nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-022-31538-3 OPEN

Model-guided engineering of DNA sequences with predictable site-specific recombination rates

Qiuge Zhang [1], Samira M. Azarin [1] & Casim A. Sarkar [2]✉

Site-specific recombination (SSR) is an important tool in synthetic biology, but its applications are limited by the inability to predictably tune SSR reaction rates. Facile rate manipulation could be achieved by modifying the DNA substrate sequence; however, this approach lacks rational design principles. Here, we develop an integrated experimental and computational method to engineer the DNA attachment sequence attP for predictably modulating the inversion reaction mediated by the recombinase Bxb1. After developing a qPCR method to measure SSR reaction rate, we design, select, and sequence attP libraries to inform a machine-learning model that computes Bxb1 inversion rate as a function of attP sequence. We use this model to predict reaction rates of attP variants in vitro and demonstrate their utility in gene circuit design in *Escherichia coli*. Our high-throughput, model-guided approach for rationally tuning SSR reaction rates enhances our understanding of recombinase function and expands the synthetic biology toolbox.

[1] Department of Chemical Engineering and Materials Science, University of Minnesota, Minneapolis, MN 55455, USA. [2] Department of Biomedical Engineering, University of Minnesota, Minneapolis, MN 55455, USA. ✉email: csarkar@umn.edu

Site-specific recombination (SSR) technology relies on recombinases that can precisely recognize two DNA sites, form an intermediate complex, cut, swap, and recombine the DNA in a new configuration, resulting in gene insertions, deletions, or inversions[1]. Based on the active residue in the catalytic domain, the recombinase superfamily is divided into two groups: tyrosine recombinases and serine recombinases[2]. Each group can be further subdivided by directionality (bidirectional/unidirectional) for tyrosine recombinases or by size (large/small) for serine recombinases[3]. Among these four recombinase subgroups, the large serine recombinases (LSRs) are considered one of the most powerful tools in synthetic biology based on the following properties[4]:

Irreversibility: LSRs have non-identical recognition sites typically known as attB (attachment site bacteria) and attP (attachment site phage) and yield hybrid product sites attL and attR. LSRs cannot target the hybrid attL and attR sites to regenerate attP and attB, resulting in an exceptionally stable DNA recombination product, which is in contrast to the commonly used Cre-lox and FLP-FRT systems[5]. This feature is important in applications such as human cell genome editing[6] and gene circuits for data storage in living cells[7].

Simplicity: In contrast to some tyrosine recombinases such as λ integrase, which requires long attP sites (>200 bp), supercoiled DNA structure, and other factors to stabilize DNA bending, serine recombinases have short DNA sites (<50 bp) and no required DNA topology or cofactors[8].

Efficiency: LSRs such as Bxb1, TP901 and φC31 have been demonstrated to be efficient in both prokaryotic and eukaryotic cells, in gene therapy[9], memory circuit design[10,11], and genome editing[12].

However, SSR applications in gene circuit design have been largely limited to the creation of logic gates and memory, with a focus on the end products and not the rates at which they are produced[7,10,11]. The ability to predictably control SSR reaction rates would enable rational tuning of timescales in the aforementioned applications and would also enable advanced circuit designs that rely on differential SSR reaction rates. An understanding of the DNA determinants of such processes could not only lead to improvements in wild-type recombination rates but could also provide a suite of parts that could be coupled together to enable higher-order information processing in genetic circuits via kinetic control. Here, we focused on understanding and engineering the DNA inversion reaction mediated by the mycobacteriophage integrase Bxb1[13], though given the shared functional mechanisms of LSRs, our approach should be readily translatable to other LSRs as well.

Previous engineering approaches for regulating biochemical reaction rates have focused on altering key amino acid residues in the enzyme[14]. However, rational protein design is limited by the lack of high-resolution recombinase-DNA complex crystal structures. To our knowledge, only one structure of a DNA-bound LSR, Listeria phage A118 integrase, has been reported and the resolution of the protein-DNA interface is not sufficiently high[15]. Despite ongoing efforts to understand the interactions between amino acid residues of recombinases and nucleotides, static recombinase-DNA complex crystal structures cannot provide sufficient information to understand DNA sequence determinants of SSR rates[15–19]. In addition to direct protein-DNA contacts, charge/shape complementarity and water-mediated interactions contribute to the SSR rate of different DNA sequences[20]. Furthermore, mutations in the recombinase can alter protein stability or solubility, confounding efforts to rationally tune reaction rates via enzyme engineering. Although engineering a recombinase is in principle possible using a library-based approach, this would be more challenging due to the lack of a high-throughput enzyme selection method for SSR rates. Additionally, for gene circuit applications, it is easier to create multiple DNA substrates for a single recombinase than to create multiple recombinases for the same substrate. Therefore, we focused on engineering the DNA attachment sites and developing a method to rationally design DNA attachment site sequences to modulate the SSR reaction rate. Previously, the impact of single and double base substitutions in the Bxb1-attP site on specificity and directionality was reported[21]. Another group used a high-throughput approach to identify DNA specificity determinants by selecting saturation mutagenesis DNA site libraries[22]. Although essential for better understanding the recombination mechanism and revealing potential off-target substrates, these studies focused only on DNA sequence specificity. Considering the potential application of LSRs in the design of genetic circuits, it would also be useful to be able to rationally tune their reaction kinetics. Through this mode of predictable kinetic control at the DNA level, it would be possible to use recombinases in applications beyond genetic memory, such as coordination of protein expression dynamics and temporal ordering of genetically encoded processes.

We therefore sought to develop a method to programmably tune the Bxb1 reaction rate via the DNA attachment sequences. However, a method to accurately measure SSR reaction rates or a platform to screen DNA sequences for recombination on a large scale has not been well established. In this study, we developed a qPCR-based method for quantifying relative SSR reaction rate as well as a platform for profiling SSR reaction rates of selected sequences from a designed DNA library using next-generation sequencing (NGS). Then, we constructed a machine-learning model to quantify the contributions of different nucleotide substitutions in attP to the overall SSR reaction rate. Finally, using assays both in vitro and in *E. coli*, we demonstrated accurate model predictions of rates of DNA inversion. Our study enables rational modulation of SSR reaction rates, providing a form of kinetic control for predictably tuning synthetic genetic circuits and gene editing.

Results

DNA library design. For our initial DNA library design, we first identified nucleotide positions in the attP and attB DNA attachment sites of Bxb1 that could potentially be substituted to vary the reaction rate while maintaining recognition specificity. As shown in Fig. 1a, during the SSR reaction, the attP and attB attachment sites each bind a Bxb1 dimer. After synapsis through the interaction between the coiled-coil (CC) motifs at the different sites, the DNA sequence is cleaved at the center and then rotated 180°. After rotation, the conformation of the CC groups bound together on the same DNA strand is much more thermodynamically stable; thus, the ligation step is essentially irreversible and drives the overall reaction from the attP and attB substrate sites to the attL and attR product sites[23]. Notably, the attP and attB sites for Bxb1 have non-identical sequences, with both having two quasi-half-sites attP-L/attP-R and attB-L/attB-R (Fig. 1b). A Bxb1 monomer bound to the attP-L half-site has direct contact with the DNA site sequence via the zinc ribbon domain (ZD), the recombinase domain (RD), and the linker connecting ZD and RD (Fig. 1c)[15]. The shorter length of the attB site forces the linker to adopt a different conformation when bound to attB half sites[15]. As shown in Fig. 1d (underlined), the four half-sites attP-L, attP-R, attB-L, and attB-R have ~50% conserved bases. Previous studies have demonstrated that the homology positions are highly conserved for specificity[24]. Therefore, we hypothesized that these conserved positions may directly interact with protein residues and are necessary for DNA

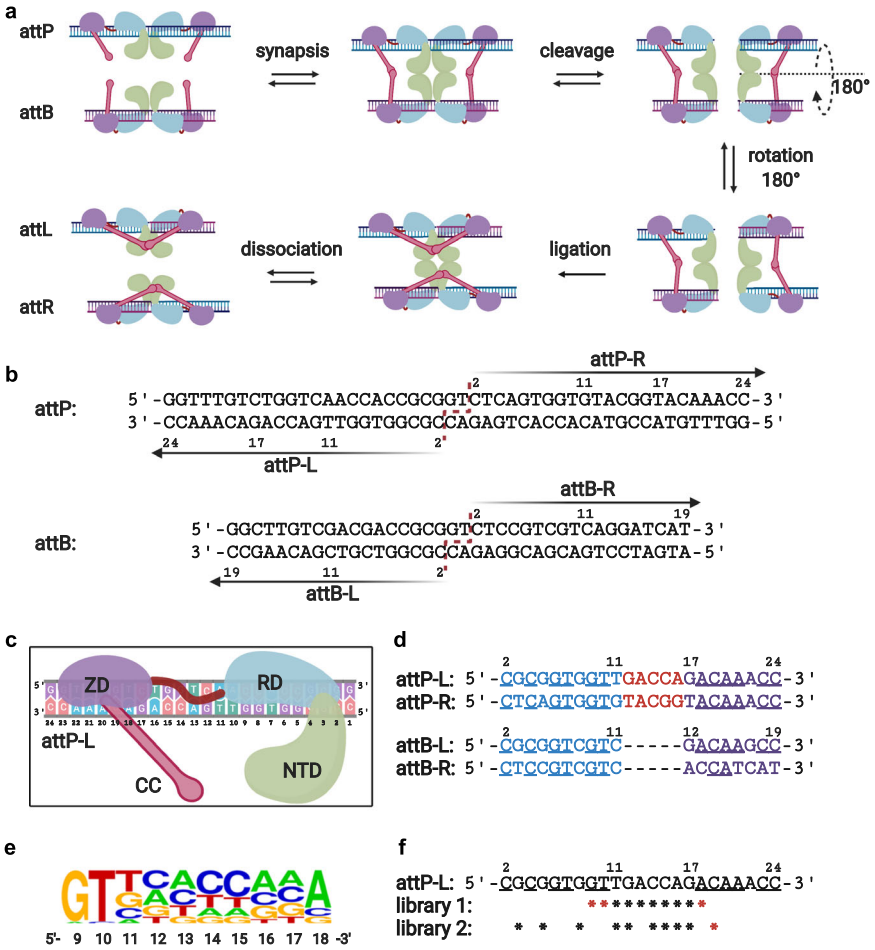

Fig. 1 Bxb1 attP-L DNA library design. a DNA inversion mediated by Bxb1 recombinase. Attachment sites attP and attB are bound by Bxb1 dimers, followed by synapsis through intermolecular binding of coiled-coil motifs (CC, pink). DNA sequences are then cleaved in the middle by the recombinase domain and rotated 180°. The resulting sticky ends, which are complementary, are ligated. After rotation, Bxb1 forms a more stable conformation, with CC motifs interacting within the same Bxb1 dimer on a given DNA sequence. Finally, Bxb1 dissociates from the two newly generated attachment sites, attL and attR, and the intervening DNA sequence is inverted. **b** DNA sequences of Bxb1 attP and attB sites. Cleavage happens at the central bases GT in both (red dashed line). Both attP and attB have two quasi-palindromic half sites attP-L/attP-R and attB-L/attB-R (also referred to as attP/attP′and attB/attB′ half-sites). **c** In addition to the coiled-coil domain (CC, pink), a Bxb1 monomer includes the N-terminal domain (NTD, green), recombinase domain (RD, blue), and zinc ribbon domain (ZD, purple). A linker (red) connects RD and ZD and adopts different conformations when bound to attP or attB half-sites due to their differential lengths. **d** Comparison of attP and attB sites and their features. Four half-site sequences are aligned from center to end, with numbering from the 5'-terminus to the 3'-terminus. Underlined nucleotides indicate positions that are conserved across at least three half sites. Positions bound to RD or ZD are shown in blue or purple, respectively, and bases in red are aligned with the linker. **e** WebLogo of base frequency from position 9 to position 18 in library 1 after an *E. coli*-based selection. **f** Designed saturation mutagenesis libraries 1 and 2, where an asterisk represents a degenerate position in the library (with equal representation of bases A, C, G, and T at each such position; Supplementary Method 1 and Supplementary Table 4). Figures 1a, c were created with BioRender.com. Source data are provided as a Source Data file.

recognition, whereas bases at other asymmetric positions might be substituted to vary reaction rates.

To test this hypothesis, we chose to perform saturation mutagenesis on the attP-L half-site based on the following considerations. First, the attP site has more asymmetric positions than attB (nucleotides in red, Fig. 1d), potentially expanding the tunable range of reaction rates. Second, considering limitations of transformation efficiency and NGS read depth in the selection step, the number of positions for saturation mutagenesis had to be limited to 10 nucleotides (library size = 4^{10} ~ 1 million). Lastly, within the attP site, substituting bases in both the attP-L and attP-R half-sites would be more likely to result in a loss of specificity[21]. We therefore opted to keep the attP-R half-site unchanged, introducing mutations only within the attP-L half-site.

In previous studies, base conservation at homologous positions was characterized in vitro by gel electrophoresis and reaction rates could not be quantitively measured due to the lack of a highly sensitive quantification method[22]. Since SSR can be impacted by multiple factors, including direct protein-DNA contacts and long-distance interactions, we designed a selection method in *E. coli* to ensure that our selected attP-L variants function in vivo. This selection method entails inducing a DNA inversion reaction that confers expression of chloramphenicol acetyltransferase, with selected attP-L variants then recovered by plating on chloramphenicol-containing agar plates (Supplementary Fig. 1). We constructed a DNA library containing random bases at 10 continuous positions from 9 to 18 in attP-L, including symmetric sites and asymmetric sites (library 1 in Fig. 1f). From initial selections with library 1 in *E. coli*, 20 colonies with SSR functionality in DNA flipping were Sanger sequenced and three highly conserved bases—9 G, 10 T, and 18 A—were identified (Supplementary Fig. 1 and Supplementary Method 1). These

conserved bases are shared by at least three out of the four wild-type half-sites (Fig. 1e), suggesting that these positions are essential for maintaining SSR specificity, while other positions 11 T, 12 G, 13 A, 14 C, 15 C, 16 A, 17 G were tolerant to nucleotide substitutions.

Given these observations, we hypothesized that the homologous positions among the four half-sites are indispensable for SSR recognition, while the asymmetric positions could provide more flexibility in tuning the reaction rates. We therefore designed a new saturation mutagenesis DNA library (library 2, Fig. 1f) to identify variants with a broader range of reaction rates by altering the asymmetric positions (3 G, 5 G, 8 G, 11 T, 12 G, 14 C, 15 C, 16 A, 17 G), while keeping one putatively conserved base, 19 C, as a control in the selection.

DNA flipping rate assay and library selection in vitro. Traditional methods for characterizing DNA recombination events are based on differential electrophoretic mobility of substrates and products[18,21,23], but they are not amenable to high-throughput experiments. Furthermore, quantification of band intensities in gels is not sensitive enough to detect low product concentrations at the beginning of the reaction, when the initial rate can capture differences in the intrinsic efficiency of DNA variants. In this work, we developed a qPCR-based method to accurately measure the initial SSR reaction rate by quantifying the recombinant DNA product concentration. In our in vitro experiments, we identified a 5-min incubation time as optimal: this time period is long enough to generate and robustly quantify flipped products and is also short enough to maintain an approximately constant substrate concentration. These conditions, in contrast to prior approaches, enable estimation of the initial reaction rate and, consequently, the reaction rate constant (Supplementary Method 2). As shown in Fig. 2a, two primers were designed to selectively amplify the flipped product DNA but not the substrate. Briefly, we constructed a linear DNA fragment flanked by attP and attB sites positioned in opposite directions. Bxb1 binds to attP/attB sites and then inverts the flanked DNA segment. Our designed primers are complementary to the adjacent sequences of the cleavage site on the attP sequence, and the inversion reaction changes both the orientation of one of the primers and the DNA strand to which it anneals, such that PCR results in exponential amplification of the inverted template, which, importantly, includes the attP-L sequence. For unflipped DNA substrate, both primers extend in the same direction on the same strand, so no exponential amplification is possible by PCR. We used qPCR to measure the percentage of flipping with different Bxb1 concentrations for wild-type (WT) attP sites, and the result was consistent with those from DNA gel electrophoresis (Supplementary Fig. 2). To test the sensitivity of this quantification method, a standard curve of C_q values and flipped DNA template concentration was plotted (Fig. 2b). The total copy number of unflipped and flipped DNA fragments was fixed at 10^9 per qPCR reaction volume (20 µl) and the ratio of flipped to unflipped DNA was varied (0:1, 1:10^7, 1:10^6, 1:10^5, 1:10^4, 1:10^3, 1:10^2, 1:10, 1:0). As shown in Fig. 2b, our qPCR approach accurately measured the flipped DNA percentage ($r^2 = 0.998$). Additionally, the lower bound of the linear range was ~10^3 flipped DNA copies, indicating that this method has a high sensitivity and can detect a flipping percentage as low as 0.0001%. Compared with previous methods based on differences in electrophoretic mobility[21] or exonuclease digestion[22], our method is facile and can accurately measure initial reaction rates.

With this sensitive and accurate qPCR-based rate assay, we selected for DNA variants with high recombination efficiencies using a DNA library as the substrate. To prepare the library, we used oligonucleotides containing a degenerate base (N = equimolar A, C, G, and T) at designated positions (Fig. 2a). Then, recombinant Bxb1 was added to the DNA library and incubated in reaction buffer. The attP-L variants with efficient SSR functionality underwent flipping of the DNA sequence flanked by the attP and attB sites, whereas the remaining inefficient or nonfunctional DNA molecules stayed unflipped. From this mixture, the flipped DNA product was selectively amplified by qPCR and sequenced by NGS. Sequence variants that led to faster reaction rates produced more flipped DNA products and thus appeared more frequently in the NGS pool. By ranking the frequency of the flipped DNA sequences, we obtain a list of sequences in order of their flipping rate.

To optimize the selection conditions, we tested Bxb1 concentrations ranging from 1 nM to 300 nM. At high enzyme concentrations, we were unable to distinguish the reaction rates of attP-L variants due to the lower selection pressure. Further, low enzyme concentrations resulted in very low flipped percentages, even below the sensitivity of qPCR detection. As shown in Fig. 2c, qPCR was unable to detect flipped DNA in library 1 at the lowest Bxb1 concentrations (1 nM and 2 nM), whereas under the same conditions, 0.002% and 0.017% of flipped DNA substrate was detected in library 2. The flipped percentages for library 2 were higher than library 1 at all other Bxb1 concentrations as well. This observation was consistent with our hypothesis in library design, as library 2 (with only one putatively conserved position, 19 C) was expected to have more functional sequences than library 1 (with three conserved positions, 9 G, 10 T, and 18 A). Last, in measuring the flipped percentage under the most permissive reaction conditions (300 nM Bxb1 and 1 h incubation time), we found that 33% of library 1 and 51% of library 2 variants possessed SSR competence. Although Bxb1 is considered to have high specificity, this result suggests that the half-site attP-L is not highly resistant to nucleotide substitution when paired with three unchanged WT half-sites, attP-R, attB-L, and attB-R. More importantly, it revealed the potential to modulate DNA recombination rate by altering the bases at specific positions in attP-L.

After sequencing the flipped libraries using NGS, we quantified the percentage of wild-type bases at each nucleotide position (Supplementary Tables 5 and 6) and plotted them as an enrichment heat map. The enrichment score for a given base was defined as the ratio between its percentage in the selected library and its percentage in the naïve library. As shown in Fig. 2d, e, the high enrichment scores of WT bases at positions 9, 10, and 19 supported our hypothesis that homologous positions tended to be functionally conserved. However, for other positions, especially 5, 8, 12, and 17, WT bases were more amenable to substitution, with some substitutions even preferred over WT bases under stringent selection conditions (Supplementary Tables 5 and 6). As shown in Fig. 2d, e, these results demonstrated that our in vitro selection approach using qPCR and NGS was able to identify optimal base substitutions, consistent with our preliminary in vivo selections in *E. coli*. In addition, these substitutions were more pronounced at lower Bxb1 concentrations. Next, to investigate the relationship between the entire attP-L sequence and its corresponding SSR reaction rate, we analyzed the frequency of individual attP-L variants from the NGS results and then converted the frequency into an SSR rate constant, k_{flip}.

Model training and calculation of weight scores. For each individual attP variant in library 2, we counted the frequency of occurrence from the refined NGS results. As shown in Fig. 3a, as the Bxb1 concentration decreased, fewer variants were selected and appeared in the sequencing results, and their frequencies varied more between the different selected DNA sequences due to the

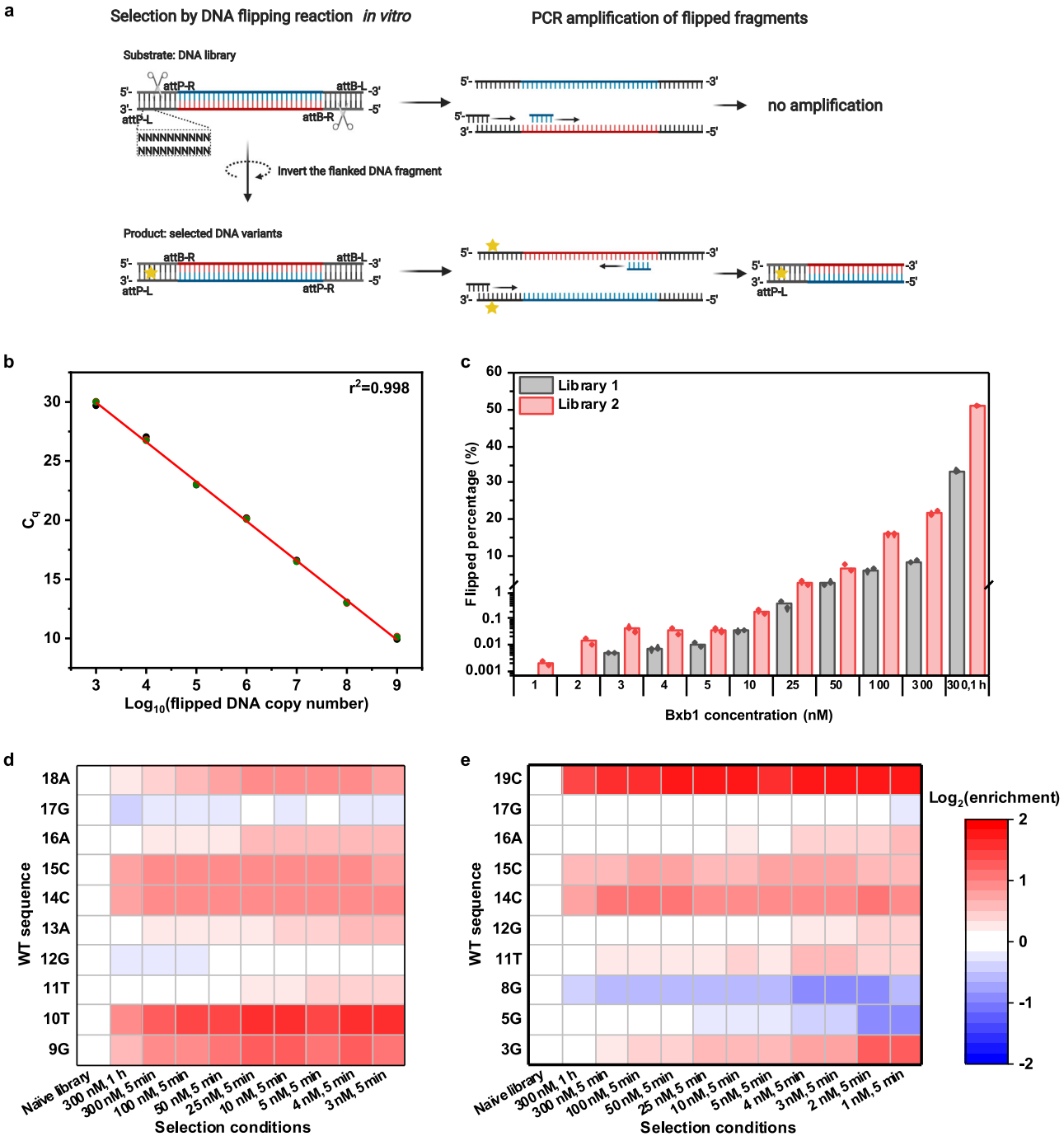

Fig. 2 Quantification of DNA inversion and attP-L library selections by qPCR. a Library selection procedure. First, a DNA site-saturation library is constructed by designed oligonucleotides with equimolar mixed nucleotides (N) at 10 designated positions (shown as asterisks in Fig. 1f). The 10 consecutive Ns shown here depict library 1; library 2 is conceptually analogous but is not illustrated here. Second, sequences with DNA recombination functionality (yellow star) undergo inversion in the reaction solution. Last, the flipped sequences are selectively enriched by qPCR. For flipped DNA fragments, the black and blue primers anneal to two different DNA strands, resulting in amplified products (bottom); for unflipped DNA, the primers anneal to the same strand and no PCR product can be amplified (top). **b** Standard curve for qPCR quantification of flipped DNA fragments. The templates for qPCR are mixtures of flipped and unflipped DNA with different ratios and the same total copy number 10^9 (slope = −3.378, intercept = 40.169, $r^2 = 0.998$). Three technical replicates were performed. **c** qPCR quantification of DNA library flipping. Bxb1 at concentrations varying from 1 nM to 300 nM was added to the reaction buffer with 10 nM DNA library and then incubated at 30 °C for 5 min. The reaction was terminated by denaturation at 80 °C for 20 min. The sample with 300 nM Bxb1 and incubated for 1 h at 30 °C was considered the reaction with the most permissive conditions. Two biological replicates were performed for all conditions except for the condition of 300 nM Bxb1 and 1 h incubation. **d**, **e** Heat maps of wild-type nucleotide enrichment as a function of selection stress, which increases from left to right in each map (**d** library 1; **e** library 2). Figure 2a was created with BioRender.com. Source data are provided as a Source Data file.

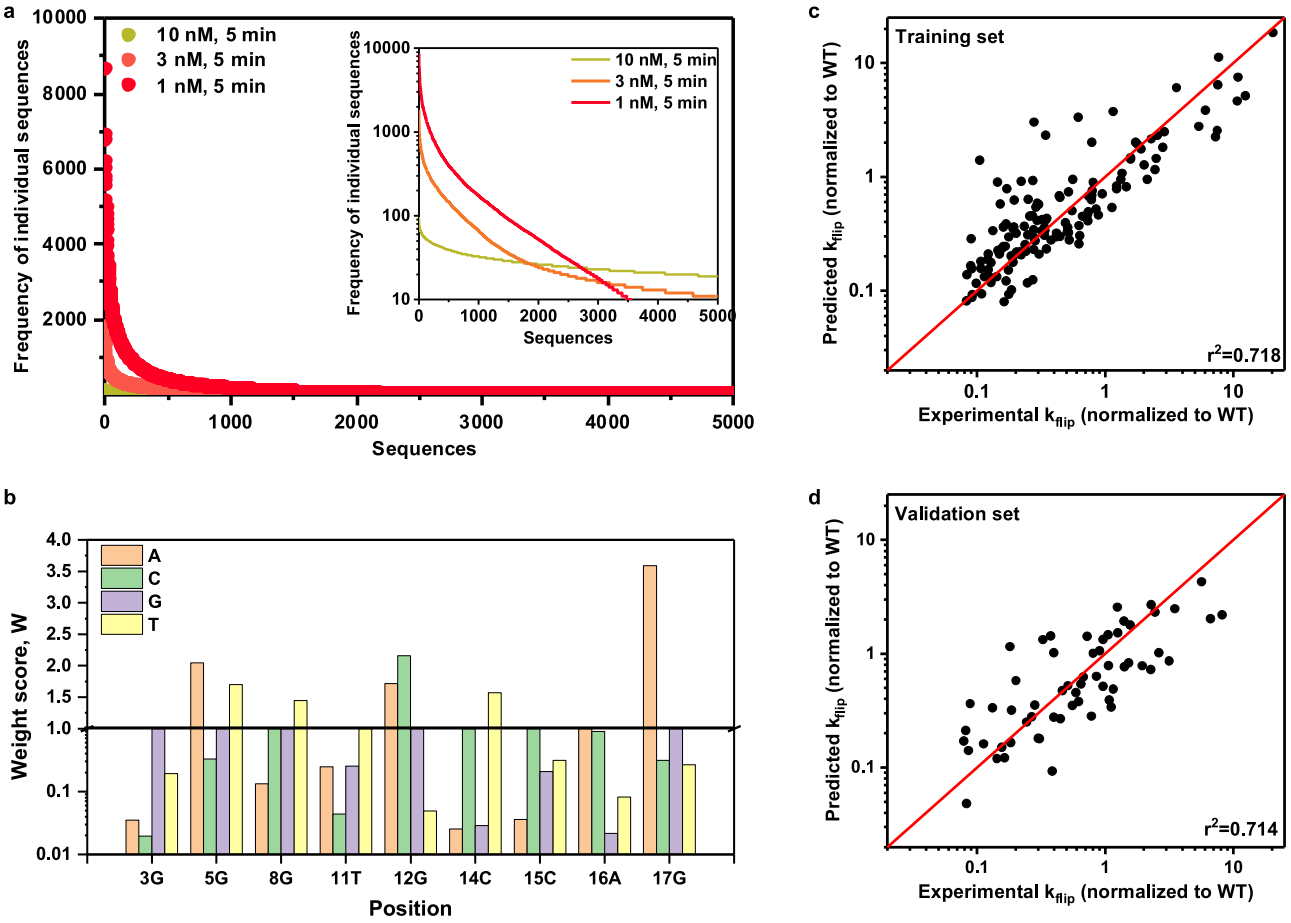

Fig. 3 Model fitting of NGS data. a Occurrence frequency of individual sequences from library 2 selected with different Bxb1 concentrations. **b** Weight scores generated from linear regression. Experimental log(k_{flip}) values are derived from the occurrence frequency in NGS results (Supplementary Method 2). Weight scores of the reference WT sequences are fixed as 1 (i.e., $W_{i1} = 1$). For the *y*-axis in 3b, 0.01-1 is on a $\log_{10}$ scale and 1–4 is on a linear scale. **c** Predicted k_{flip} values from the model correlate with the experimental k_{flip} values for 140 randomly chosen training sequences (70% of 200 randomly chosen sequences from the top 3000 hits in a selection with 1 nM Bxb1). **d** Predicted k_{flip} values from the model correlate with the experimental k_{flip} values for 60 randomly chosen validation sequences (the remaining 30% of these 200 randomly chosen sequences). Source data are provided as a Source Data file.

same depth of 5×10^5 reads during sequencing. To demonstrate that frequency and reaction rate have the correlation we expected, we selected 8 sequences, including WT, from the 3000 most frequently occurring sequences and performed SSR assays on each individual sequence. For these tested sequences, the frequencies in the NGS results and the flipped percentages in the SSR assays showed a good correlation ($r^2 = 0.83$; Supplementary Fig. 3c). Given this experimentally confirmed relationship, we converted the occurrence frequency to a more biologically relevant value, a flipping rate constant (k_{flip}), to represent the intrinsic activity of the attP-L variants (Supplementary Method 2). In brief, under selection conditions of low Bxb1 and DNA concentrations, SSR reactions with different DNA variants as substrates can be considered independent, with the reaction rate for an individual sequence governed by a first-order reaction with a rate constant k_{flip}. Thus, for each individual attP-L variant, its frequency of occurrence within the total reads was proportional to its percentage of flipped DNA after a period of reaction time, so its flipping rate can be calculated from its k_{flip} value and a reaction time of 5 min.

We initially posited that k_{flip} might be reasonably predicted by a model that assumes independent contributions of each DNA base in the attP-L sequence (Eq. (1)). To test our hypothesis and quantify the contributions of different bases at each nucleotide position, we converted this into a linear model (Eq. (2)) and defined weight score parameters to quantify the relative contribution of each possible nucleotide substitution.

$$k_{\mathrm{flip}} = \prod_{i=1}^{9} \left\{ \sum_{j=1}^{4} (W_{ij} \cdot X_{ij}) \right\} \quad (1)$$

$$\log(k_{\mathrm{flip}}) = \sum_{i=1}^{9} \left\{ \log\left(\sum_{j=1}^{4} (W_{ij} \cdot X_{ij}) \right) \right\} \quad (2)$$

In this model, the input data are the DNA sequences (X_{ij}) and the output data are the sequence-specific rate constants (k_{flip}), converted from frequencies in the NGS results. We sought to determine the weight scores (W_{ij}) that would yield the best fit by linear regression using the least squares method (described in Supplementary Method 2)[25]. First, the input sequence as a character string was converted into a binary two-dimensional matrix X_{ij} using the one hot encoding method[26,27] (Supplementary Table 7), which assumes independent contributions of each nucleotide without prior classification and thus simplifies the calculation of the overall reaction rate of each sequence. We expressed the contribution of different bases to the overall flipping rate constant k_{flip} by Eq. (1), which assumed that the overall k_{flip} was the product of the weight scores W_{ij} of the nine positions. By taking the logarithm of both sides in Eq. (1), the resulting Eq. (2) is linear with respect to all log(W_{ij}) values since only one X_{ij} value is non-zero for a given value of i.

As shown in the inset in Fig. 3a, the frequency of sequences outside of the 3000 most enriched sequences was very low (<28).

Given the lower reliability of these frequencies and the more limited utility of these sequences with very low reaction rates, we only analyzed the top 3000 sequences under selection with 1 nM Bxb1 and determined the corresponding nucleotide weight scores W_{ij} by linear regression (Fig. 3b). The W_{i1} values, corresponding to the WT reference sequence, were set to 1, which in turn produces an overall $k_{flip,WT} = 1$. Analyzing the fitted W_{ij} for all position bases, we found that the weight scores of most substitutions were smaller than WT base at each position, which is consistent with the expectation that most mutations would be less efficient than WT even though library 2 was designed to largely exclude highly conserved positions. Specifically, some substitutions resulted in moderate loss of DNA recombination function ($0.1 < W < 1$), including 3 T, 5 G, 8 A, 11 A, 11 G, 15 G, 15 T, 16 C, 17 C, and 17 G. Other substitutions such as 3 A, 3 C, 11 C, 12 T, 14 A, 14 G, 15 A, 16 G, and 16 T greatly disrupted DNA recombination ($W < 0.1$). Notably, several substitutions—5 A, 5 T, 8 T, 12 A, 12 C, 14 T, and 17 A—had a $W > 1$, suggesting higher flipping rates than the WT sequence. To examine whether the weight scores obtained from one set of sequences can accurately predict the reaction rates for another set of sequences, we randomly split 200 sequences from the library into a training set (140 sequences) and a validation set (60 sequences). The weight scores derived from the training sequences (Fig. 3c) were able to accurately predict the DNA recombination efficiency of the validation sequence set (Fig. 3d). Additionally, the frequency profiles from NGS are consistent with the simulated frequency curves based on flipping rate constants predicted by the linear regression model (Supplementary Fig. 3b).

Model evaluation. To further evaluate the model and to exclude the possibility of overfitting, we randomly selected different numbers of sequences (from 10 to 3000) from the 3000 most enriched sequences to train our model, and then analyzed the variation of the weight scores (Fig. 4a) and prediction errors (Fig. 4b) as a function of sequence number. The prediction error for both the training and validation data sets was defined as the difference between the predicted output and the experimental output of $\log(k_{flip})$. As seen in Fig. 4b, the prediction error of the validation set was much larger than that of the training set when the total number of sequences was <100, suggestive of overfitting. This conclusion is also supported by the highly variable W values when the sequence number is <100 (Fig. 4a). As more sequences were used for model training, the prediction set error and the validation set error converged to an acceptable value ($\log(k_{flip}) < 0.1$) around 100–200 sequences (Figs. 4a, b), suggesting that this number is sufficient to quantitatively predict the performance of each individual sequence with a non-negligible reaction rate in a DNA library, which in this case is the top ~3000 enriched sequences. The model is also able to classify sequences with very low reaction rates and is in principle capable of quantitatively predicting their rates (given the independence and convergence of W_{ij} values from model training), but the minimal NGS frequency of selected sequences outside of the top 3000 and the sensitivity limit of our qPCR method prevent accurate experimental quantification and validation of k_{flip} values in this regime.

When analyzing sequences selected with 1 nM Bxb1 for 5 min, we obtained weight scores that were good predictors of the overall reaction rate constant, k_{flip}. Given that the specific selection conditions can influence the weight scores, as evidenced in Fig. 2d, e, we analyzed DNA sequences selected under different Bxb1 concentrations. In comparing the results of weight scores obtained under 1 nM, 3 nM, and 10 nM Bxb1 selection conditions, we found that for most of the positions, the values obtained under different selection conditions were similar (Supplementary Fig. 4), indicating that nucleotide substitutions that are intrinsically more or less efficient than wild type in facilitating SSR are robustly captured by our model. The magnitude of the weight scores decreased as the concentration of Bxb1 increased from 1 nM to 3 nM to 10 nM in the selections (Supplementary Fig. 4). This is not unexpected since increasing the enzyme concentration can offset biochemical advantages or disadvantages that specific nucleotide substitutions generate in enzyme-substrate affinity or catalytic turnover.

As previously noted, the logarithm of k_{flip} is proportional to the sum of the sequence weight scores in our model, which assumes independence of the contributions of individual nucleotides to the overall reaction rate constant. In protein engineering, multiple residue substitutions may have cooperative interactions or cause significant conformational changes at the binding or catalytic site; however, for the nucleotide substitutions here, significant structural changes at the protein-DNA interface are less likely to occur due to the rigid double-stranded helical structure of the DNA and the tolerance of recombinases to some sequence substitutions in their target DNA[28]. Furthermore, based on homology modeling of the Bxb1-DNA crystal structure complex (Supplementary Fig. 5), we

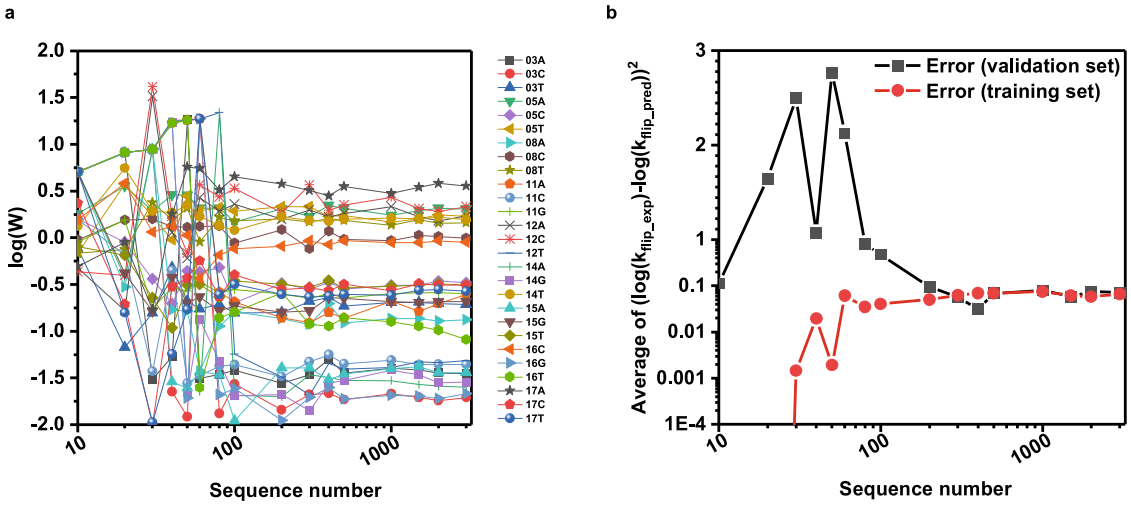

Fig. 4 Sensitivity of the model to input sequence number. a Generated weight score values, W, as a function of the size of the data set (the number on the *x*-axis is total data set, 70% of which are training data and 30% of which are validation data). **b** Difference between predicted $\log(k_{flip})$ and experimental $\log(k_{flip})$ of the training data set (red) and validation data set (black). Source data are provided as a Source Data file.

found that the high conservation at certain nucleotide positions is due to strong recombinase interactions, both from the methyl groups on thymine and specific hydrogen bond donors and acceptors in the major groove[21–23]. In contrast, nucleotides amenable to substitutions have long-range interactions with flexible linker loops, such as water-mediated and electrostatic interactions[20]. At these positions, each nucleotide makes a roughly independent contribution to the long-range interaction to change the overall free energy of protein-DNA binding. Although this observation has not been reported previously for recombinase-DNA interactions, this independent base contribution has been demonstrated for other well-studied protein-DNA interactions, including transcription factor-promoter interactions, and modeled using similarly defined position weight matrices[29,30]. Despite its predictive capability, our model alone cannot elucidate why specific nucleotide substitutions tune the reaction rate. While the model highlights specific putative interactors, a high-resolution structure of the Bxb1-DNA complex would help to mechanistically interpret the model weight scores.

Experimental model validation and applications. To experimentally validate that the predicted k_{flip} values derived from DNA libraries can accurately predict the reaction rate of sequences from the library, we tested 12 individual attP-L variants from the 3000 most enriched sequences in library 2 using our in vitro inversion reaction assay. We identified a panel of attP-L variants with different predicted flipping rates, including three sequences predicted to be more efficient than WT (Fig. 5A, top), and inserted them into linear DNA fragments. To test the predictive capability of the model, we added 10 nM Bxb1 recombinase to the reaction buffer and measured the percentage of DNA flipping by qPCR for each sequence after a 5-min incubation at 30 °C. The predicted and measured flipped percentage values for these 12 sequences, including the most efficient attP-L mutant (Sequence 1), showed good correlation (Fig. 5a, bottom). We also demonstrated that our model trained by 140 random sequences (Fig. 3) can predict the reaction rates of other mutant sequences with a range over four orders of magnitude, which provides more tunability for synthetic circuits in which SSR is a key element in the gene network. To test whether the relative efficiencies of these attP-L sequences might hold in a different SSR reaction that was not the basis of selection, we also measured the reaction rates of a panel of these variants (including S1, S2, and S3 from Fig. 5a) in an intermolecular recombination experiment in vitro (Supplementary Fig. 6). Notably, despite the fundamentally different nature of this reaction requiring two distinct substrate molecules, we found that S1, S2, and S3 still had faster reaction rates than wild type, and attP-L sequences that were less efficient than wild type in intramolecular recombination had slower intermolecular reaction rates. The conservation of this trend between intramolecular and intermolecular reactions suggests broad applicability of our findings in tuning SSR rates.

To further confirm that our predictions of inversion rates translate into cells, we characterized flipping dynamics of a genetic circuit element in *E. coli*. For this experiment, we transformed a Bxb1 donor plasmid and a substrate plasmid into *E. coli* to monitor the DNA flipping process (Fig. 5b, top). We inserted three different attP-L sequences into the substrate plasmid to modulate the flipping rate of the promoter by Bxb1, which was monitored by mCherry expression (Fig. 5b, bottom).

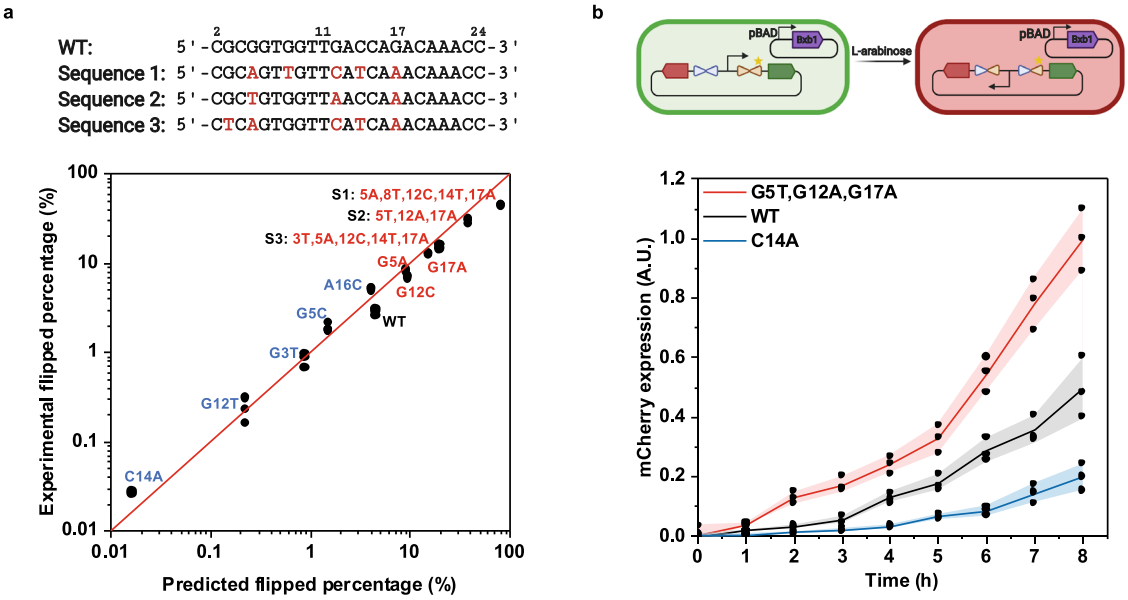

Fig. 5 Validation of individual sequences by measuring DNA flipped percentage in reaction buffer and mCherry expression in *E. coli*. a Accurate prediction of DNA flipped percentage of multiple individual sequences in the library with various predicted k_{flip} values using the linear regression model (see Fig. 3 and Supplementary Method 2). The reactions, which contained 10 nM Bxb1 and 1 nM DNA substrate, were incubated at 30 °C for 5 min and then terminated by denaturation at 80 °C for 20 min. The flipped DNA fragment concentration was measured by qPCR. The black dots represent three technical replicates. Sequences 1, 2, and 3 at the top were predicted to be the most efficient variants, each incorporating multiple mutations. **b** Three sequences with significantly different DNA flipping performances from **a**, tested in *E. coli*. Top: Donor plasmid with Bxb1 gene (purple block) and substrate plasmid are transformed into *E. coli*. DNA flipping of the promoter sequence between the GFP gene (green block) and mCherry gene (red block) gene will result in a switch from GFP expression (left) to mCherry expression (right). Bottom: The different flipping rates of these attP-L variants can be characterized by mCherry expression since it is only expressed when the promoter DNA sequence is flipped. mCherry expression level is normalized to the highest average intensity of the sequence with G5T, G12A, G17A mutations at the 8-h time point. The black dots represent three biological replicates, the lines represent the mean values of these replicates, and the shading represents one standard deviation above and below each mean. The schematic in Fig. 5b was created with BioRender.com. Source data are provided as a Source Data file.

High-copy plasmids were used as the DNA substrate so that intermediate levels of flipped plasmids could be more readily discriminated via the resulting levels of mCherry expression. Compared to wild type, the more efficient sequence (S2) rapidly expressed mCherry whereas the less efficient sequence (C14A) produced minimal fluorescence at the same growth rate, in agreement with the in vitro predictions. Thus, our model-guided DNA engineering approach can predictably tune protein expression profiles in synthetic gene circuits.

These DNA attachment sequences with programmable recombination rates can serve as useful tools for kinetic control in synthetic gene circuit design. For instance, by incorporating attP variants with predictable reaction rates into a gene circuit that co-expresses two proteins, their proportions and total expression level can be rationally tuned without the need for external control. As shown in Fig. 6a, GFP and mCherry expression levels are regulated by attP1 and attP2 recombination rates, respectively. Notably, there is product inhibition in this system, as illustrated in Fig. 6b. For example, after an attP1-mediated inversion, the intact attP2 site could still bind the Bxb1 recombinase even though a recombination event is no longer possible; the same would be true for an attP1 site after an attP2-mediated inversion. These free sites serve as decoys that reduce the effective concentration of the recombinase and, depending on the k_{flip} values for the selected attP variants, this enzyme sequestration phenomenon can modulate the total protein expression. To examine these system dynamics more quantitatively, we constructed a mathematical model based on the reactions in Fig. 6b (Supplementary Method 3) and simulated relative and total expression levels as a function of the k_{flip} values for the two attP sites (Fig. 6c). For nine selected k_{flip} combinations (numbered

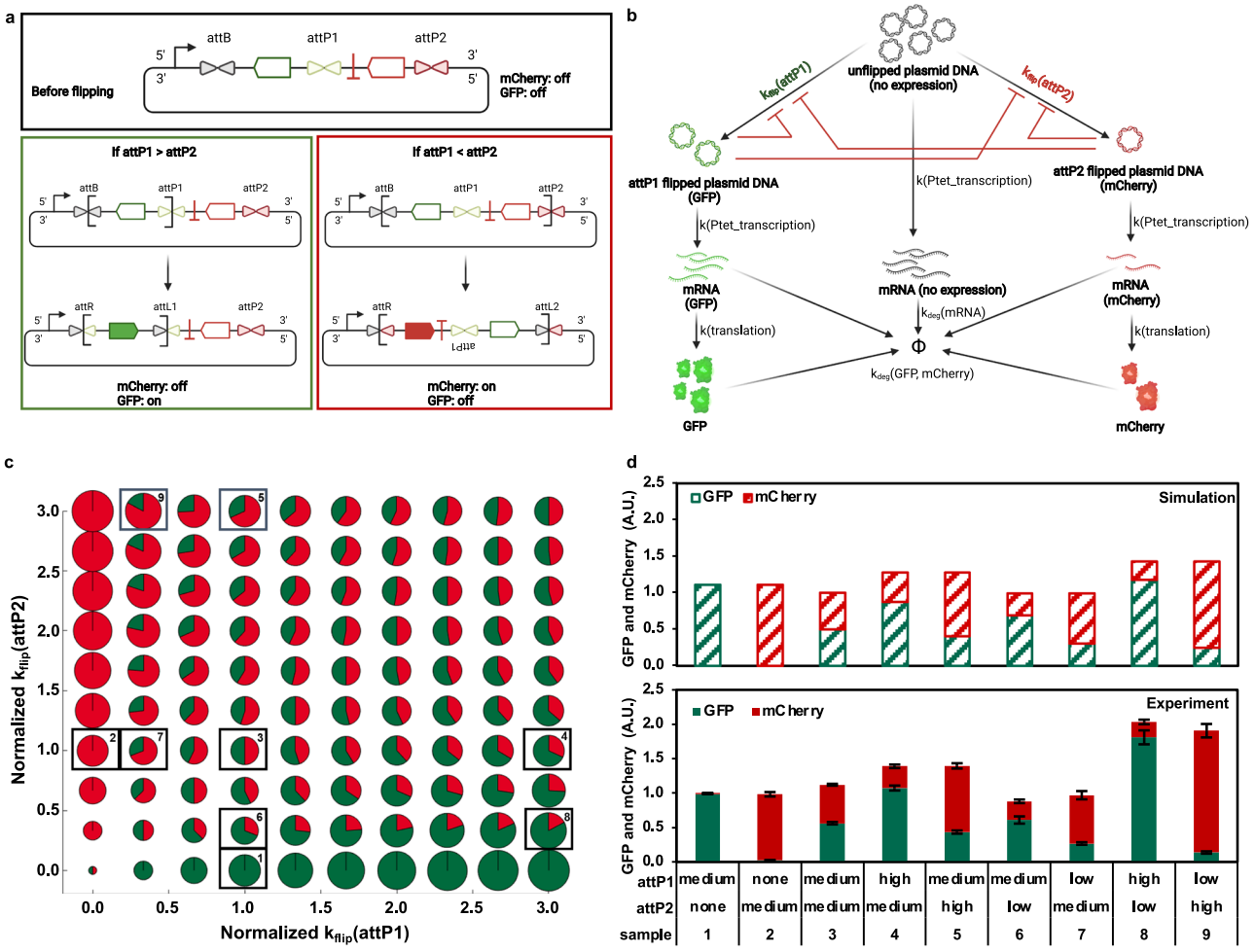

Fig. 6 Use of attP variants with predictable DNA recombination rates to coordinate co-expression of GFP and mCherry in *E. coli*. a Plasmid design for GFP and mCherry co-expression. At the top is the plasmid construct before flipping. A reversed GFP gene, flanked by attB and attP1, and a reversed mCherry gene, positioned between attP1 and attP2, are placed downstream of a promoter; attP1 and attP2 are variants with different DNA recombination rates. When the recombination rate of attP1 is larger than that of attP2 (bottom left), the GFP gene in the bracket will be preferentially flipped and transcribed. When the recombination rate of attP2 is larger than that of attP1 (bottom right), the entire construct in the bracket, which includes both the GFP and mCherry genes, will be flipped; however, due to the terminator immediately after the mCherry gene, only mCherry is expressed. The use of a high-copy plasmid ensures more robust ratiometric control of the co-expressed fluorescent proteins. **b** Schematic of the GFP/mCherry co-expression network in *E. coli*, used to construct a mechanistic model (Supplementary Method 3). **c** Bubble pie chart of simulations of the mCherry/GFP co-expression levels in *E. coli* using attP1 and attP2 variants with different DNA flipping rates. The green and red colors indicate relative proportions of GFP and mCherry expression, respectively, and the pie size indicates the total expression of fluorescent protein. The predictions tested experimentally in **d** are boxed and numbered on this pie chart. **d** Simulated (top) and experimental (bottom) GFP and mCherry expression levels for different attP1 and attP2 combinations, after 12-h incubation in M9 medium at 37 °C and normalized by cell density. Across the sample set, the low attP sequence was C14A; the medium attP sequence was WT; and the high attP sequence was S2 (G5T, G12A, G17A). The relative expression level was normalized by sample 1: WT attP1 and no attP2. Experimental data are the mean values of three biological replicates, with error bars representing standard deviations. Source data are provided as a Source Data file.

in Fig. 6c and shown as bar graphs in Fig. 6d, top), a range of relative and total expression levels are predicted. As expected, simulations with larger k_{flip}(attP1) and k_{flip}(attP2) values result in higher GFP and mCherry expression levels, respectively. More interestingly, the magnitudes of the k_{flip} values dictate the total expression level. For example, attP2 in sample 8 has a smaller recombination rate constant than that in sample 4, resulting in a smaller mCherry:GFP ratio; however, the total amount of fluorescent protein expression in sample 8 was larger than that in sample 4, due to weaker product inhibition in sample 8. These nine attP combinations were experimentally constructed and tested (Fig. 6d, bottom), showing good agreement with the model predictions in both relative and total expression levels. This example highlights the utility of differential kinetic control in gene circuit design and should have broad utility in synthetic biology to rationally tune dynamics of gene expression and memory storage.

Discussion

Previous characterization studies of site-specific recombinases have focused on the specificity of DNA recognition[21,22], but quantitative analysis of the reaction rate dependence on DNA sequence has not been reported. Our qPCR-based assay is a rapid in vitro method for capturing initial SSR reaction rates. In addition, we developed a data-driven method to predict the relative reaction rate as a function of nucleotide sequence. Using this approach, the reaction rate can be programmed over four orders of magnitude by making rational nucleotide substitutions in the attP-L half-site sequence. By combining multiple substitutions that individually make modest improvements in the recombination efficiency, we created new attP-L sequences that confer a 10-fold increase in initial reaction rate over WT (Fig. 5a). These sequences (S1: 5'-CGCAGTTGTTCATCAAACAAACC-3' and S2: 5'-CGCTGTGGTTAACCAAACAAACC-3') should be useful in improving Bxb1 recombination efficiency across a range of synthetic biology and cell engineering applications. While sequences with the highest rates of recombination could be identified from selection alone, our machine-learning model used the full set of sequencing data to more accurately predict reaction rate constants. The model revealed a fundamental biological insight that k_{flip} is composable from a linear combination of weightings of each nucleotide in the attP-L sequence, enabling precise and rational tuning of SSR rates both in vitro and in vivo. Additionally, since we determined that a much smaller library (~150 sequences) can provide convergent weight score values at each of the nine nucleotide positions in this study (Fig. 4a), future studies to analyze and engineer att sites could span more nucleotide positions at subsaturation levels of mutagenesis. Moreover, the use of model-guided design to rationally tune Bxb1 recombination reaction rates enables differential kinetic control of multiple processes by the same enzyme, which can allow for stoichiometric control of the expression of multiple proteins as shown here, and also temporal ordering of events in a synthetic genetic program without the need for multiple enzymes.

This quantitative data-driven approach can be extended to other recombinases and DNA or RNA modifying enzymes for which an appropriate selection system can be developed. Importantly, this method can be applied without a detailed mechanistic or structural understanding of the complex interaction between protein and DNA. In fact, the computed weight scores, which are derived from kinetic experiments, can reveal indirect long-range interactions that are important for the observed dynamics but may not be captured by static crystal structures. Deriving the corresponding weight score matrices for other such enzymes can therefore enhance our mechanistic understanding of their function, further broaden the genetic editing toolbox, and improve the ability to design tunable artificial circuits.

Methods

General methods. All cloning was performed in *E. coli* XL1-Blue (Agilent, 200130). *E. coli* strain BL21(DE3) was used for recombinant Bxb1 protein expression and strain TOP10 pro (F'[lacIq Tn10(tetR)] mcrA Δ(mrr-hsdRMS-mcrBC) φ80lacZΔM15 ΔlacX74 deoR nupG recA1 araD139 Δ(ara-leu)7697 galU galK rpsL(StrR) endA1 λ-, a generous gift from Dr. James J. Collins) was used for all experimental characterization of site-specific recombination. *E. coli* cells were grown in LB medium (Fisher Scientific, DF0446173), SOB medium (Teknova, S0210), or M9 medium containing M9 salt (Sigma-Aldrich, M6030), 0.4 % glycerol (Sigma-Aldrich, G7757), 0.2 % casamino acids (MP Biomedicals, 113060012), 2 mM $MgSO_4$ (Fisher Scientific, BP213), 0.1 mM $CaCl_2$ (Fisher Scientific, BP510), and 0.34 g/L thiamine hydrochloride (Sigma-Aldrich, T4625). Antibiotic concentrations of 100 μg/ml carbenicillin (Teknova, C2105) and 15 μg/ml chloramphenicol (Sigma-Aldrich, C0378) were used to maintain plasmids in *E. coli*. Isopropyl β-D-1-thiogalactopyranoside (IPTG, Millipore Sigma, 420322), anhydrotetracycline hydrochloride (aTc, AdipoGen, CDX-A0197-M500), and L-arabinose (Sigma-Aldrich, A3256) were used to induce gene expression.

Oligonucleotides were purchased from Integrated DNA Technologies. Phusion high-fidelity DNA polymerase from New England Biolabs (NEB, M0530) was used for all PCR amplifications, and PowerUp SYBR Green master mix (Thermo Fisher Scientific, A25741) was used for qPCR. T4 DNA ligase and restriction enzymes for gene cloning were purchased from NEB. Plasmid and DNA fragments were purified using kits from Qiagen according to the manufacturer's instructions. The sequences were verified by Sanger sequencing by ACGT.

Construction of the attP-L library. DNA oligonucleotides containing the randomized attP-L DNA sequences were used as the reverse primers for PCR amplification. After purification, the DNA fragment with the attP-L variants at the end was digested by HindIII-HF and ligated to another DNA fragment with the complementary HindIII-HF sticky end. Thus, the attP-L library was in the middle of a linear DNA fragment after ligation. The ligation product was purified by gel extraction and was used as the substrate for library selection. Sequence information and the DNA oligonucleotides can be found in Supplementary Table 1. The quality of the constructed naïve library was verified by NGS, as shown in Supplementary Tables 5 and 6.

Expression and purification of Bxb1 recombinase. The Bxb1 gene was cloned from a plasmid (Addgene #123132) and inserted into the expression vector pET22, which appends a C-terminal $(His)_6$ tag. Primers are listed in Supplementary Table 1. The plasmid pET22-Bxb1 was transformed into BL21(DE3) cells. A sequence-verified colony was grown in 5 ml LB medium overnight at 37 °C. Then the overnight cell culture was diluted to $OD_{600} = 0.05$ in 2x YT medium and grown at 30 °C until $OD_{600} = 0.6$. To induce Bxb1 expression, 0.4 mM IPTG was added and the incubation temperature was decreased to 16 °C. After a 15-h incubation, the cells were harvested by centrifuging the culture at 4000 *g* for 30 min.

The cell pellet was resuspended in lysis buffer and lysed by sonication. The cell lysate was then centrifuged to remove cell debris. Protein purification was performed using Ni-NTA resin from Qiagen. After equilibrating the 1 ml 50% resin using 5 ml lysis buffer (50 mM Tris-HCl, pH 8, 300 mM NaCl, 10 mM imidazole, 1 mM DTT), the cell lysate was loaded onto the column. The resin was washed with 25 ml wash buffer 1 (50 mM Tris-HCl, pH 8, 300 mM NaCl, 10 mM imidazole, 1 mM DTT) and 2.5 ml wash buffer 2 (50 mM Tris-HCl, pH 8, 800 mM NaCl, 65 mM imidazole, 1 mM DTT). Then the protein was eluted with 2 ml of elution buffer (50 mM Tris-HCl, pH 8, 800 mM NaCl, 200 mM imidazole, 1 mM DTT). Finally, the protein was rebuffered and concentrated via centrifugal ultrafiltration using a column with 10-kDa cutoff (Millipore). The purified protein was stored at −20 °C in storage buffer (50 mM Tris-HCl, pH 8, 300 mM NaCl, 1 mM DTT, 1 mM EDTA, and 50% glycerol). The protein size was verified by SDS-PAGE and the concentration was quantified by BCA assay.

In vitro recombination assays. The DNA recombination reactions were performed in PCR tubes in a reaction volume of 50 μl. Purified linear DNA fragments containing the Bxb1 attachment DNA sites in the middle were used as substrate. The reaction buffer consisted of 50 mM NaCl, 10 mM Tris-HCl, pH 7.9, 10 mM $MgCl_2$, and 100 μg/ml BSA. To test various Bxb1 concentrations, the 1 μg/μl concentrated Bxb1 stock was serially diluted to 10 times the desired final concentration using the same storage buffer, so that 5 μl 10x Bxb1 enzyme could be added to a 50 μl reaction. The reaction was incubated at 30 °C and terminated by heating at 80 °C for 20 min.

After recombination, the percentage of flipped DNA substrate was quantified using qPCR to measure the number of flipped molecules. The recombination mixture was diluted using Milli-Q water so that the total DNA concentration was 10^9 copies/μL. In a 20 μL qPCR reaction, 0.5 μM primer, 1 μL of diluted recombinant reaction mixture (10^9 copies of total DNA), and 10 μl of PowerUp SYBR Green master mix were added. PCR cycles were set up as follows: preincubation at 95 °C for 2 min, then 40 cycles were repeated with denaturation at 95 °C for 15 s, annealing at 60 °C for 15 s,

and extension at 72 °C for 15 s. Dissociation curve analysis was performed to test qPCR sensitivity. The same peak position in the dissociation curves indicated that the unflipped DNA substrate did not interfere with the PCR reaction. Through appropriate primer design, qPCR selectively amplified only the flipped DNA template. The primer sequences are listed in Supplementary Table 1.

Amplicon preparation for sequencing. Two rounds of PCR were performed to barcode the selected DNA libraries. In the first round, a forward primer and reverse primer were used to introduce a constant region that would serve as an annealing site in the second round of PCR. In the second round, the Nextera DNA CD index (Illumina) was used as the primers. The 5' end of the index was used for binding between the amplicon and the sequencing flow cell, the 3' end was used for annealing to the PCR product from the first round, and the middle 8 bases, i5 or i7, were used for barcoding different samples. DNA samples were quantified with the Quant-iT PicoGreen dsDNA detection kit (Thermo Fisher Scientific) according to the manufacturer's instructions. Sequencing was performed on an Illumina MiniSeq using a paired-end read kit following the manufacturer's instructions.

NGS data analysis. The FASTQ sequencing results generated from the MiniSeq were written to .csv files in MATLAB (Mathworks). To refine the results, reads that did not match at consistent positions due to sequencing errors were filtered out. For the remaining sequences, duplicate sequences were removed so that only unique sequences were listed. For each distinct sequence, its frequency of occurrence in the total number of refined sequences was counted. In addition, the enrichment of each of the four nucleotides at each randomized position was calculated. The MATLAB code is available on GitHub (https://github.com/sarkarlab/Bxb1-attP).

Validation in *E. coli*. The high-copy-number pUC vector was used for the construction of the GFP/mCherry reporter plasmid and the low copy number p15A-P_{BAD} vector was used for the construction of the Bxb1 donor plasmid, using restriction enzyme digestion and ligation. The primers, plasmids, and strains used are listed in Supplementary Tables 1, 2, and 3, respectively.

Each plasmid (100 ng) was transformed into 50 µL TOP10 pro competent cells at 42 °C for 45 s by heat shock. Then, 950 µL of pre-warmed SOB medium was added, followed by incubation at 37 °C and 200 rpm for 1 h. After recovery, the cells were centrifuged at 3000 *g* for 90 s and resuspended in M9 minimal medium with 50 µg/ml carbenicillin, 15 µg/ml chloramphenicol, and 0.1% L-arabinose inducer for Bxb1 expression. Cells were aliquoted into 96-well plates (200 µl/well) and incubated in a plate reader (Cytation 3 multimode reader) at 30 °C and a shaking speed of 807 rpm. Cell growth (OD_{600}), GFP expression (ex/em: 488 nm/509 nm), and mCherry expression (ex/em: 584 nm/610 nm) were monitored in real time for 8 h on the plate reader. For each transformation, two biological replicates and three technical replicates were performed. Strains transformed with different reporter plasmids had similar growth rates. GFP and mCherry expression levels were normalized to OD_{600} and subtracted from the basal fluorescence readings of the negative control (TOP10 pro cells without plasmid transformation).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

High-throughput sequencing data have been deposited in the NCBI Sequence Read Archive database under accession number PRJNA752141. Source data are provided with this paper.

Code availability

The MATLAB code used for data analysis is available on GitHub (https://github.com/sarkarlab/Bxb1-attP).

Received: 2 August 2021; Accepted: 22 June 2022;
Published online: 20 July 2022

References

1. Sauer, B. Site-specific recombination: developments and applications. *Curr. Opin. Biotechnol.* **5**, 521–527 (1994).
2. Smith, M. C. M. & Thorpe, H. M. Diversity in the serine recombinases. *Mol. Microbiol.* **44**, 299–307 (2002).
3. Wang, Y., Yau, Y. Y., Perkins-Balding, D. & Thomson, J. G. Recombinase technology: applications and possibilities. *Plant Cell Rep.* **30**, 267–285 (2011).
4. Van Duyne, G. D. & Rutherford, K. Large serine recombinase domain structure and attachment site binding. *Crit. Rev. Biochem. Mol. Biol.* **48**, 476–491 (2013).
5. Huang, L. C., Wood, E. A. & Cox, M. M. Convenient and reversible site-specific targeting of exogenous DNA into a bacterial chromosome by use of the FLP recombinase: The FLIRT system. *J. Bacteriol.* **179**, 6076–6083 (1997).
6. Zhu, F. et al. DICE, an efficient system for iterative genomic editing in human pluripotent stem cells. *Nucleic Acids Res.* **42**, e34 (2014).
7. Roquet, N., Soleimany, A. P., Ferris, A. C., Aaronson, S. & Lu, T. K. Synthetic recombinase-based state machines in living cells. *Science* **353**, aad8559 (2016).
8. Van Duyne, G. D. Cre Recombinase. *Microbiol. Spectr.* **3**, 1–19 (2015).
9. Dormiani, K. et al. Long-term and efficient expression of human β-globin gene in a hematopoietic cell line using a new site-specific integrating non-viral system. *Gene Ther.* **22**, 663–674 (2015).
10. Siuti, P., Yazbek, J. & Lu, T. K. Synthetic circuits integrating logic and memory in living cells. *Nat. Biotechnol.* **31**, 448–452 (2013).
11. Yang, L. et al. Permanent genetic memory with >1-byte capacity. *Nat. Methods* **11**, 1261–1266 (2014).
12. Brown, W. R. A., Lee, N. C. O., Xu, Z. & Smith, M. C. M. Serine recombinases as tools for genome engineering. *Methods* **53**, 372–379 (2011).
13. Nkrumah, L. J. et al. Efficient site-specific integration in *Plasmodium falciparum* chromosomes mediated by mycobacteriophage Bxb1 integrase. *Nat. Methods* **3**, 615–621 (2006).
14. Bornscheuer, U. T. et al. Engineering the third wave of biocatalysis. *Nature* **485**, 185–194 (2012).
15. Rutherford, K., Yuan, P., Perry, K., Sharp, R. & Van Duyne, G. D. Attachment site recognition and regulation of directionality by the serine integrases. *Nucleic Acids Res.* **41**, 8341–8356 (2013).
16. Li, H., Sharp, R., Rutherford, K., Gupta, K. & Van Duyne, G. D. Serine integrase attP binding and specificity. *J. Mol. Biol.* **430**, 4401–4418 (2018).
17. Keenholtz, R. A., Grindley, N. D. F., Hatfull, G. F. & Marko, J. F. Crossover-site sequence and DNA torsional stress control strand interchanges by the Bxb1 site-specific serine recombinase. *Nucleic Acids Res.* **44**, 8921–8932 (2016).
18. Jusiak, B. et al. Comparison of integrases identifies Bxb1-GA mutant as the most efficient site-specific integrase system in mammalian cells. *ACS Synth. Biol.* **8**, 16–24 (2019).
19. Gaj, T., Mercer, A. C., Gersbach, C. A., Gordley, R. M. & Barbas, C. F. Structure-guided reprogramming of serine recombinase DNA sequence specificity. *Proc. Natl Acad. Sci. USA* **108**, 498–503 (2011).
20. Siggers, T. & Gordân, R. Protein-DNA binding: complexities and multi-protein codes. *Nucleic Acids Res.* **42**, 2099–2111 (2014).
21. Singh, S., Ghosh, P. & Hatfull, G. F. Attachment site selection and identity in Bxb1 serine integrase-mediated site-specific recombination. *PLoS Genet.* **9**, e1003490 (2013).
22. Bessen, J. L. et al. High-resolution specificity profiling and off-target prediction for site-specific DNA recombinases. *Nat. Commun.* **10**, 1937 (2019).
23. Pokhilko, A. et al. The mechanism of φC31 integrase directionality: experimental analysis and computational modelling. *Nucleic Acids Res.* **44**, 7360–7372 (2016).
24. Wu, M. R. et al. A high-throughput screening and computation platform for identifying synthetic promoters with enhanced cell-state specificity (SPECS). *Nat. Commun.* **10**, 2880 (2019).
25. Bishop, C. *Pattern Recognition and Machine Learning* (Springer-Verlag, 2006).
26. Aoki, G. & Sakakibara, Y. Convolutional neural networks for classification of alignments of non-coding RNA sequences. *Bioinformatics* **34**, 237–244 (2018).
27. Quang, D. & Xie, X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.* **44**, e107 (2016).
28. Stormo, G. D. Modeling the specificity of protein-DNA interactions. *Quant. Biol.* **1**, 115–130 (2013).
29. Brewster, R. C., Jones, D. L. & Phillips, R. Tuning promoter strength through RNA polymerase binding site design in *Escherichia coli*. *PLoS Comput. Biol.* **8**, e1002811 (2012).
30. Siddharthan, R. Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix. *PLoS One* **5**, e9722 (2010).

Acknowledgements

We thank Victor Garcia for assistance with next-generation sequencing, and Ayako Ohoka and Jennifer Kang for helpful discussions. This work was supported by the National Institutes of Health (R01DK114453 to S.M.A. and C.A.S. and R35GM136309 to C.A.S.).

Author contributions

Q.Z., S.M.A., and C.A.S. designed the research; Q.Z. performed the research; Q.Z., S.M.A., and C.A.S. analyzed the data; and Q.Z., S.M.A., and C.A.S. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-31538-3.

Correspondence and requests for materials should be addressed to Casim A. Sarkar.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.