Nucleic Acids Research, 2025, **53**, gkaf691
https://doi.org/10.1093/nar/gkaf691
NAR Breakthrough Article

Identification of cognate recombination directionality factors for large serine recombinases by virtual pulldown

Heewhan Shin[1], Alexandria Holland[2], Abdulrazak Alsaleh[2], Alyssa D. Retiz[1], Ying Z. Pigli[1], Oluwateniola T. Taiwo-Aiyerin[2], Tania Peña Reyes[1], Adebayo J. Bello[2], Jialiang Quan[1], Weixin Tang[1], Femi J. Olorunniji[2,*], Phoebe A. Rice[1,*]

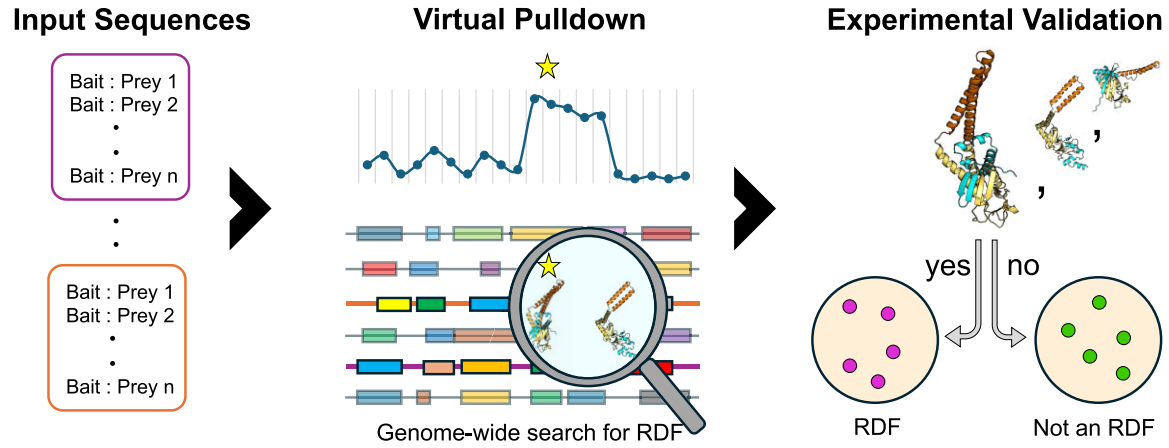
[1]Department of Biochemistry & Molecular Biology, The University of Chicago, Chicago, IL 60637, United States
[2]School of Pharmacy and Biomolecular Sciences, Faculty of Health, Innovation, Science, and Technology, Liverpool John Moores University, James Parsons Building, Byrom Street, Liverpool L3 3AF, United Kingdom

*To whom correspondence should be addressed. Email: F.J.Olorunniji@ljmu.ac.uk
Correspondence may also be addressed to Phoebe A. Rice. Email: price@uchicago.edu

Abstract

Integrases from the "large serine" family are simple, highly directional site-specific DNA recombinases that have great promise as synthetic biology and genome editing tools. Integrative recombination (mimicking phage or mobile element insertion) requires only integrase and two short (~40–50) DNA sites. The reverse reaction, excisive recombination, does not occur until it is triggered by the presence of a second protein termed a recombination directionality factor (RDF), which binds specifically to its cognate integrase. Identification of RDFs has been hampered due to their lack of sequence conservation and lack of synteny with the phage integrase gene. Here we use AlphaFold2-multimer to identify putative RDFs for more than half of a test set of 98 large serine recombinases, and experimental methods to verify predicted RDFs for 6 of 9 integrases chosen as test cases. We find no universally conserved structural motifs among known and predicted RDFs, yet they are all predicted to bind a similar location on their cognate integrase, suggesting convergent evolution of function. Our methodology greatly expands the available genetic toolkit of cognate integrase–RDF pairs.

Graphical abstract



Introduction

Large serine recombinases (LSRs), which are encoded by many lysogenic bacteriophages (and some other mobile genetic elements), have great potential as genetic tools [1–11]. In their natural setting, this family of site-specific DNA recombinases catalyzes unidirectional site-specific recombination between an ~50-bp bacteriophage ("phage") attachment site (*attP*) and an ~40-bp bacterial site (*attB*), resulting in the insertion of the phage genome into the host chromosome (Fig. 1A). The resulting prophage can then be passively replicated as part of the bacterial chromosome [12]. Upon activation of the phage's lytic phase, a second phage-encoded protein, the recombination directionality factor (RDF), binds the integrase protein and alters its preferred reaction direction to greatly

Received: February 25, 2025. **Revised:** June 26, 2025. **Editorial Decision:** June 27, 2025. **Accepted:** July 18, 2025

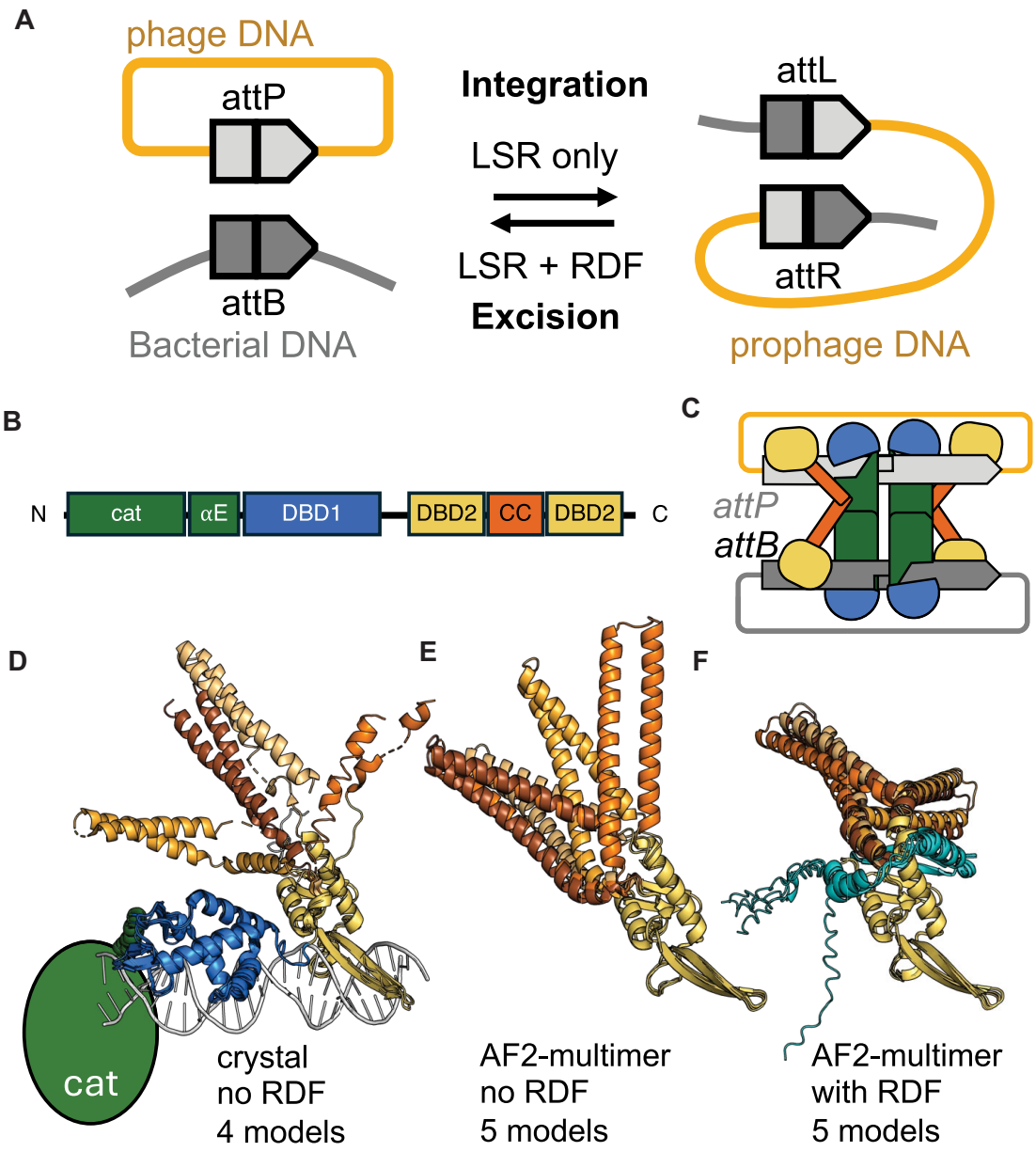


Figure 1. Recombination mechanism and domain structure. (**A**) Integrative recombination requires only the LSR and its cognate *att* sites but reverse reaction (excisive recombination) requires the presence of a second protein, the cognate RDF. (**B**) Linear cartoon of the domain structure of LSR domain structure. CAT: catalytic domain; αE: long helix connecting the CAT to the DNA-binding domains (DBDs); DBD1: DNA-binding domain 1 (also known as recombinase domain); DBD2: DNA-binding domain 2 (also known as Zn-binding domain); CC: coiled coil. (**C**) Cartoon of the proposed structure of the substrate complex for integrative recombination, as proposed by Van Duyne *et al.* [38]. (**D**) Structure of the DBDs of LI integrase bound to half an *attP* site [40]. The proteins from the four copies in the asymmetric unit are superimposed on one another, the second half of DBD2 as a guide. Domains are colored as in panel (C) except that varying shades of orange are used for the CCs to highlight their flexibility. (**E**) AlphaFold2-multimer modeling of LI integrase. DBD2 and the CC for five models are superimposed and colored as in panel (D). (**F**) AlphaFold2-multimer modeling of LI integrase in complex with its RDF [20]. The RDF (turquoise) partially restrains the flexibility of the CC when compared to the experimental or the AlphaFold2-multimer models in the absence of the RDF.

favor the excision reaction [12–14]. Although LSRs are readily identified in genomic sequences due to their conserved sequence motifs, RDFs are not, and the relatively few that are known were identified primarily through painstaking genetic work [14–26]. No conserved sequence motifs are apparent among them, nor do they show consistent synteny with their cognate integrase genes. In fact, some are moonlighting DNA replication proteins [27, 28]. Here we describe and experimentally verify a new AlphaFold2-multimer-based method for rapid identification of RDFs—essentially, a virtual pulldown approach [29–31].

Many features of LSRs render them useful as genetic tools. Unlike CRISPR–Cas systems, they have evolved to catalyze the insertion of large payloads, and their recombination mechanism leaves the DNA product with not a single broken phosphodiester bond. In contrast to bacteriophage integrases from the mechanistically very different tyrosine family [32], serine integrases do not require host proteins and their *attP* sites are much smaller [33]. Although LSRs lack the programmable sequence specificity of RNA-guided systems, their lack of target flexibility can be alleviated by a "drag-and-drop" procedure that uses a CRISPR–Cas-derived system to insert an LSR's *attB*

site, or by choosing from the large array of already characterized LSRs with differing sequence specificity [9, 10, 32, 34]. This toolkit was greatly expanded by the recent publication of a list of over 60 LSRs that can catalyze insertions into the human genome [10]. Furthermore, for some applications, such as the SIRA method for assembly of large replicons, the availability of multiple LSRs with orthogonal sequence specificity is a key feature that facilitates multiplexing [35, 36]. When the cognate RDFs are known for particular LSRs, their versatility as tools expands significantly. For example, RDF-mediated reaction reversal can be used for modular editing of assembled replicons, and LSR–RDF pairs can be used to create living logic gates [6].

LSRs constitute a branch of the "serine" family of site-specific DNA recombinases, which share a conserved catalytic domain and overall mechanism of recombination (Fig. 1A) [37]. Each crossover site in the DNA is bound by a dimer of the recombinase, after which the two DNA-bound dimers are brought together by a regulatory apparatus that varies widely among systems but is an intrinsic part of the crossover site-bound protein subunits in the LSR case. The serine nucleophile in the active site of each subunit then attacks a particular phosphodiester bond, displacing a 3′ OH to create a reaction intermediate in which both DNA duplexes have double-strand breaks with 2-nt 3′ overhangs, and each 5′ end is covalently linked to a recombinase subunit. The recombinase tetramer then swivels internally to realign the broken ends, and the DNAs are religated by the reverse of the DNA cleavage reaction. Serine recombinases have modular domain organizations. Members of the "large" branch carry the catalytic domain at their N-termini, followed by two DBDs. For simplicity, we refer to these as DBD1 and DBD2, although DBD1 is often termed "recombinase domain (RD)" and DBD2 "zinc-binding domain (ZD)."

The favored reaction direction for LSRs is determined by energetic differences between the substrate and product conformations of the protein–DNA complexes rather than by a net change in chemical bond energy [37]. Protein–protein interactions are mediated not only by the catalytic domains but also by a coiled coil (CC) with a hydrophobic tip that is inserted within DBD2 [38, 39]. Whether CC–CC interactions among subunits preferentially stabilize synaptic tetramers or product dimers is determined by the positioning of DBD2, which differs between *attP* and *attB* sites, and by the presence or absence of the RDF (Fig. 1B; see [13 40] for details of the mechanism). For integration, CC–CC interactions between dimers bound to *attP* and *attB* stabilize a synaptic tetramer, but after recombination they rearrange to conformationally lock the *attL*- and *attR*-bound product dimers [13, 41]. For excision, the RDF changes the trajectories of the CCs [24, 42], and CC–CC interactions can stabilize synapsis between *attL*- and *attR*-bound dimers, and then rearrange to preferentially stabilize at least the *attP*-bound product dimers and possibly the *attB*-bound ones [13].

Biochemical studies found that RDFs bind to the CC and/or DBD2 of their cognate integrases [20, 23, 24], and our recent structural studies of the phage SPbeta integrase system show that RDF binding anchors the DBD2-proximal portion of the CC to DBD2, locking out flexibility in a hinge at the CC–DBD2 junction [13]. This interaction redirects the trajectory of the CC but still allows flexibility in a second hinge closer to the CC tip.

No other experimental structures are currently available for integrase–RDF complexes. However, recent AI-based advances in protein structure prediction can now help to fill these knowledge gaps [29, 30]. We found that AlphaFold2-multimer predicts a similar binding site at the DBD2–CC junction for known LSR–RDF pairs, despite the lack of any conserved structural motif among those RDFs.

We then automated AlphaFold2-multimer to efficiently predict RDFs for LSRs whose RDFs were previously unknown, using the large collections of active LSRs identified by Durrant *et al.* [10] and Yang *et al.* [34] as test cases. Finally, we showed that RDFs for six of the nine integrases picked for *in vivo* verification do indeed function as predicted, including two functional RDFs for one of these integrases. We further verified the function of two of these pairs *in vitro* using purified proteins. The virtual pulldown workflow described here will lead to identification of the RDFs for many known integrases and for those yet to be discovered. This will give access to a larger pool of integrase–RDF pairs available for fundamental studies on the reaction mechanism of the integrase–RDF system, and for building orthogonal integrase-based genetic circuits for synthetic biology applications.

Materials and methods

Virtual pulldowns

Preparation of paired input files for AlphaFold2-multimer predictions

We used the collections of LSRs organized by Durrant *et al.* and Yang *et al.* as test cases for our procedure [10, 34]. Those two studies were chosen because they provided a large but defined number of test cases with putative *att* sites and with information on *in vivo* activity. For each LSR listed in their supplementary tables, we obtained the corresponding genomic data (DNA and protein coding sequences) from the National Center for Biotechnology Information database. To find the *attL* and *attR* sites that mark the ends of the inserted form of the prophage or mobile element, we utilized the first 15–20 and last 15–20 nucleotides of the *attB* and *attP* sequences given in [10, 34] to search within the genome sequence and identify the prophage region, and in some cases corroborated that range using Phaster [43].

The protein coding sequences within the identified prophage region were extracted as prey sequences. For the bait sequence, we predicted the LSR's structure with ESMFold (https://esmatlas.com/resources?action=fold; [44]) and then truncated the sequence to include only the second DNA-binding domain (DBD2). The truncated "bait" LSR sequence was paired with each prey sequence in a FASTA format for AlphaFold2-multimer predictions. In some cases where the initial pulldown failed to produce a clear "hit," it was repeated using both DBDs or the intact LSR. Python scripts used for this and later steps can be found in Supplementary data.

Multimer predictions

LocalColabFold v.1.5.2 was installed on Beagle-3, a shared GPU cluster at the University of Chicago's Research Computing Center, following the steps described at https://github.com/YoshitakaMo/localcolabfold [29, 30, 45]. All predictions were performed using default parameters except that we increased number of prediction recycles from 3 to 5 "–use-gpu-relax –num-recycle 5 –num-models 5." We note that the general power of the virtual pulldown approach was independently proposed by Yu *et al.* [31] while we were developing our method and experimentally testing our results. How-

ever, theirs is a more generalized approach that requires a pre-prepared list of protein sequences as input, and they did not focus on our question (RDF identification) nor experimentally test any newly proposed protein–protein interactions.

Assessment of output files

LocalColabFold generates predicted aligned error (PAE) plots with predicted scores and models (in JSON and PDB format, respectively). For efficient assessment, all PAE plots were concatenated; pTM and ipTM scores were extracted from the output JSON files, and these values were plotted using gnuplot. pTM is the predicted template modeling score for the individual protein structure models and the ipTM the interface pTM for complexes; the higher (closer to 1) scores indicate more confident predictions [29, 30, 46]. Predicted models exhibiting low PAE and high pTM and ipTM scores were visually evaluated in PyMOL. Initial hits were chosen as those prey proteins that gave the highest ipTM value, which usually exceeded the pTM value. Potential hits were rejected if they (i) were not predicted to interact with DBD2, (ii) were predicted to interact with the tip of the CC, as that hydrophobic patch is known to be required for self interactions, or (iii) were predicted to interact with the DNA-binding surface of DBD2 [39]. The latter proteins were most likely DNA mimics involved in the host–virus arms race, and our results suggest that virtual pulldowns may provide a new way to discover such proteins [47].

Structure figures

All structure figures were made using PyMOL [48].

Testing of predictions

To experimentally test our predictions, we chose a total of nine integrases: Nm60, Bt24, Cb16, Dn29, Enc3, Pc01, and Pa03 from the Durrant *et al*. list as well as Int10 and Int30 from the Yang *et al*. list [10, 34]. These examples were chosen based on the reported activity of the integrase, or intriguing features of the predicted RDFs such as unusually large size (Pa03 and Pc01), the possibility of two RDFs for the same integrase (Ints 10 and 30), or an unusual predicted interaction with the DBD2-proximal portion of the CC (Enc3; see the "Results" section for more details).

To experimentally test the activities of the predicted RDFs, we fused the RDF to the C-terminus of the integrase using flexible peptide linkers. We have previously shown that such integrase–RDF fusions ensure the 1:1 stoichiometry required for synapsis and recombination [3, 27, 49]. This approach avoids the risk of uneven expression of the two proteins if they were expressed from different vectors as separate proteins. In addition, integrase–RDF fusions ensure optimal binding of the RDF to the integrase, which could result in a false negative if the RDF has low binding affinity to the integrase. Candidate RDFs were tested using a slight variation of the *in vivo* inversion assay that we and others have previously used to study integrase recombination reactions *in vivo* [49–51] (see diagram in Fig. 6). Each assay requires two co-transformed plasmids: (i) a test plasmid carrying the *att* sites for the integrase in question (in inverted orientation) flanking a promoter that drives expression of either green fluorescent protein (GFP) or red fluorescent protein (RFP), depending on its orientation, and (ii) an expression vector for the integrase or integrase–RDF fusion of interest.

The activity of two integrase–RDF pairs (Nm60 and Int30) and was additionally verified in *in vitro* assays.

Vectors encoding integrase and integrase–RDF fusion proteins

Coding sequences for all integrases and integrase–RDF fusions were cloned between NdeI and XhoI sites in pBAD33, which carries an arabinose-inducible promoter, a p15a origin, and chloramphenicol resistance. The use of an inducible expression system prevented potential issues with toxicity of constitutively expressed integrases, and the use of integrase–RDF fusions (previously described for known integrase–RDF pairs [52]) alleviated potential issues with the stoichiometry of the two proteins that could occur if they were expressed separately. In the fusion constructs, each integrase and its putative RDF were covalently joined together using an 18-residue linker (TSGSGGSGGSGGSGRSGT) between the C-terminal residue of the integrase and the second amino acid residue in the candidate RDF. The amino acid sequences of the proteins of interest were reverse-translated and the DNA sequences codon-optimized for expression in *Escherichia coli* and ordered as gene fragments from Twist Biosciences. Each integrase gene has a SpeI site before the stop codon, which adds the first two amino acids (TS) of the linker sequence to the C-terminus of the protein. RDF genes with the linker added to their N-termini were then cloned in-frame between the SpeI and XhoI sites in the integrase vectors.

All new plasmids were verified by sequencing.

Prediction of core attP and attB recombination sites

To determine the *attP* and *attB* sequences for integrases of interest, we used the information provided by Durrant *et al*. and Yang *et al*. [10, 34] in combination with genomic sequences. Figure 5 shows the four half-sites for each integrase, using the half-site sequence alignment approach of Van Duyne and Rutherford to identify potential binding motifs for the two DBDs and to determine the most likely location of the central dinucleotide [38]. The outer ends of the *att* sites used in our test vectors each contained at least 5 additional bp of genomic sequence beyond what is shown in Fig. 5.

In vivo recombination reactions

In vivo recombination assays (inversion) of each integrase–putative RDF pair were carried out using the invertible promoter reporter system depicted in Fig. 6A. The plasmid substrates were based on the previously reported pϕC31-invPB and pϕC31-invRL ([50]; see Figs 4 and 6). The backbone of these vectors carries a pSC101 origin and a kanamycin resistance marker that are compatible with those on our protein expression vectors. The recombination sites (*attP* and *attB*) for each integrase were cloned into pϕC31-invPB to replace the *att* sites for ϕC31 integrase using gene fragments (Twist Biosciences) covering the entire ∼560-bp invertible segment. Substrate plasmids containing *attL* and *attR* for each integrase were made by integrase-mediated recombination.

Escherichia coli DS941 strains [53] containing the substrate plasmid (kanamycin selection) and the arabinose-inducible expressing vector (chloramphenicol) for each integrase or integrase–RDF fusion were prepared in advance such that recombination activities can be initiated when integrase expression is induced by addition of arabinose to a growing culture. To prepare the recombination strain, competent *E. coli* DS941 cells were transformed with the inversion substrate plasmid and the integrase expression vector and grown for 16 h in LB

media in the presence of kanamycin (50 μg/ml) and chloramphenicol (25 μg/ml) selection.

To assay *attP* × *attB* recombination activity of each integrase, fresh overnight culture of each recombination strain was grown at 37°C in LB media to mid-log phase in the presence of kanamycin (50 μg/ml) and chloramphenicol (25 μg/ml). Integrase or integrase–RDF expression was induced for 2 h by addition of 0.2% arabinose, after which cultures were diluted 1:1000 in fresh LB media containing 0.2% glucose to repress further integrase expression and grown for 16 h with shaking at 37°C. Cultures were diluted 1/100 000 in fresh LB media and spread on LB agar plates containing kanamycin (50 μg/ml). Slight variations were later introduced in assaying Int10, Dn29, Enc3, Pc01, and Pa03: cells were grown in 0.2% glucose to repress expression before switching to arabinose for induction, induction was continued overnight, and DNA from the overnight culture was miniprepped and retransformed before plating.

Recombination of *attP* and *attB* sites inverts the orientation of the promoter, thereby switching gene expression from RFP to GFP (Fig. 6B), and generating *attR* and *attL* sites. Colonies expressing GFP or RFP were imaged in a Gel Doc™ imaging system (Bio-Rad, Hercules, CA, USA). The differences in fluorescence intensity seen are most likely primarily due to the different *att* sites placed between the promoter and the coding region, some of which can efficiently hairpin and may terminate transcription or translation.

Following recombination, individual colonies expressing GFP were grown in liquid culture for 16 h with kanamycin selection. Plasmid DNAs were extracted from the liquid culture and separated on 1.0% agarose gel. Bands corresponding to supercoiled substrate (inversion products bearing *attR* × *attL* sites) were cut out of the gel, purified, and used to retransform *E. coli* cells. Plasmid DNA samples were subsequently extracted from the cultures and sequenced to verify that they contain the expected *attR* and *attL* sites for each integrase. The verified recombination products of *attP* × *attB* reactions (*attR* × *attL* plasmid substrates) were then used for the experiments where the activities of the candidate RDFs were tested. In those assays, recombination of *attR* × *attL* sites leads to inversion of the promoter resulting in expression of RFP, while stopping GFP production (Fig. 6).

Protein expression and purification

The expression and purification of Nm60 integrase and Nm60 integrase–RDF fusion was as described previously [52]. The DNA sequences encoding the integrase and the integrase–RDF fusion were cloned between NdeI and XhoI sites in pET28a(+) and the plasmids were used to transform *E. coli* BL21(DE3)pLysS strain. The strain for each protein was grown at 37°C in LB to an optical density of 0.8, before cooling the cultures to 20°C and protein expression induced with 0.5 mM IPTG (Isopropyl β-d-1-thiogalactopyranoside). Protein expression was allowed to continue for 16 h at 20°C. Each protein was purified by nickel affinity chromatography and bound proteins were eluted with an imidazole gradient buffer system. Samples of fractions corresponding to peaks were analyzed on sodium dodecyl sulfate–polyacrylamide gel electrophoresis and chosen fractions containing the desired proteins were dialyzed against protein dilution buffer (25 mM Tris–HCl, pH 7.5, 1 mM DTT (dithiothreitol), 1 M NaCl and 50% glycerol), and stored at −20°C. Dilutions of the integrase and the integrase–RDF fusion were made into the same buffer for *in vitro* recombination reactions.

In vitro recombination reactions

Recombination assays of excisive (*attR* × *attL*) and integrative (*attP* × *attB*) activities were carried out on the same plasmid substrates used for the *in vivo* reactions (Fig. 7A). Plasmid DNA substrates for Nm60 integrase were grown in *E. coli* DS941 and purified using Qiagen miniprep kit. *In vitro* recombination of supercoiled plasmid substrates and analysis of recombination products were carried out using conditions similar to those described previously [36, 49]. Typically, recombination reactions were carried out by adding integrase (0.5–1.0 μM, 5 μl) to a 30 μl solution containing the plasmid substrate (25 μg/ml), 50 mM Tris–HCl (pH 7.5), 100 μg/ml BSA, 5 mM spermidine, and 0.1 mM ethylenediaminetetraacetic acid. Reactions were allowed to proceed at 30°C for 2 h, after which the integrases were denatured by heating the reaction at 80°C for 10 min to stop all recombination activities. The samples were cooled and treated with the restriction endonuclease XhoI (New England Biolabs) to facilitate analysis of recombination products. Following the restriction endonuclease treatment, the reaction products were treated with 0.1% sodium dodecyl sulfate and protease K before reaction products were separated by 1.2% agarose gel electrophoresis [36, 49].

Results

Previously identified RDFs: variable structures bind a conserved LSR feature

We first examined several previously identified LSR–RDF pairs, using AlphaFold2-multimer to predict the structures of their complexes. The close match between the predicted and experimental structure for SPbeta lends confidence to this approach (Fig. 2) [13]. Because protein structure can be more conserved than sequence, we had expected to find a small conserved structural motif within the RDFs that would mediate interaction with their cognate LSRs. Figure 2 shows that we did not: some RDFs are predicted to utilize only alpha helices as interaction motifs (e.g. phiRv1), some to use both helices and beta strands (e.g. Bxb1 and SPbeta), and some to use loops (e.g. phiC31). However, despite sharing no structural homology at all, these disparate RDFs were all predicted to bind DBD2 at or near its junction with the CC. Confidence in these predictions can be seen in the PAE in Supplementary Fig. S1. Supplementary Fig. S2 shows that these models also agree with existing biochemical data [14, 23, 24].

Virtual pulldowns identify known RDFs

Given the difficulties in sequence-based RDF prediction and the apparent success of AI-based structure predictions, we asked whether AlphaFold2-multimer could be used to identify RDFs. The examples shown in Fig. 2 were chosen as positive controls. Structures for each and every individual ORF (open reading frame) within a given phage were predicted in complex with DBD2 of the relevant LSR (see the "Materials and methods" section). For four out of these five examples, the complex predicted with the highest confidence was indeed that with the previously identified RDF (based on the ipTM score) [29, 30, 46]. Predictions for the fifth, phage Bxb1, were ambiguous until the procedure was repeated using both DBDs

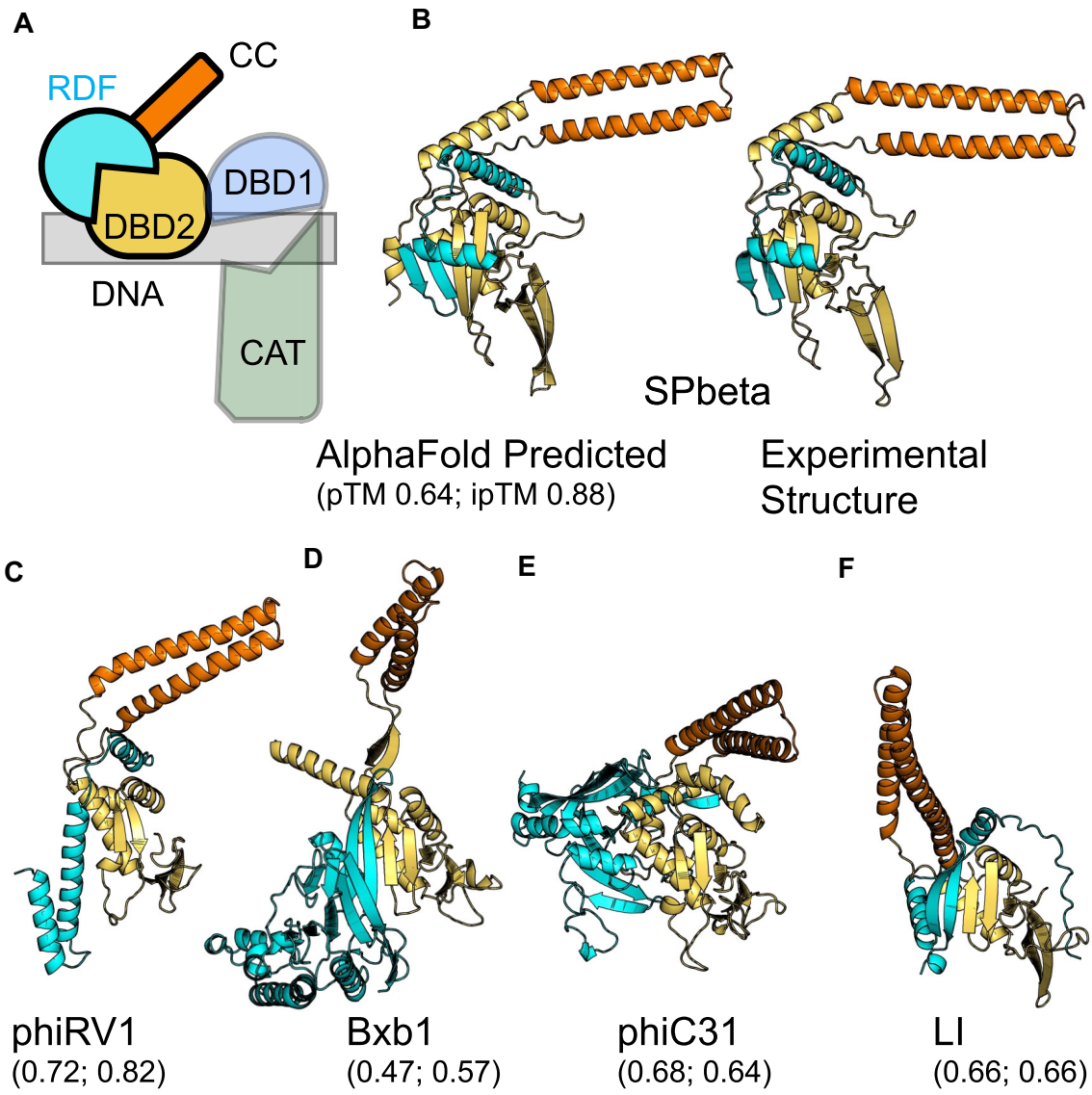


Figure 2. Predicted structures for known LSR–RDF pairs. (**A**) Cartoon similar to those in Fig. 1, but in the same relative orientation as the structural cartoons in panels (B–F). (**B**) AlphaFold-predicted versus experimental structure of SPbeta Int–RDF complex. (**C**–**F**) Predicted complexes of previously identified RDFs in complex with their cognate integrases. Only DBD2 and the CC inserted within it are shown for the LSR partner but the entire predicted structure for each RDF is shown.

as "bait" even though the RDF is only predicted to interact with DBD2. For consistency, the virtual pulldowns shown for all five examples in Fig. 3 and Supplementary Fig. S3 were carried out using the last 400 amino acids of the LSR, which ensured including both DBDs in the bait.

Prediction of new RDFs by virtual pulldown

Next we used our virtual pulldown procedure to try to predict RDFs for all of the LSRs listed in supplementary table 2 of Durrant *et al.* [10] as well as the 34 new LSRs identified by Yang *et al.* [34]. Of the 68 active integrases listed by Durrant *et al.*, we could not find the appropriate mobile element or prophage range for three (Ec06, Kp03, Sa10). In our first round, in which we used only DBD2 of each of the remaining 65 LSRs as prey, we were able to make confident predictions in 33 cases (Supplementary Table S1). In examining the failures, we noted that one integrase (Efs2) was in fact a DDE recombinase [54] not an LSR, and that many have additional structured protein at their C-termini that would be expected to block the RDF binding region noted above. Noting our findings for the Bxb1 test case, we repeated the failed pulldowns using a larger fragment of the integrase protein as prey, resulting in 10 more RDFs predicted with moderate to high confidence.

In total, we were able to predict putative RDFs for 43 of the 64 integrases listed by Durrant *et al.* [10], and to speculatively predict that at least some of the 24 of the remaining integrases may not in fact have a cognate RDF, in some cases perhaps due to substantial C-terminal extensions of the integrase protein that block the potential RDF binding site. The predicted complexes chosen for experimental testing are shown in Fig. 4, with corresponding PAE plots in Supplementary Fig. S4.

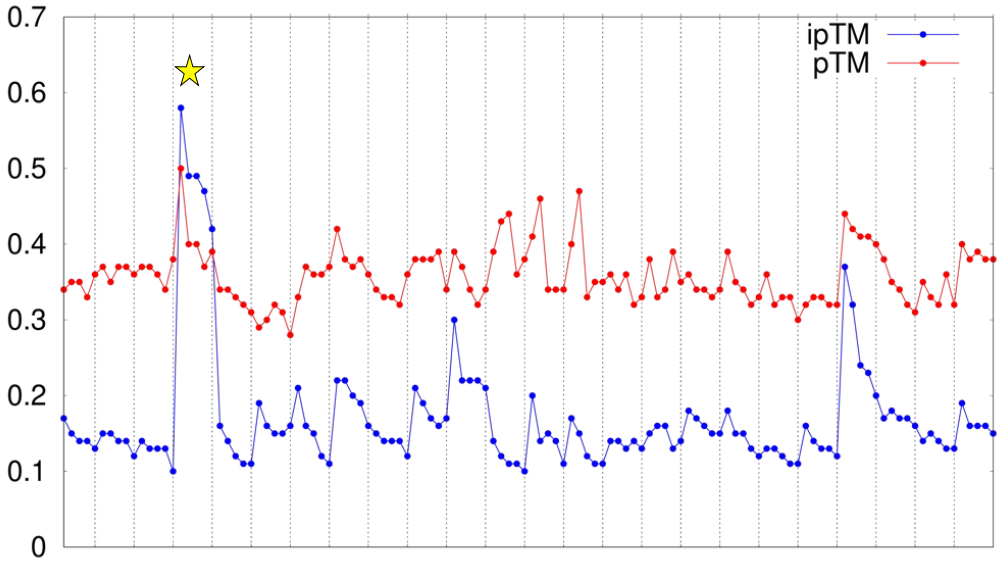


Figure 3. A scatter plot of the virtual pulldown for the phage-like element A118. Five models were predicted for the complex of each element-encoded protein with the A118 integrase (dotted vertical lines separate sets of models). The plot shows two measures of confidence for each—pTM (red) and ipTM (blue) scores—which were calculated using AlphaFold2-multimer v3 model in ColabFold v.1.5.2 [30, 45, 46]. The highest-ranking model for the known RDF is highlighted with a yellow star. See Supplementary Fig. S3 scatter plots for the other LSR–RDF pairs shown in Fig. 2 and for the PAE plots for the each of these known LSR–RDF pairs.

Many of the LSRs (57% or 38 out of 67) were from genomic elements that Phaster deemed unlikely to be bacteriophages—although Phaster may have mis-categorized them, they may instead be other genomic islands with different control mechanisms. However, some of those LSRs for which we were able to confidently predict an RDF also appeared to be from non-bacteriophage mobile genetic elements. We were also able to predict putative RDFs for many of the 34 integrases by Yang *et al.* (Supplementary Table S1). In some cases, our procedure predicted more than one putative RDF, in which case all are listed.

The predicted RDFs adopt a variety of folds (Fig. 4). They are clearly not predicted to fully restrain the mobility of the CC. However, we noted a strong tendency of the RDF to interact not only with DBD2 itself but also with the DBD2-proximal segment of the CC (which is sometimes replaced with beta strands; e.g. see Bxb1 in Fig. 2), suggesting that it may partially restrain it, as seen experimentally for the SPbeta RDF [13].

Experimental verification of newly predicted RDFs

To verify our method, we focused on predicted RDFs for nine integrases: seven from the Durrant *et al.* list (Nm60, Bt24, Cb16, Dn29, Enc3, Pc01, and Pa03) and two from the Yang *et al.* list (Int10 and Int30). Nm60, Bt24, and Cb16 were the most active integrases shown in fig. 2I of [10] for which we could predict a cognate RDF. Int10 and Int30 were chosen because two potential RDFs were predicted for both, and we were curious as to whether or not both were active with their corresponding integrase. Pa03 and Pc01 were chosen because of the unusually large size of their putative RDFs. Enc3 was chosen because of an unusual mode of RDF interaction: the DBD2-proximal portion of the CC, like that of Bxb1 (Fig. 1), is predicted to be a pair of beta strands, and the Enc3 RDF is predicted to add a third strand to that pair (Supplementary Fig. S5). Pairwise sequence identity among these test integrases was mostly under 20%, and the two most closely related pairs were Nm60 and Int30 (56% identical) and Bt24 and Dn29 (55% identical) (Supplementary Fig. S5). However, even these pairs showed differences in the conserved nucleotides found in their predicted *att* sites (Fig. 5). While the RDFs for these pairs of integrases also shared sequence homology (Supplementary Fig. S5), attempts to align larger collections of RDF sequences failed.

Confirmation of integrase and att site functionality

First, we verified the functionality of our integrase and substrate constructs. Figure 6B (panel 1) shows the activities of Nm60 integrase and its fusion recombinase on *attP* × *attB* substrate. The results show Nm60 integrase mediated the complete switching of gene expression from RFP to GFP following inverting the orientation of the promoter as a result of *attP* × *attB* recombination reaction (Fig. 6A). Similar results were obtained for the other integrases tested (Supplementary Fig. S6). This confirms that the core sequences that we identified are correct and functional for *attP* × *attB* recombination.

Following successful *attP* × *attB* recombination by the integrases, we isolated the recombinant DNA plasmid product and used this as *attR* × *attL* substrate for the excisive recombination reaction. The plasmids were sequenced to ascertain *attR* and *attL* sites were formed and to confirm the central dinucleotide crossover point we deduced based on half-site alignment (Fig. 5). The sequenced *attR* × *attL* sites for the tested integrases agreed with the predicted crossover points. Finally, we found that, as expected of LSRs, the integrases were essentially inactive on their cognate *attR* × *attL* substrates (Fig. 6 and Supplementary Fig. S6).

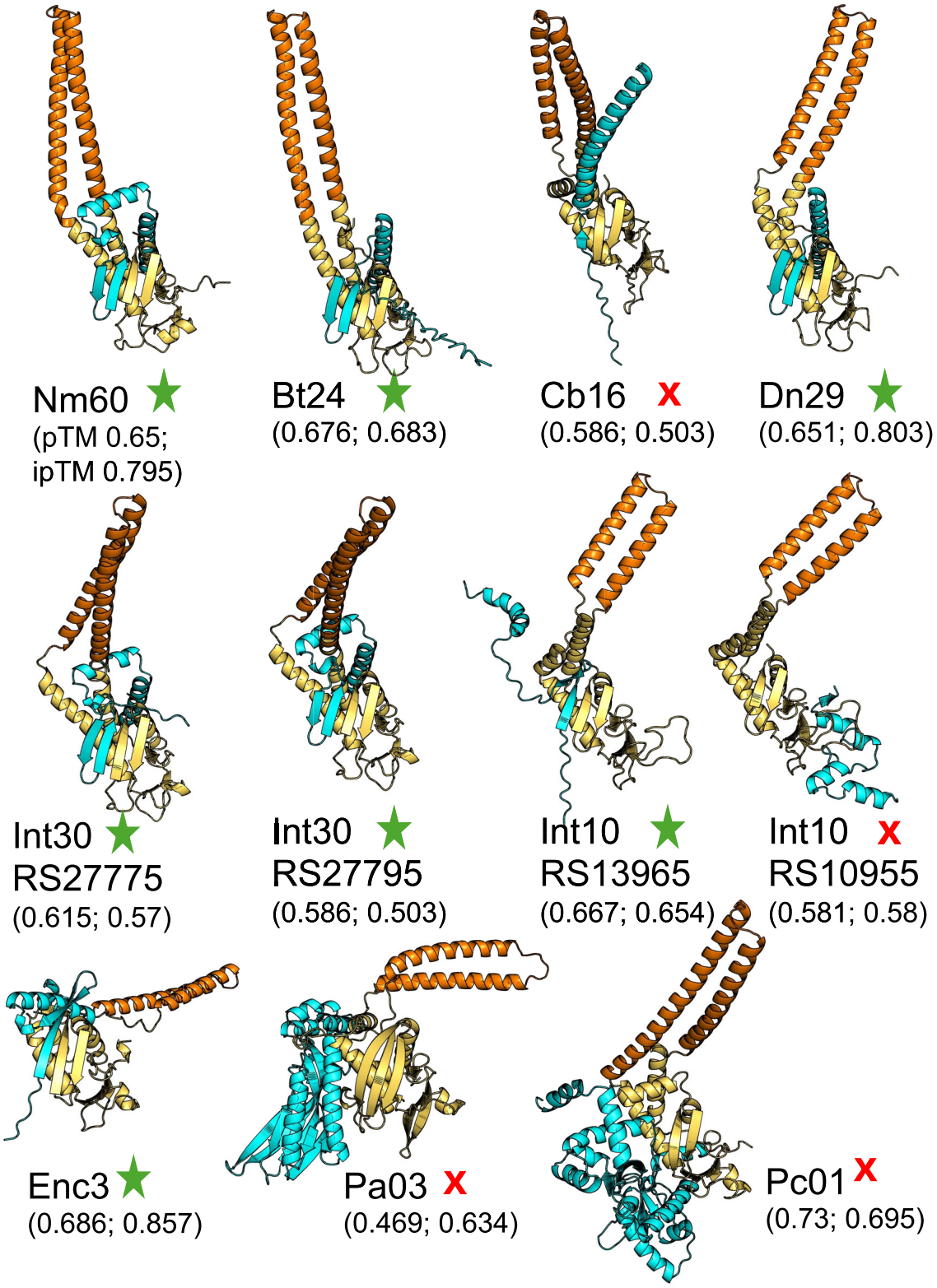


Figure 4. Predicted DBD2–RDF complex structures for the 11 LSR–RDF pairs tested experimentally. DBD2 of the LSR is shown in yellow, the CC in orange, and the RDF in cyan. If more than one potential RDF was predicted, a partial locus tag is added to disambiguate. Functional RDFs are marked with stars and false positives with X's. PAE plots for each pair are shown in Supplementary Fig. S5. The confidence scores are given beneath each model; closer to 1 denotes higher confidence.

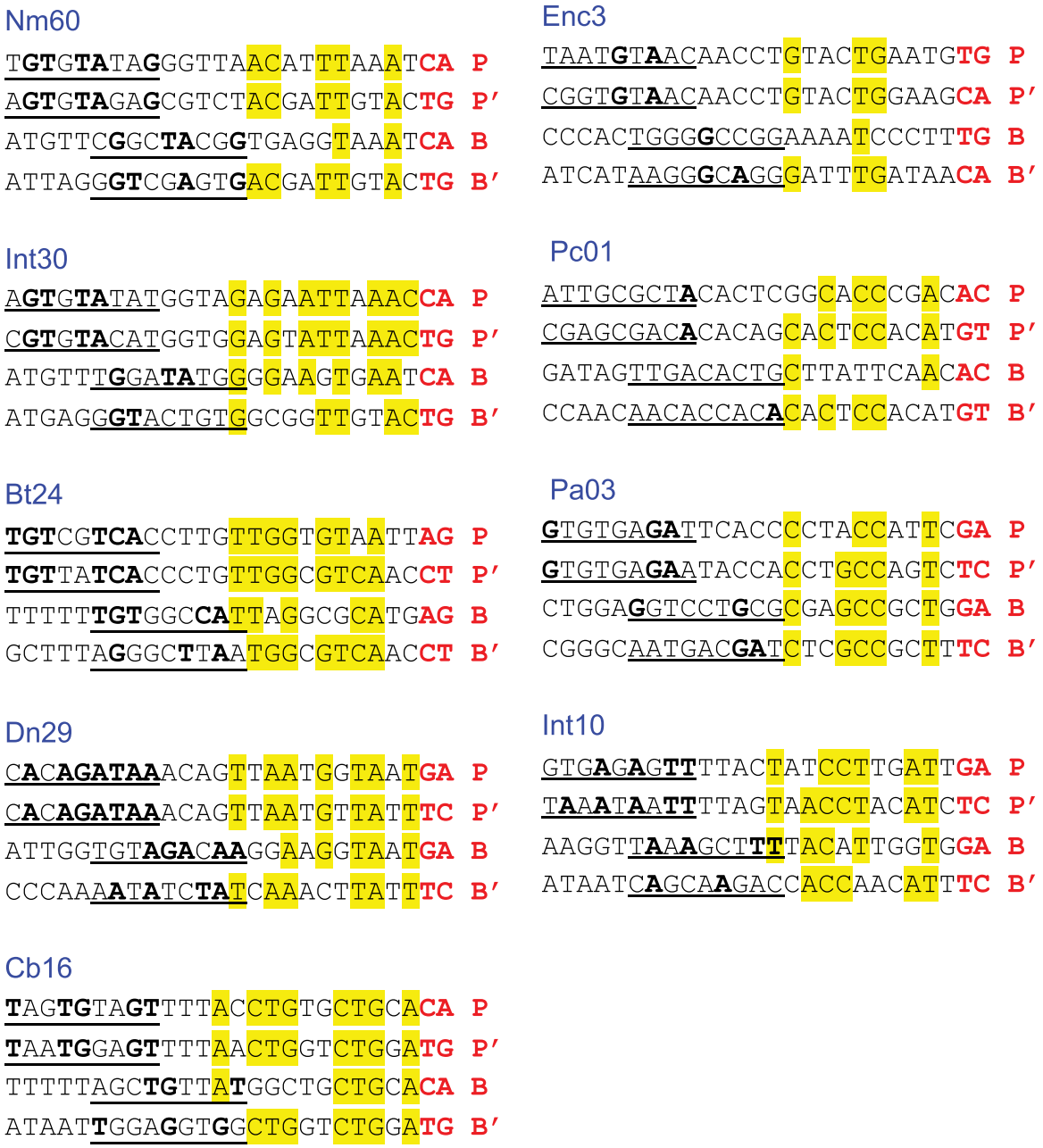


Figure 5. Alignment of LSR half-site sequences. The central dinucleotides are shown in bold and red font. DBD1 (also called RD) is expected to bind the innermost 12 bases, and DBD2 (also called ZD) is expected to bind the outer underlined segment. Bases that are conserved in three or four positions are highlighted in yellow in the DBD1 motif and in bold in the DBD2 motif. This convention follows [38].

In vivo assay of RDF function

A functional RDF should inhibit the *attP* × *attB* recombination by its cognate LSR and activate *attL* × *attR* recombination [12]. Figure 6B and Supplementary Fig. S6 show that this is indeed the case for the single predicted RDFs for Nm60, Bt24, Dn29, and Enc3, for both predicted RDFs for Int30, and for one of the two predicted for Int10. For experiments using these LSR–RDF fusion constructs, the final colonies were predominately red regardless of whether the starting substrate was the *attP* × *attB* one (red) or the *attL* × *attR* one (green). However, the predicted RDFs for Cb16, Pc01, and Pa03, and one of the two predicted Int10 RDFs failed to trigger *attL* × *attR* recombination. For the Cb16 and the second Int10–RDF fusions, the final colonies were predominately green regardless of which starting substrate was used (i.e. they behaved as if no RDF was present). For the Pc01 and Pa03 integrase–RDF fusions, the final colony color was primarily that of the starting plasmid, indicating that fusion of these larger proteins to the integrase's C-terminus inhibited integrase activity. Overall, these results confirm the efficacy of our virtual pulldown approach, with the caveat that 4 of 11 tested hits appear to be false positives.

In all *in vivo* experiments, plasmids were extracted and sequenced to confirm that the *att* sites in the product plasmids are as expected and in agreement with orientation of the promoter directing the expression of GFP or RFP as shown in Fig. 6. Results were fully consistent with the predictions in Fig. 5.

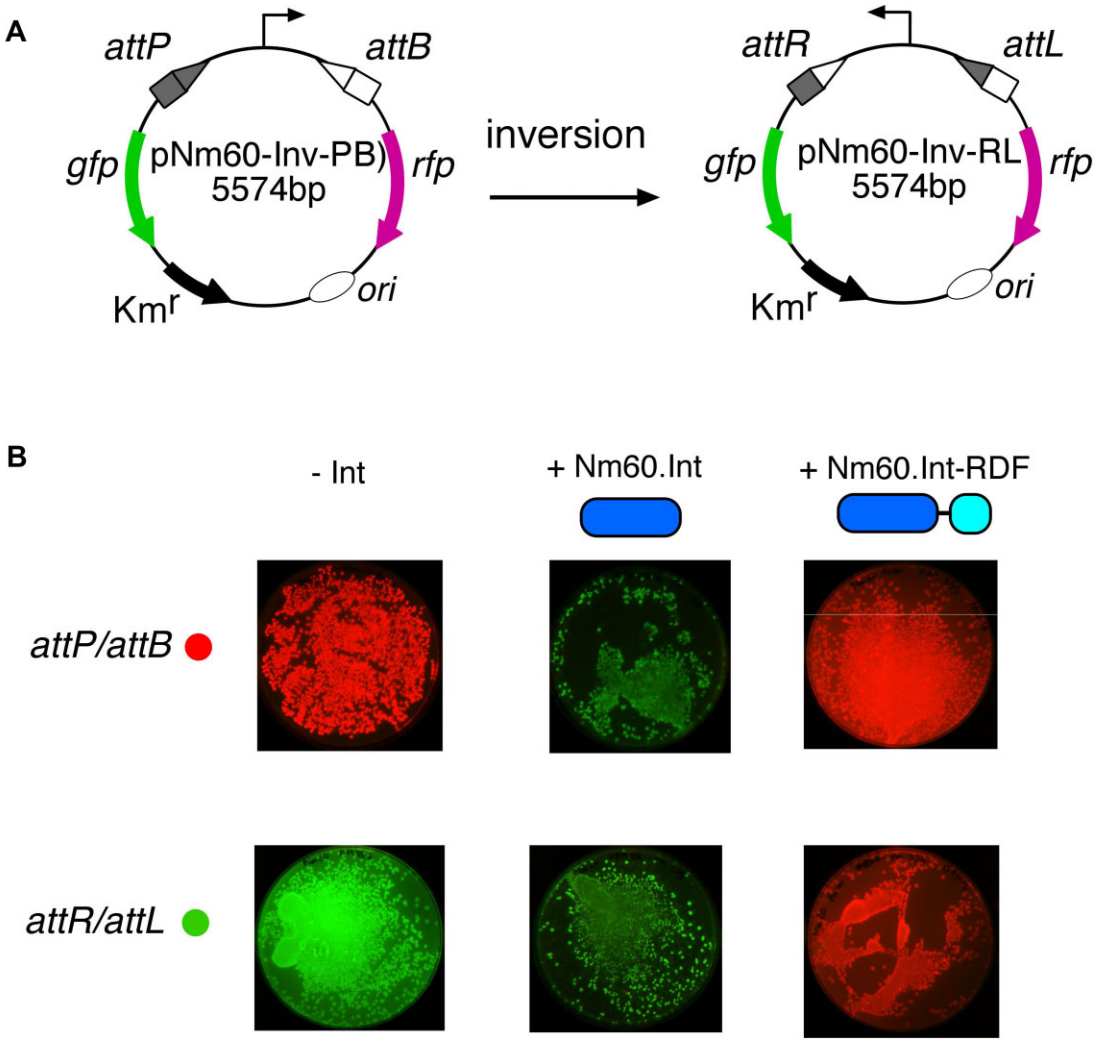


Figure 6. *In vivo* recombination reactions of Nm60 integrase and its fusion with the RDF. (**A**) Schematic illustration of the *in vivo* recombination (inversion) assay. In the plasmid pNm60-invPB, the promoter, flanked by *attP* and *attB* sites, constitutively drives the expression of an *rfp* gene (pink arrow). To prevent transcriptional read-through to the *gfp* gene (green arrow), a terminator sequence is inserted upstream of the promoter. LSR-catalyzed recombination (inversion) of *attP* × *attB* to give *attR* × *attL* products (plasmid pNm60-invRL) flips the orientation of the promoter to allow the expression of GFP, and block RFP production. (**B**) Recombination activities of Nm60 integrase and its fusion to the RDF on *attP* × *attB* (pNm60-invPB) and *attR* × *attL* (pNm60-invRL) substrates. The integrase is depicted as a long oval (blue) and the RDF as a short oval (cyan).

A secondary screen for false positives

Our virtual pulldown used only the second DBD, in the absence of DNA, as bait. We wondered whether some of the false-positive binding partners might interfere with formation of a full integrase monomer–DNA complex even though they were not predicted to bind to the core of the DBD2 DNA binding surface. AlphaFold3 can now predict protein–DNA complexes, although the code is not available for automation, and it docks proteins randomly with regard to DNA sequence [55]. We therefore asked whether AlphaFold3 would predict similar RDF–DBD2 interactions if given full sequences for the integrase, the putative RDF, and a 56-nt random-sequence DNA duplex. Of the 11 test cases shown in Fig. 4, the predicted RDF–DBD2 interactions were similar to those shown in Fig. 4 for all of the true positives and for the false positive Cb16, but not for the other three false positives (Int10 RS10955, Pa03, and Pc01). AlphaFold3 also failed to properly dock the known RDFs for Bxb1 and PhiC31 onto their respective DBD2s, but it should be noted that these are special cases where the RDF is a moonlighting DNA-binding protein [27]. With that caveat in mind, manually rescreening hits from a virtual pulldown using AlphaFold3 could therefore help eliminate false positives.

In vitro recombination activities of Nm60 integrase and Nm60 integrase–RDF fusion

Next, we tested the *in vitro* recombination activities of Nm60 integrase and its integrase–RDF fusion in both *attP* × *attB* and *attR* × *attL* reactions. The proteins were expressed and purified as previously described for ϕC31 and Bxb1 integrases and their RDF fusions [52]. As shown in Fig. 7A, recombination of pϕNm60-invPB (*attP* × *attB*) to give pϕNm60-invRL (*attR* × *attL*) is accompanied by inversion of the DNA segment flanked by the *att* sites. Treatment of the recombination reaction product with the restriction endonuclease, XhoI, gave distinct restriction patterns for recombined and non-recombined DNAs.

The activities of the two proteins on *attP* × *attB* substrate are shown in Fig. 7B (first panel). Nm60 integrase catalyzed the conversion of the substrate to the product, with more ac-

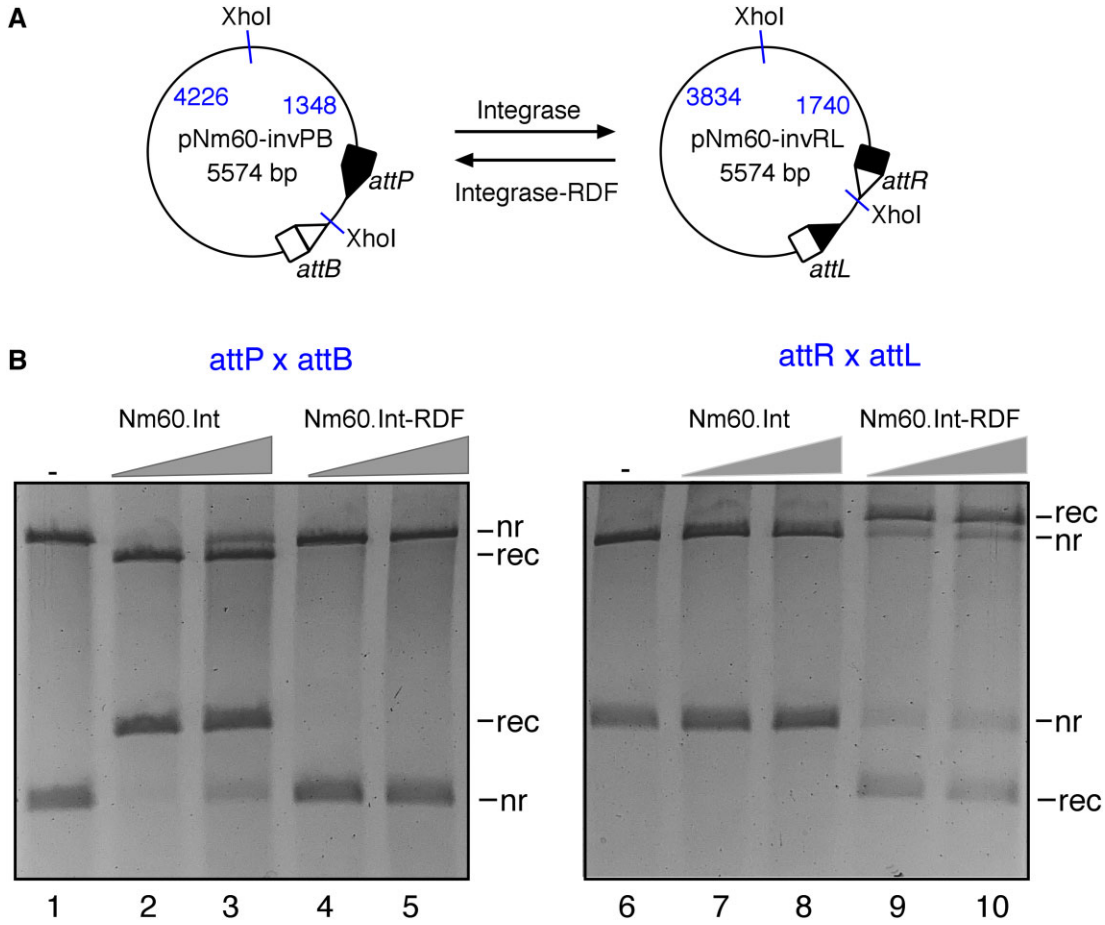


Figure 7. *In vitro* recombination reactions of Nm60 integrase and its fusion with the RDF. (**A**) Schematic illustration of the *in vitro* recombination (inversion) assay. LSR-catalyzed recombination (inversion) of *attP* and *attB* sites in pNm60-invPB gives rise to *attR* and *attL* sites in the product plasmid (pNm60-invRL), and vice versa. (**B**) Reactions were carried out for 2 h as described in the "Materials and methods" section. Integrase and integrase–RDF fusions were used at two final concentrations of 50 and 100 nM. Reaction products were digested with the restriction endonuclease, XhoI, prior to 1.2% agarose gel electrophoresis. As illustrated in panel (A), digestion of the plasmids with XhoI gives different restriction patterns for recombined and non-recombined DNAs. The bands on the gel are labeled *nr* (non-recombinant, i.e. substrate) and *rec* (recombination product).

tivity seen in the 50 nM integrase reaction than the one with 100 nM integrase. As expected, fusion of the RDF to the integrase inhibited *attP* × *attB* recombination with no DNA corresponding to the recombinant product visible on the gel. The second panel shows the results of the *attR* × *attL* reaction. As expected, purified Nm60 integrase failed to catalyze the reaction, while the integrase–RDF fusion recombined *attR* × *attL* to give *attP* × *attB*. These *in vitro* recombination activities agree with the *in vivo* results described above (Fig. 6). Similar results were found for purified Int30 and Int30–RDF fusion proteins (Supplementary Fig. S7).

Discussion

Using AlphaFold2-multimer to perform virtual pulldowns, we identified putative RDFs for ∼60% of our test set of integrases (Supplementary Table S1) [10]. For six of the nine integrases chosen for experimental testing, at least one predicted RDF stimulated *attL* × *attR* recombination and inhibited *attP* × *attB* recombination, as expected of an RDF. For Int30, two functional RDFs were found. Because they have highly similar sequences and predicted structures, it is unsurprising that both were active as RDFs. However, it is surprising that a single phage encodes two functional RDFs.

This demonstrates the feasibility of our approach, which can obviate the need for painstaking genetic experiments to find the cognate RDF for a given LSR. It also adds to the growing body of reports demonstrating the power of the AI-based virtual pulldown approach [31]. In fact, a similar approach was reported in finding an RDF for the Yin element of *Clostridium botulinum* [56].

Although 4 of the 11 putative new RDFs were inactive when tested, we found that secondary screening using AlphaFold3 could have removed three of these from the testing pipeline. The PAE plots for these three predictions were not the most convincing of the tested set—they were chosen instead for potentially interesting features: Pc01 and Pa03 because the unusually large size of the predicted RDFs suggested that they could be moonlighting proteins with additional non-RDF functions, and Int10 RS10955 was tested because it was a second hit for the same integrase. In comparing the two potential RDFs tested for Int10, we noted that the non-functional one, while predicted to interact with DBD2, was not predicted to interact with the DBD2–CC junction region as seen for the functional RDFs. It remains unclear how we could have eliminated the fourth false positive, the non-functional predicted RDF for Cb16: it passed the secondary screen, is predicted to bind DBD2–CC junction, and the ipTM

of 0.503 for that prediction, while lower than that for many, was the same as that for one of the functional Int30 RDFs.

The two predicted RDFs for Int30 are highly similar in their sequences and predicted structures. Hence, it is unsurprising that both proteins were active as RDFs, although it is surprising that a single phage encodes two functional RDFs.

There are many possible explanations for why we could not confidently predict RDFs for some of the LSRs in our test set. First, AlphaFold2-multimer may have returned false negatives, especially for unusual sequences due to a shallow multiple sequence alignment or for unusual cases of moonlighting RDFs—i.e. cases where both proteins have multiple homologs in the database, but most of the integrase homologs use a different RDF, and thus the sequence databases lack a sufficient co-evolutionary signature. Second, our bait lists only included ORFs found between the ends of the integrase-encoding phage or mobile element, but the relevant RDF may be encoded elsewhere, as seen for the PLE2 element [26]. Finally, there may not be an RDF: e.g. the LSRs encoded by the staphylococcal SCC*mec* element can catalyze recombination with roughly equal efficiency on a variety of *att* site pairings [57]. How that element's excision and integration are regulated is poorly understood [57]. We also note that although the *attP* × *attB* recombination activity of most of the LSRs in our test set was previously verified [10, 34], their activity on a full range of *att* site pairs remains untested.

AlphaFold2 models of the SCC*mec* LSRs showed that they do have additional beta strands at the C-terminus of DBD2 that could block the RDF binding site here. However, further studies are needed to determine whether the existence of such C-terminal extensions is an accurate predictor of a lack of a cognate RDF and a lack of directionality, or whether some LSRs with such extensions have found an alternate solution to directionality.

Access to a larger pool of characterized LSR–RDF pairs

The virtual pulldown workflow and predictions described here will greatly facilitate identification of the RDFs for known integrases and for those yet to be discovered. This will give access to a larger pool of integrase–RDF pairs available for fundamental studies on the reaction mechanism of the integrase–RDF system. Insights gained from studying several structurally diverse RDFs and how they interact with their cognate integrases could provide a better understanding of the quintessential properties of an RDF. This knowledge could be used to iteratively create a set of structure and property profiles to find common themes among the highly variable RDFs and their LSR interactions, and thus a better understanding of what is important for optimal function.

AlphaFold2-multimer predicts that despite their structural diversity, all the verified RDFs that we looked at are predicted to bind their cognate integrase at the DBD2–CC junction (Figs 2 and 4) while one of the false positives was predicted to bind elsewhere on DBD2, further supporting the hypothesis that restraining or altering the trajectory of the CC is critical to RDF function [13, 24]. In comparing the predicted RDF–DBD2 interactions, we find no universal interaction motif. Most but not all add at least one additional beta strand to DBD2's beta sheet—a common protein–protein interaction motif that may serve as an "anchor" to hold the RDF to the core of DBD2. Additionally, most but not all are predicted to include a segment that could interact with the DBD2-proximal portion of the CC. Given the difficulties that AlphaFold has in predicting which conformation of a flexible protein occurs in which context, and that it can under-predict the presence of hinges within helices or fraying at the ends of helices, experimental structural information is required to confidently predict exactly how these RDFs alter the DNA-bound complexes of their cognate integrases. We envisage that the availability of a wide range of known RDFs will provide a foundation for the development of a universal method for designing synthetic RDFs or RDF-independent integrases engineered to catalyze the excisive *attR* × *attL* recombination. Additional structural information will be required to achieve these goals.

Genetic circuits and logic gates

The potential applications of large serine integrases in building genetic circuits and logic gates have been explored in prokaryotic and eukaryotic systems [6, 7, 51, 58–60]. These examples have been built using the same set of few characterized LSRs and their RDFs. The new RDFs characterized here and the virtual pulldown protocol for identifying new ones will enable the design and testing of multiplex genetic circuits built from orthogonal LSR–RDF modules. It is anticipated that these constructs will be used in designing more complex cellular operations with applications in engineering biology.

Acknowledgements

Author contributions: Heewhan Shin (Investigation [equal], Methodology [equal], Software [lead], Writing—review & editing [supporting]), Alexandria Holland (Investigation [equal]), Abdulrazak Alsaleh (Investigation [equal]), Alyssa D. Retiz (Investigation [equal]), Ying Z. Pigli (Investigation [equal]), Oluwateniola T. Taiwo-Aiyerin (Investigation [equal]), Tania Peña Reyes (Investigation [equal]), Adebayo J. Bello (Investigation [equal]), Jialiang Quan (Investigation [supporting]), Weixin Tang (Investigation [supporting]), Femi J. Olorunniji (Conceptualization [equal], Funding acquisition [equal], Methodology [equal], Writing—review & editing [lead]), and Phoebe A. Rice (Conceptualization [equal], Funding acquisition [equal], Methodology [equal], Writing—original draft [lead]).

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

This work was supported by the National Science Foundation and UK Research and Innovation (collaborative grants NSF/BIO 2223480 and UKRI/BBSRC BB/X012085/1 to P.A.R. and F.J.O.). Funding to pay the Open Access publication charges for this article was provided by the BBSRC BB/X012085/1.

Data availability

The data underlying this article have been deposited in Mendeley (output of AlphaFold2-multimer); VirtualPulldown_predicted_structures_full_length_LSI (https://data.mendeley.com/datasets/ky6ptcbr6g/1), VirtualPulldown_predicted_structures_DBD2 (https://data.mendeley.com/datasets/x8b3zwn2pf/1), and Int1_34 (https://data.mendeley.com/datasets/8ktg473htr/1). Plasmids used for the *in vivo* assays are being deposited with Addgene.

References

1. Olorunniji FJ, Rosser SJ, Stark WM. Site-specific recombinases: molecular machines for the Genetic Revolution. *Biochem J* 2016;**473**:673–84. https://doi.org/10.1042/BJ20151112
2. Colloms SD, Merrick CA, Olorunniji FJ *et al.* Rapid metabolic pathway assembly and modification using serine integrase site-specific recombination. *Nucleic Acids Res* 2014;**42**:e23. https://doi.org/10.1093/nar/gkt1101
3. Olorunniji FJ, Merrick C, Rosser SJ *et al.* Multipart DNA assembly using site-specific recombinases from the large serine integrase family. *Methods Mol Biol* 2017;**1642**:303–23.
4. Huang H, Chai C, Yang S *et al.* Phage serine integrase-mediated genome engineering for efficient expression of chemical biosynthetic pathway in gas-fermenting *Clostridium ljungdahlii*. *Metab Eng* 2019;**52**:293–302. https://doi.org/10.1016/j.ymben.2019.01.005
5. Roquet N, Soleimany AP, Ferris AC *et al.* Synthetic recombinase-based state machines in living cells. *Science* 2016;**353**:aad8559. https://doi.org/10.1126/science.aad8559
6. Bonnet J, Yin P, Ortiz ME *et al.* Amplifying genetic logic gates. *Science* 2013;**340**:599–603. https://doi.org/10.1126/science.1232758
7. Gomide MS, Sales TT, Barros LRC *et al.* Genetic switches designed for eukaryotic cells and controlled by serine integrases. *Commun Biol* 2020;**3**:255. https://doi.org/10.1038/s42003-020-0971-8
8. Anzalone AV, Gao XD, Podracky CJ *et al.* Programmable deletion, replacement, integration and inversion of large DNA sequences with twin prime editing. *Nat Biotechnol* 2022;**40**:731–40. https://doi.org/10.1038/s41587-021-01133-w
9. Yarnall MTN, Ioannidi EI, Schmitt-Ulms C *et al.* Drag-and-drop genome insertion of large sequences without double-strand DNA cleavage using CRISPR-directed integrases. *Nat Biotechnol* 2023;**41**:500–512. https://doi.org/10.1038/s41587-022-01527-4
10. Durrant MG, Fanton A, Tycko J *et al.* Systematic discovery of recombinases for efficient integration of large DNA sequences into the human genome. *Nat Biotechnol* 2023;**41**:488–99. https://doi.org/10.1038/s41587-022-01494-w
11. Murphy KC, Nelson SJ, Nambi S *et al.* ORBIT: a new paradigm for genetic engineering of mycobacterial chromosomes. *mBio* 2018;**9**:e01467-18. https://doi.org/10.1128/mBio.01467-18
12. Smith MCM. Phage-encoded serine integrases and other large serine recombinases. *Microbiol Spectr* 2015;**3**. https://doi.org/10.1128/microbiolspec.MDNA3-0059-2014
13. Shin H, Pigli Y, Peña Reyes T *et al.* Structural basis of directionality control in large serine integrases. bioRxiv, https://doi.org/10.1101/2025.01.03.631226, 13 January 2025, preprint: not peer reviewed.
14. Ghosh P, Wasil LR, Hatfull GF. Control of phage Bxb1 excision by a novel recombination directionality factor. *PLoS Biol* 2006;**4**:e186. https://doi.org/10.1371/journal.pbio.0040186
15. Ramaswamy KS, Carrasco CD, Fatma T *et al.* Cell-type specificity of the *Anabaena* fdxN-element rearrangement requires *xisH* and *xisI*. *Mol Microbiol* 1997;**23**:1241–9. https://doi.org/10.1046/j.1365-2958.1997.3081671.x
16. Breüner A, Brøndsted L, Hammer K. Novel organization of genes involved in prophage excision identified in the temperate lactococcal bacteriophage TP901-1. *J Bacteriol* 1999;**181**:7291–7. https://doi.org/10.1128/JB.181.23.7291-7297.1999
17. Bibb LA, Hatfull GF. Integration and excision of the *Mycobacterium tuberculosis* prophage-like element, φRv1. *Mol Microbiol* 2002;**45**:1515–26. https://doi.org/10.1046/j.1365-2958.2002.03130.x
18. Khaleel T, Younger E, McEwan AR *et al.* A phage protein that binds φC31 integrase to switch its directionality. *Mol Microbiol* 2011;**80**:1450–63. https://doi.org/10.1111/j.1365-2958.2011.07696.x
19. Zhang L, Zhu B, Dai R *et al.* Control of directionality in *Streptomyces* phage φBT1 integrase-mediated site-specific recombination. *PLoS One* 2013;**8**:e80434. https://doi.org/10.1371/journal.pone.0080434
20. Mandali S, Gupta K, Dawson AR *et al.* Control of recombination directionality by the *Listeria* phage A118 protein Gp44 and the coiled-coil motif of its serine integrase. *J Bacteriol* 2017;**199**:e00019-17. https://doi.org/10.1128/JB.00019-17
21. Abe K, Shimizu S-Y, Tsuda S *et al.* A novel non prophage(-like) gene-intervening element within gerE that is reconstituted during sporulation in *Bacillus cereus* ATCC10987. *Sci Rep* 2017;**7**:11426. https://doi.org/10.1038/s41598-017-11796-8
22. Serrano M, Kint N, Pereira FC *et al.* A recombination directionality factor controls the cell type-specific activation of σ^K and the fidelity of spore development in *Clostridium difficile*. *PLoS Genet* 2016;**12**:e1006312. https://doi.org/10.1371/journal.pgen.1006312
23. Abe K, Takahashi T, Sato T. Extreme C-terminal element of SprA serine integrase is a potential component of the "molecular toggle switch" which controls the recombination and its directionality. *Mol Microbiol* 2021;**115**:1110–21. https://doi.org/10.1111/mmi.14654
24. Fogg PCM, Younger E, Fernando BD *et al.* Recombination directionality factor gp3 binds φC31 integrase via the zinc domain, potentially affecting the trajectory of the coiled-coil motif. *Nucleic Acids Res* 2018;**46**:1308–20. https://doi.org/10.1093/nar/gkx1233
25. Suzuki S, Yoshikawa M, Imamura D *et al.* Compatibility of site-specific recombination units between mobile genetic elements. *iScience* 2020;**23**:100805. https://doi.org/10.1016/j.isci.2019.100805
26. McKitterick AC, Seed KD. Anti-phage islands force their target phage to directly mediate island excision and spread. *Nat Commun* 2018;**9**:2348. https://doi.org/10.1038/s41467-018-04786-5
27. Alsaleh A, Holland A, Shin H *et al.* Large serine integrases utilise scavenged phage proteins as directionality cofactors. *Nucleic Acids Res* 2025;**53**:gkaf050. https://doi.org/10.1093/nar/gkaf050
28. Savinov A, Pan J, Ghosh P *et al.* The Bxb1 gp47 recombination directionality factor is required not only for prophage excision, but also for phage DNA replication. *Gene* 2012;**495**:42–8. https://doi.org/10.1016/j.gene.2011.12.003
29. Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9. https://doi.org/10.1038/s41586-021-03819-2
30. Evans R, O'Neill M, Pritzel A *et al.* Protein complex prediction with AlphaFold-multimer. bioRxiv, https://doi.org/10.1101/2021.10.04.463034, 10 March 2022, preprint: not peer reviewed.
31. Yu D, Chojnowski G, Rosenthal M *et al.* AlphaPulldown—a Python package for protein–protein interaction screens using AlphaFold-Multimer. *Bioinformatics* 2023;**39**:btac749. https://doi.org/10.1093/bioinformatics/btac749
32. Stark WM. Making serine integrases work for us. *Curr Opin Microbiol* 2017;**38**:130–6. https://doi.org/10.1016/j.mib.2017.04.006
33. Thorpe HM, Smith MCM. *In vitro* site-specific integration of bacteriophage DNA catalyzed by a recombinase of the

resolvase/invertase family. *Proc Natl Acad Sci USA* 1998;**95**:5505–10. https://doi.org/10.1073/pnas.95.10.5505
34. Yang L, Nielsen AAK, Fernandez-Rodriguez J *et al.* Permanent genetic memory with >1-byte capacity. *Nat Methods* 2014;**11**:1261–6.
35. Merrick CA, Wardrope C, Paget JE *et al.* Rapid optimization of engineered metabolic pathways with serine integrase recombinational assembly (SIRA). *Methods Enzymol* 2016;**575**:285–317.
36. Abioye J, Lawson-Williams M, Lecanda A *et al.* High fidelity one-pot DNA assembly using orthogonal serine integrases. *Biotechnol J* 2023;**18**:e2200411. https://doi.org/10.1002/biot.202200411
37. Grindley NDF, Whiteson KL, Rice PA. Mechanisms of site-specific recombination. *Annu Rev Biochem* 2006;**75**:567–605. https://doi.org/10.1146/annurev.biochem.73.011303.073908
38. Van Duyne GD, Rutherford K. Large serine recombinase domain structure and attachment site binding. *Crit Rev Biochem Mol Biol* 2013;**48**:476–91. https://doi.org/10.3109/10409238.2013.831807
39. Gupta K, Sharp R, Yuan JB *et al.* Coiled-coil interactions mediate serine integrase directionality. *Nucleic Acids Res* 2017;**45**:7339–53. https://doi.org/10.1093/nar/gkx474
40. Rutherford K, Yuan P, Perry K *et al.* Attachment site recognition and regulation of directionality by the serine integrases. *Nucleic Acids Res* 2013;**41**:8341–56. https://doi.org/10.1093/nar/gkt580
41. Sun YE, Aspinall L, Joseph AP *et al.* Structural basis of DNA recombination catalysis and regulation by φC31 integrase. bioRxiv, https://doi.org/10.1101/2025.05.02.651858, 2 May 2025, preprint: not peer reviewed.
42. Chen Y-W, Su B-Y, Van Duyne GD *et al.* The influence of coiled-coil motif of serine recombinase toward the directionality regulation. *Biophys J* 2023;**122**:4656–69. https://doi.org/10.1016/j.bpj.2023.11.009
43. Arndt D, Marcu A, Liang Y *et al.* PHAST, PHASTER and PHASTEST: tools for finding prophage in bacterial genomes. *Brief Bioinform* 2019;**20**:1560–7. https://doi.org/10.1093/bib/bbx121
44. Lin Z, Akin H, Rao R *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;**369**:1123–30. https://doi.org/10.1126/science.ade2574
45. Mirdita M, Schütze K, Moriwaki Y *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* 2022;**19**:679–82. https://doi.org/10.1038/s41592-022-01488-1
46. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. *Proteins Funct Bioinforma* 2004;**57**:702–10. https://doi.org/10.1002/prot.20264
47. Wang HC, Ho C-H, Hsu KC *et al.* DNA mimic proteins: functions, structures, and bioinformatic analysis. *Biochemistry* 2014;**53**:2865–74. https://doi.org/10.1021/bi5002689
48. Schrödinger, LLC. The PyMOL Molecular Graphics System, version 1.8. https://www.pymol.org/
49. MacDonald AI, Baksh A, Holland A *et al.* Variable orthogonality of serine integrase interactions within the φC31 family. *Sci Rep* 2024;**14**:26280. https://doi.org/10.1038/s41598-024-77570-9
50. Olorunniji FJ, Lawson-Williams M, McPherson AL *et al.* Control of φC31 integrase-mediated site-specific recombination by protein trans-splicing. *Nucleic Acids Res* 2019;**47**:11452–60. https://doi.org/10.1093/nar/gkz936
51. Zhao J, Pokhilko A, Ebenhöh O *et al.* A single-input binary counting module based on serine integrase site-specific recombination. *Nucleic Acids Res* 2019;**47**:4896–909. https://doi.org/10.1093/nar/gkz245
52. Olorunniji FJ, McPherson AL, Rosser SJ *et al.* Control of serine integrase recombination directionality by fusion with the directionality factor. *Nucleic Acids Res* 2017;**45**:8635–45. https://doi.org/10.1093/nar/gkx567
53. Summers DK, Sherratt DJ. Resolution of ColE1 dimers requires a DNA sequence implicated in the three-dimensional organization of the cer site. *EMBO J* 1988;**7**:851–8. https://doi.org/10.1002/j.1460-2075.1988.tb02884.x
54. Rice P, Kiyoshi M. Structure of the bacteriophage Mu transposase core: a common structural motif for DNA transposition and retroviral integration. *Cell* 1995;**82**:209–20. https://doi.org/10.1016/0092-8674(95)90308-9
55. Abramson J, Adler J, Dunger J *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold3. *Nature* 2024;**630**:493–500. https://doi.org/10.1038/s41586-024-07487-w
56. Douillard FP, Portinha IM, Derman Y *et al.* A novel prophage-like insertion element within *yabG* triggers early entry into sporulation in *Clostridium botulinum*. *Viruses* 2023;**15**:2431. https://doi.org/10.3390/v15122431
57. Misiura A, Pigli YZ, Boyle-Vavra S *et al.* Roles of two large serine recombinases in mobilizing the methicillin-resistance cassette SCC*mec*. *Mol Microbiol* 2013;**88**:1218–29. https://doi.org/10.1111/mmi.12253
58. Franco RAL, Brenner G, Zocca VFB *et al.* Signal amplification for cell-free biosensors, an analog-to-digital converter. *ACS Synth Biol* 2023;**12**:2819–26. https://doi.org/10.1021/acssynbio.3c00227
59. Guiziou S, Maranas CJ, Chu JC *et al.* An integrase toolbox to record gene-expression during plant development. *Nat Commun* 2023;**14**:1844. https://doi.org/10.1038/s41467-023-37607-5
60. Bonnet J, Subsoontorn P, Endy D. Rewritable digital data storage in live cells via engineered control of recombination directionality. *Proc Natl Acad Sci USA* 2012;**109**:8884–9. https://doi.org/10.1073/pnas.1202344109

Received: February 25, 2025. **Revised:** June 26, 2025. **Editorial Decision:** June 27, 2025. **Accepted:** July 18, 2025