGACGTTGTGTAAG
	Rev	GATCCTTACACAACGTCGAATATGAATCGAACGCGAAGCGTTCTGATTCAGTATGGGCCGTTGTGTAAG
attC$_{r11}$_ VTS-UCSmut	Fw	AATTCTTACACAACGGCCCATACTGAATCAGAACCTAAAACATTCGATTCATATGCTACGTTGTGTAAG
	Rev	GATCCTTACACAACGTAGCATATGAATCGAATGTTTTAGGTTCTGATTCAGTATGGGCCGTTGTGTAAG
*attC*_ linker1	Fw	CTAGCGGAGTAACCAATAGTAGCAACCAGAACATGTCAAACATGTCTGGTTCTAACTCGAGTTACTCCGGTACGGTAC
	Rev	CGGAGTAACTCGAGTTAGAACCAGACATGTTTGACATGTTCTGGTTGCTACTATTGGTTACTCCG
*attC*_ linker2	Fw	CTAGCGGAGTAACACAGATGAGCTCGAACACCAACGGATCGTTGGGTTCGACTCAAAACGGTTACTCCGGTAC
	Rev	CGGAGTAACCGTTTTGAGTCGAACCCAACGATCCGTTGGTGTTCGAGCTCATCTGTGTTACTCCG
*attC*_ linker3	Fw	CTAGCTCCGTAACTAATAGTTCCCAGAACACCAATGGCTCATTGGGTTCTGGAAATAACGGTTACGGAGGTAC
	Rev	CTCCGTAACCGTTATTTCCAGAACCCAATGAGCCATTGGTGTTCTGGGAACTATTAGTTACGGAG
LacZ-OE-PCR	Fw	CTTCCGGCTCGTATGTTGTG
	Rev	GCCTGACTGGCGGTTAAATTGCC
*LacZ-attC*_rev1	Fw	ACGGGGTGAACAGTTGCAACCATCTGTGGTGCAACTTTCGATGGGTTGGATACGGCCAGGACAGTCGTTTGCCGTCTGAATTTGACC
	Rev	TCCAACCCATCGAAAGTTGCACCACAGATGGTTGCAACTGTTCACCCCGTTGGATATAATTCGCGTCTGGCCTTCCTGTAGCC
*LacZ-attC*_rev2	Fw	CGGAGAGCTGGCTGGAGTGCGATCTCCCGGAAAGCGACACCGTTGTGGTCCCCTCAAACTGGCAGATGCACGGTTACGATGC
	Rev	CCGGGAGATCGCACTCCAGCCAGCTCTCCGGAAGAGCTTCTGGTTGTGGAAACCAGGCAAAGCGCCATTCGCCATTCAGG
*LacZ-attC*_rev3	Fw	GGAGAGCGCCGGGGAGCTCTGGATCACCGTTAGAGTAGTGCAACCGAACGCGACCGCATGG
	Rev	GGCTGGGGTAGCTCTGGGAGCTCTGTTAGAGGTTTACCTTGTGGAGCGACATCC
*LacZ-attC*_rev4	Fw	GTCAACTAGCGATAACTGTCGATGTTGAGGTGGCGAGCGATACACCGCATCCGGCGCGG
	Rev	GTCAACTAGCGATAACTGTCGATGTTGAGGTGGCGAGCGATACACCGCATCCGGCGCGG
LacZ-attCaadA7	Fw	TGAATTAAGCCGCGCCGCGAAGCGGCGTCGGCTTGAATGAATTGTTATAACTCGCGTCTGGCCTTCCTGTAGCCAGCTTTCATC
	Rev	AACAATTCATTCAAGCCGACGCCGCTTCGCGGCGCGGCTTAATTCAAGCGTTATAACCGGCCAGGACAGTCGTTTGCCGTCTGAATTTGACC

pKAC_linker_fwd	CCGCATCTGTCCAACTTCCG
pKAC_linker_rev	CACGCCGATATTCATGTCGC
attC$_{r0}$_library *	GTCCTAAGGTAGCGAA$N_2N_4N_3N_2N_2N_4N_1N_1N_2N_3N_3N_1N_3N_3N_4N_4N_1N_2N_2N_2N_1N_4N_3N_3N_1N_4N_4N_2N_3N_1N_3N_4N_4N_2N_2N_4N_2N_3N_1N_1N_2N_2N_1N_4N_3N_3N_4N_1N_1N_1N_3N_1N_3N_4$GTTAGGCAGTAGGGATAACAG
Library_Gibson1	TGGAGAGGGTGAAGGTGATGACATAACTATAACGGTCCTAAGGTAGCGAA
Library_Gibson2	GCTCTAGAACTAGTGGATCCAGTATTACCCTGTTATCCCTACTGCCTAAC
Library_Gibson3	CATCACCTTCACCCTCTCCAGTCGACGCCGGCCAGCCTCGCAGAGCAGGA
Library_Gibson4	GGATCCACTAGTTCTAGAGCGGCCGCCACCGCGGTGGAGC
SWbeg	CCGTCACAGGTATTTATTCGGCG
SWend	CCTCACTAAAGGGAACAAAAGCTG

*: N1, N2, N3 and N4 correspond to custom oligonucleotide mixes:
N1 = 96.1% A, 1.3% C, 1.3% G, 1.3% T
N2 = 96.1% C, 1.3% A, 1.3% G, 1.3% T
N3 = 96.1% G, 1.3% C, 1.3% A, 1.3% T
N4 = 96.1% T, 1.3% C, 1.3% G, 1.3% A

B.

Plasmid number	Plasmid description	Plasmid properties and construction
p3938	pBAD::*intI1*	*ori*ColE1; [Carb^R] (*22*)
p929	pSU38Δ::*attI1*	*ori*p15A; [Kan^R] (*42*)
p4383	pSW23T	*oriV*$_{R6K\gamma}$, *oriT*$_{RP4}$;[Cm^R] (*15*)
p4849	pSW23T::*attP*	p4383 with *attP* cloned into SacII site of p4383
p9276	pSW23T::*attC*$_{r0}$	Annealing of primers and cloning into p4849
p9277	pSW23T::*attC*$_{r1}$	Annealing of primers and cloning into p4849
p9278	pSW23T::*attC*$_{r2}$	Annealing of primers and cloning into p4849
p9279	pSW23T::*attC*$_{r3}$	Annealing of primers and cloning into p4849
p9280	pSW23T::*attC*$_{r4}$	Annealing of primers and cloning into p4849
p9281	pSW23T::*attC*$_{r5}$	Annealing of primers and cloning into p4849
p9282	pSW23T::*attC*$_{r6}$	Annealing of primers and cloning into p4849
pG582	pSW23T::*attC*$_{r7}$	Annealing of primers and cloning into p4849
pG583	pSW23T::*attC*$_{r8}$	Annealing of primers and cloning into p4849
pG584	pSW23T::*attC*$_{r9}$	Annealing of primers and cloning into p4849
pG585	pSW23T::*attC*$_{r10}$	Annealing of primers and cloning into p4849
pG586	pSW23T::*attC*$_{r11}$	Annealing of primers and cloning into p4849
pG587	pSW23T::*attC*$_{r12}$	Annealing of primers and cloning into p4849
pG588	pSW23T::*attC*$_{r13}$	Annealing of primers and cloning into p4849
p9895	pSW23T::*attC*$_{L1}$	Annealing of primers and cloning into p4849
p9896	pSW23T::*attC*$_{L2}$	Annealing of primers and cloning into p4849
p9897	pSW23T::*attC*$_{L3}$	Annealing of primers and cloning into p4849
p9919	pKAC::T25-*attC*$_{L1}$-T28	Annealing of primers and cloning into p9922

p9920	pKAC::T25-*attC*$_{L2}$-T28	Annealing of primers and cloning into p9922
p9921	pKAC::T25-*attC*$_{L3}$-T28	Annealing of primers and cloning into p9922
p9922	pKAC::p5	*ori*p15A; [Kan^{R}] (18)
p9949	pKAC::T25-*attC*$_{L1}$FS-T28	p9919 digested by NheI, incubated with Klenow fragment to introduce a frameshift by complementing the overhangs, and re-ligated.
pG768	pSW23T::*attC*$_{r2}$_6bp	Annealing of primers and cloning into p4849
pG771	pSW23T::*attC*$_{r2}$_6bp_shifted	Annealing of primers and cloning into p4849
pG775	pSW23T::*attC*$_{r6}$_6bp	Annealing of primers and cloning into p4849
pG776	pSW23T::*attC*$_{r6}$_6bp_shifted	Annealing of primers and cloning into p4849
pG779	pSW23T::*attC*$_{r11}$_6bp	Annealing of primers and cloning into p4849
pG780	pSW23T::*attC*$_{r11}$_6bp_shifted	Annealing of primers and cloning into p4849
pG770	pSW23T::*attC*$_{r2}$_5-6bp_Gshifted	Annealing of primers and cloning into p4849
pG769	pSW23T::*attC*$_{r2}$_7bp_Tshifted	Annealing of primers and cloning into p4849
pG777	pSW23T::*attC*$_{r11}$_VTSmut	Annealing of primers and cloning into p4849
pG778	pSW23T::*attC*$_{r11}$_VTS-UCSmut	Annealing of primers and cloning into p4849
p1370	pSU38Δ::*pLac-lacZ*	Lab collection; unpublished
p2713	pSW23T::*attI1*	Lab collection; unpublished
pN232	pSU38Δ::*pLac-lacZ::attC*$_{reg1}$	Amplification of *attC*$_{reg1}$ and flanking regions by two overlap extension (OE-PCR) and two *attC*-specific primers, followed by digestion with BamHI and ClaI; ligation with p1370 digested with BamHI and ClaI.
pN235	pSU38Δ::*pLac-lacZ::attC*$_{reg2}$	Amplification of *attC*$_{reg2}$ and flanking regions by two overlap extension (OE-PCR) and two *attC*-specific primers, followed by digestion with BamHI and ClaI; ligation with p1370 digested with BamHI and ClaI.
pN021	pSU38Δ::*pLac-lacZ::attC*$_{reg3}$	Amplification of p1370 with primers introducing *attC*$_{reg3}$; ligation of blunt ends.

pN009	pSU38Δ::*pLac-lacZ::attC*$_{reg4}$	Amplification of p1370 with primers introducing *attC*$_{reg4}$; ligation of blunt ends.
pN070	pSU38Δ::*pLac-lacZ::attC*$_{aadA7}$	Amplification of p1370 with primers introducing *attC*$_{aadA7}$; Gibson assembly.

Table S3. Performance of ML algorithms.

(**A**) Manually curated feature lists. (**B**) Performance of the three ML algorithms. Letters indicate feature lists that were used in the manual feature selection method. (**C**). Feature importance values of the Random Forest Regression algorithm.

A

Feature list	Number of features	Features
A	4	Ensemble ΔG, Ensemble diversity, MFE ΔG, MFE frequency
B	5	H-bond number, G:T number, ΔG(bs)-ΔG(ts), ΔG(bs)/ΔG(ts), pfold
C	63	Nature of each base
D	220	Positional entropy and base pairing probability for each nucleotide, pairwise base pairing probabilities
Total	292	

B

Regression Methods	**FS Method**	**#features**	**Mean Correlation**	**Mean Abs error**	**Mean square error**	**varScore**
DecisionTree		7	0.59	0.14	0.035	0.17
RidgeRegressor	PCA	50	0.73	0.112	0.02	0.53
SVR		42	0.77	0.101	0.017	0.59
DecisionTree		48	0.62	0.133	0.032	0.24
RidgeRegressor	k-best-features	46	0.69	0.119	0.023	0.47
SVR		48	0.68	0.12	0.023	0.47
DecisionTreeRegressor			0.54	0.147	0.039	0.08
RidgeRegressor	A	4	0.66	0.123	0.024	0.43
SVR			0.66	0.121	0.024	0.44
DecisionTreeRegressor			0.54	0.145	0.038	0.11
RidgeRegressor	B	5	0.46	0.154	0.033	0.21
SVR			0.52	0.146	0.031	0.27
DecisionTreeRegressor			0.63	0.127	0.031	0.29
RidgeRegressor	C	54	0.55	0.141	0.03	0.3
SVR			0.56	0.141	0.029	0.32
DecisionTreeRegressor			0.65	0.128	0.031	0.29
RidgeRegressor	D	215	0.7	0.12	0.022	0.49
SVR			0.75	0.109	0.019	0.56
DecisionTreeRegressor			0.56	0.144	0.038	0.12
RidgeRegressor	A+B	9	0.68	0.121	0.121	0.47
SVR			0.69	0.119	0.023	0.47
DecisionTreeRegressor			0.65	0.125	0.029	0.32
RidgeRegressor	A+B+C	63	0.73	0.112	0.02	0.53
SVR			0.73	0.112	0.02	0.53

C

Feature	Importance	Feature	Importance	Feature	Importance
pos_entr_7D	1.53E-02	bp_proba_3_61D	4.02E-03	bp_18D	2.97E-04
pos_entr_40D	1.50E-02	pos_entr_14D	4.00E-03	bp_proba_20_42D	2.96E-04

pos_entr_57D	1.42E-02	pos_entr_10D	3.95E-03	base_16C	2.90E-04
bp_proba_7_57D	1.42E-02	bp_proba_4_59D	3.94E-03	bp_7D	2.89E-04
bp_proba_23_41D	1.40E-02	ddG_BOT_TOPB	3.88E-03	bp_proba_19_43D	2.78E-04
pos_entr_22D	1.36E-02	pos_entr_16D	3.80E-03	base_23C	2.69E-04
pos_entr_41D	1.33E-02	bp_proba_16_49D	3.73E-03	bp_proba_28_33D	2.67E-04
pos_entr_23D	1.33E-02	bp_proba_25_37D	3.66E-03	base_34C	2.66E-04
pos_entr_21D	1.25E-02	base_54C	3.52E-03	bp_proba_1_62D	2.66E-04
pos_entr_32D	1.24E-02	bp_proba_24_38D	3.49E-03	bp_14D	2.64E-04
bp_proba_24_37D	1.21E-02	bp_proba_2_62D	3.13E-03	bp_proba_10_54D	2.61E-04
pos_entr_59D	1.18E-02	bp_proba_1_63D	3.07E-03	bp_proba_11_53D	2.61E-04
pos_entr_31D	1.17E-02	bp_proba_24_39D	2.29E-03	bp_proba_26_37D	2.58E-04
pos_entr_30D	1.16E-02	base_39C	2.14E-03	base_51C	2.47E-04
pos_entr_8D	1.13E-02	bp_20D	1.62E-03	bp_8D	2.45E-04
bp_proba_8_56D	1.12E-02	base_37C	1.61E-03	base_63C	2.42E-04
pos_entr_58D	1.11E-02	bp_proba_18_47D	1.50E-03	bp_proba_28_35D	2.41E-04
pos_entr_42D	1.10E-02	base_45C	1.35E-03	base_21C	2.32E-04
pos_entr_56D	1.09E-02	bp_proba_16_47D	9.66E-04	bp_41D	2.24E-04
pos_entr_5D	1.08E-02	bp_proba_17_45D	9.45E-04	bp_2D	2.23E-04
bp_proba_23_40D	1.06E-02	bp_43D	9.35E-04	bp_proba_2_61D	2.12E-04
pos_entr_60D	1.04E-02	bp_proba_14_50D	9.15E-04	bp_proba_4_61D	2.11E-04
pos_entr_4D	1.02E-02	bp_37D	9.04E-04	bp_proba_13_50D	2.09E-04
bp_proba_6_58D	9.88E-03	bp_proba_2_63D	8.85E-04	base_40C	2.08E-04
bp_proba_21_42D	9.44E-03	base_20C	8.53E-04	base_28C	2.07E-04
pos_entr_6D	9.43E-03	bp_33D	8.43E-04	bp_62D	2.06E-04
pos_entr_39D	9.20E-03	GT_numberB	8.18E-04	base_58C	2.03E-04
bp_proba_20_43D	9.03E-03	base_12C	7.41E-04	bp_proba_22_42D	2.01E-04
pos_entr_43D	8.92E-03	bp_proba_17_48D	7.28E-04	base_55C	1.86E-04
pos_entr_27D	8.83E-03	bp_proba_9_54D	7.26E-04	bp_proba_13_51D	1.85E-04
pos_entr_18D	8.73E-03	bp_proba_11_54D	7.21E-04	base_26C	1.79E-04
pos_entr_37D	8.69E-03	bp_29D	7.20E-04	base_41C	1.72E-04
Boltz_diversityA	8.68E-03	bp_9D	7.00E-04	base_43C	1.72E-04
bp_proba_22_41D	8.62E-03	bp_38D	6.93E-04	bp_21D	1.66E-04
pos_entr_3D	8.48E-03	bp_proba_24_40D	6.51E-04	base_61C	1.64E-04
pos_entr_61D	8.46E-03	base_30C	6.30E-04	bp_proba_5_58D	1.62E-04
bp_proba_18_46D	8.40E-03	bp_proba_12_53D	6.16E-04	bp_57D	1.62E-04
pos_entr_20D	8.34E-03	bp_24D	6.10E-04	bp_42D	1.59E-04
pos_entr_34D	7.99E-03	base_32C	6.08E-04	bp_54D	1.55E-04
pos_entr_44D	7.90E-03	base_49C	5.90E-04	bp_13D	1.55E-04
pos_entr_35D	7.89E-03	bp_proba_14_51D	5.86E-04	base_35C	1.52E-04
pfoldB	7.61E-03	bp_proba_15_50D	5.80E-04	bp_56D	1.43E-04

pos_entr_28D	7.59E-03	base_24C	5.73E-04	base_22C	1.43E-04
pos_entr_26D	7.53E-03	bp_40D	5.65E-04	bp_19D	1.42E-04
pos_entr_45D	7.52E-03	bp_44D	5.28E-04	base_42C	1.42E-04
bp_proba_9_55D	7.45E-03	bp_25D	5.28E-04	bp_5D	1.39E-04
pos_entr_19D	7.44E-03	bp_proba_12_52D	5.20E-04	bp_10D	1.32E-04
pos_entr_36D	7.41E-03	base_18C	5.17E-04	bp_6D	1.24E-04
bp_proba_23_39D	7.39E-03	bp_proba_21_41D	5.16E-04	bp_proba_3_60D	1.24E-04
pos_entr_33D	7.36E-03	bp_45D	5.15E-04	bp_22D	1.14E-04
bp_proba_18_45D	7.34E-03	bp_proba_20_44D	5.02E-04	bp_proba_26_35D	9.85E-05
pos_entr_29D	7.32E-03	base_47C	5.01E-04	bp_53D	9.67E-05
bp_proba_17_47D	7.06E-03	bp_proba_23_38D	4.98E-04	bp_4D	9.27E-05
bp_proba_25_38D	6.96E-03	bp_34D	4.98E-04	bp_52D	9.17E-05
bp_proba_19_45D	6.77E-03	bp_proba_17_46D	4.97E-04	bp_3D	8.83E-05
bp_proba_19_44D	6.74E-03	bp_55D	4.95E-04	bp_58D	7.86E-05
pos_entr_55D	6.67E-03	base_31C	4.95E-04	bp_61D	7.34E-05
bp_proba_26_36D	6.37E-03	base_17C	4.90E-04	bp_51D	6.89E-05
pos_entr_9D	6.31E-03	base_13C	4.86E-04	bp_59D	6.52E-05
bp_proba_28_34D	6.14E-03	bp_35D	4.76E-04	base_56C	6.07E-05
MFE_freqA	6.13E-03	base_14C	4.73E-04	bp_proba_6_57D	5.50E-05
bp_proba_18_44D	6.06E-03	base_38C	4.61E-04	base_36C	5.35E-05
pos_entr_38D	5.97E-03	base_53C	4.50E-04	base_27C	5.19E-05
bp_proba_5_59D	5.82E-03	bp_27D	4.47E-04	base_59C	5.18E-05
bp_proba_7_56D	5.82E-03	bp_50D	4.31E-04	bp_proba_9_56D	4.94E-05
bp_proba_16_48D	5.82E-03	base_44C	4.06E-04	bp_60D	4.74E-05
bp_proba_5_60D	5.75E-03	base_46C	4.04E-04	base_57C	4.71E-05
pos_entr_25D	5.73E-03	bp_proba_12_51D	4.03E-04	base_60C	4.63E-05
bp_proba_27_35D	5.56E-03	base_25C	3.98E-04	bp_proba_8_55D	3.53E-05
pos_entr_46D	5.53E-03	base_19C	3.95E-04	bp_12D	3.30E-05
bp_proba_29_34D	5.50E-03	bp_28D	3.87E-04	bp_11D	3.15E-05
bp_proba_15_48D	5.48E-03	bp_proba_13_52D	3.84E-04	bp_31D	2.51E-05
pos_entr_63D	5.46E-03	bp_proba_27_36D	3.75E-04	bp_32D	2.01E-05
pos_entr_50D	5.44E-03	bp_23D	3.72E-04	bp_30D	1.72E-05
pos_entr_17D	5.39E-03	bp_39D	3.71E-04	bp_proba_7_58D	8.29E-06
pos_entr_2D	5.38E-03	bp_16D	3.70E-04		
bp_proba_4_60D	5.35E-03	base_29C	3.70E-04		
bp_proba_22_40D	5.34E-03	base_15C	3.68E-04		
pos_entr_24D	5.22E-03	bp_49D	3.67E-04		
pos_entr_1D	5.19E-03	bp_36D	3.61E-04		
dG_ratio_BOT_TOPB	5.10E-03	bp_26D	3.60E-04		
bp_proba_14_49D	5.01E-03	base_33C	3.60E-04		

pos_entr_54D	4.75E-03	bp_15D	3.59E-04
pos_entr_62D	4.66E-03	bp_47D	3.49E-04
pos_entr_52D	4.65E-03	bp_proba_10_53D	3.41E-04
bp_proba_29_33D	4.65E-03	bp_48D	3.39E-04
pos_entr_49D	4.50E-03	base_62C	3.36E-04
pos_entr_15D	4.49E-03	bp_17D	3.35E-04
bp_proba_15_49D	4.47E-03	bp_proba_10_55D	3.33E-04
bp_proba_8_57D	4.38E-03	base_50C	3.33E-04
pos_entr_11D	4.35E-03	bp_46D	3.29E-04
pos_entr_53D	4.34E-03	base_48C	3.24E-04
bp_proba_6_59D	4.27E-03	base_52C	3.21E-04
bp_proba_25_36D	4.25E-03	base_11C	3.19E-04
pos_entr_51D	4.22E-03	bp_63D	3.19E-04
pos_entr_12D	4.20E-03	bp_1D	3.14E-04
pos_entr_47D	4.12E-03	bp_proba_11_52D	3.06E-04
bp_proba_27_34D	4.10E-03	bp_proba_21_43D	3.03E-04
pos_entr_13D	4.09E-03	bp_proba_3_62D	3.00E-04
pos_entr_48D	4.08E-03	base_10C	2.99E-04

REFERENCES AND NOTES

1. F. J. Olorunniji, S. J. Rosser, W. M. Stark, Site-specific recombinases: Molecular machines for the genetic revolution. *Biochem. J.* **473**, 673–684 (2016).
2. S. K. Sharan, L. C. Thomason, S. G. Kuznetsov, D. L. Court, Recombineering: A homologous recombination-based method of genetic engineering. *Nat. Protoc.* **4**, 206–223 (2009).
3. J. A. Escudero, C. Loot, A. Nivina, D. Mazel, The integron: Adaptation on demand. *Microbiol. Spectr.* **3**, MDNA3–0019–2014 (2015).
4. J. A. Escudero, C. Loot, V. Parissi, A. Nivina, C.Bouchier, D. Mazel, Unmasking the ancestral activity of integron integrases reveals a smooth evolutionary transition during functional innovation. *Nat. Commun.* **6**, 10937 (2016).
5. M. V. Francia, J. C. Zabala, F. de la Cruz, J. M. García Lobo, The IntI1 integron integrase preferentially binds single-stranded DNA of the attC site. *J. Bacteriol.* **181**, 6844–6849 (1999).
6. M. Bouvier, G. Demarre, D. Mazel, Integron cassette insertion: A recombination process involving a folded single strand substrate. *EMBO J.* **24**, 4356–4367 (2005).
7. N. D. F. Grindley, K. L. Whiteson, P. A. Rice, Mechanisms of site-specific recombination. *Annu. Rev. Biochem.* **75**, 567–605 (2006).
8. A. Nivina, J. A. Escudero, C. Vit, D. Mazel, C. Loot, 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 (2016).
9. M. Bouvier, M. Ducos-Galand, C. Loot, D. Bikard, D. Mazel, Structural features of single-stranded integron cassette attC sites and their role in strand selection. *PLOS Genet.* **5**, e1000632 (2009).
10. R. M. Hall, D. E. Brookes, H. W. Stokes, Site-specific insertion of genes into integrons: Role of the 59-base element and determination of the recombination cross-over point. *Mol. Microbiol.* **5**, 1941–1959 (1991).
11. C. Johansson, M. Kamali-Moghaddam, L. Sundström, Integron integrase binds to bulged hairpin DNA. *Nucleic Acids Res.* **32**, 4033–4043 (2004).
12. C. Frumerie, M. Ducos-Galand, D. N. Gopaul, D. Mazel, The relaxed requirements of the integron cleavage site allow predictable changes in integron target specificity. *Nucleic Acids Res.* **38**, 559–569 (2010).

13. D. MacDonald, G. Demarre, M. Bouvier, D. Mazel, D. N. Gopaul, Structural basis for broad DNA-specificity in integron recombination. *Nature* **440**, 1157–1162 (2006).
14. A. Moura, M. Soares, C. Pereira, N. Leitão, I. Henriques, A. Correia, INTEGRALL: A database and search engine for integrons, integrases and gene cassettes. *Bioinformatics* **25**, 1096–1098 (2009).
15. G. Demarre, A.-M. Guerout, C. Matsumoto-Mashimo, D. A. Rowe-Magnus, P. Marliere, D. Mazel, A new family of mobilizable suicide plasmids based on broad host range R388 plasmid (IncW) and RP4 plasmid (IncPα) conjugative machineries and their cognate *Escherichia coli* host strains. *Res. Microbiol.* **156**, 245–255 (2005).
16. D. Mazel, B. Dychinco, V. A. Webb, J. Davies, Antibiotic resistance in the ECOR collection: Integrons and identification of a novel *aad* gene. *Antimicrob. Agents Chemother.* **44**, 1568–1574 (2000).
17. R. A. George, J. Heringa, An analysis of protein domain linkers: Their classification and role in protein folding. *Protein Eng.* **15**, 871–879 (2002).
18. G. Karimova, J. Pidoux, A. Ullmann, D. Ladant, A bacterial two-hybrid system based on a reconstituted signal transduction pathway. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 5752–5756 (1998).
19. N. Dautin, G. Karimova, A. Ullmann, D. Ladant, Sensitive genetic screen for protease activity based on a cyclic AMP signaling cascade in *Escherichia coli*. *J. Bacteriol.* **182**, 7060–7066 (2000).
20. J. Cury, T. Jové, M. Touchon, B. Néron, E. P. Rocha, Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucleic Acids Res.* **44**, 4539–4550 (2016).
21. D. Yu, H. M. Ellis, E.-C. Lee, N. A. Jenkins, N. G. Copeland, D. L. Court, An efficient recombination system for chromosome engineering in Escherichia coli. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 5978–5983 (2000).
22. G. Demarre, C. Frumerie, D. N. Gopaul, D. Mazel, Identification of key structural determinants of the IntI1 integron integrase that influence *attC* × *attI1* recombination efficiency. *Nucleic Acids Res.* **35**, 6475–6489 (2007).
23. C. M. Bishop, Pattern recognition. *Mach. Learn.* **128**, 1–58 (2006).
24. R. Lorenz, S. H. Bernhart, C. H. zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, I. L. Hofacker, ViennaRNA package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).

25. C. Loot, D. Bikard, A. Rachlin, D. Mazel, Cellular pathways controlling integron cassette site folding. *EMBO J.* **29**, 2623–2634 (2010).
26. C. Loot, A. Nivina, J. Cury, J. A. Escudero, M. Ducos-Galand, D. Bikard, E. P. Rocha, D. Mazel, Differences in integron cassette excision dynamics shape a trade-off between evolvability and genetic capacitance. *MBio* **8**, e02296-16 (2017).
27. R. J. Lewis, An introduction to classification and regression tree (CART) analysis, in *Annual Meeting of the Society for Academic Emergency Medicine*, (San Francisco, CA, 2000), pp. 1–14.
28. A. E. Hoerl, R. W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67 (1970).
29. V. Vapnik, *Statistical Learning Theory* (Wiley, 1998).
30. L. Breiman, Random forests. *Mach. Learn.* **45**, 5–32 (2001).
31. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
32. Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics. *Bioinformatics* **23**, 2507–2517 (2007).
33. I. Jolliffe, *Principal Component Analysis* (Wiley Online Library, 2002).
34. A. Larouche, P. H. Roy, Effect of *attC* structure on cassette excision by integron integrases. *Mob. DNA* **2**, 3 (2011).
35. H. G. Menzella, R. Reid, J. R. Carney, S. S. Chandran, S. J. Reisinger, K. G. Patel, D. A. Hopwood, D. V. Santi, Combinatorial polyketide biosynthesis by de novo design and rearrangement of modular polyketide synthase genes. *Nat. Biotechnol.* **23**, 1171–1176 (2005).
36. K. A. J. Bozhüyük, F. Fleischhacker, A. Linck, F. Wesche, A. Tietze, C. P. Niesert. H. B. Bode, De novo design and engineering of non-ribosomal peptide synthetases. *Nat. Chem.* **10**, 275–281 (2018).
37. D. Bikard, S. Julié-Galau, G. Cambray, D. Mazel, The synthetic integron: An in vivo genetic shuffling device. *Nucleic Acids Res.* **38**, e153 (2010).

38. C. Loot, M. Ducos-Galand, J. A. Escudero, M. Bouvier, D. Mazel, Replicative resolution of integron cassette insertion. *Nucleic Acids Res.* **40**, 8361–8370 (2012).
39. J. Matos, S. C. West, Holliday junction resolution: Regulation in space and time. *DNA Repair* **19**, 176–181 (2014).
40. M. S. Grieb, A. Nivina, B. L. Cheeseman, A. Hartmann, D. Mazel, M. Schlierf, Dynamic stepwise opening of integron attC DNA hairpins by SSB prevents toxicity and ensures functionality. *Nucleic Acids Res.* **45**, 10555–10563 (2017).
41. A. Mukhortava, M. Pöge, M. S. Grieb, A. Nivina, C. Loot, D. Mazel, M. Schlierf, Structural heterogeneity of attC integron recombination sites revealed by optical tweezers. *Nucleic Acids Res.* **47**, 1861–1870 (2019).
42. L. Biskri, M. Bouvier, A.-M. Guerout, S. Boisnard, D. Mazel, Comparative study of class 1 integron and *Vibrio cholerae* superintegron integrase activities. *J. Bacteriol.* **187**, 1740–1750 (2005).
43. J. G. Carbonell, R. S. Michalski, T. M. Mitchell, in *Machine Learning: An Artificial Intelligence Approach*, R. S. Michalski, J. G. Carbonell, T. M. Mitchell, Eds. (Springer, 1983), pp. 3–23.
44. P. Baldi, S. Brunak, *Bioinformatics: The Machine Learning Approach* (MIT Press, ed. 2, 2001).
45. G. Karimova, D. Ladant, A. Ullmann, L. Selig, P. Legrain, Bacterial two-hybrid system for protein-protein interaction screening, new strains for use therein, and their applications, U.S. Patent 0045237 (2002).
46. O. Espeli, L. Moulin, F. Boccard, Transcription attenuation associated with bacterial repetitive extragenic BIME elements. *J. Mol. Biol.* **314**, 375–386 (2001).
47. D. G. Gibson, L. Young, R.-Y. Chuang, J.C. Venter, C. A. Hutchison III, H. O. Smith, Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).
48. R. L. Plackett, Karl Pearson and the chi-squared test. *Int. Stat. Rev.* **51**, 59–72 (1983).