MANUSCRIPT TITLE

S-SELeCT: A Human-Evolved Serine Integrase System for Efficient Large-Cargo Genome Integration

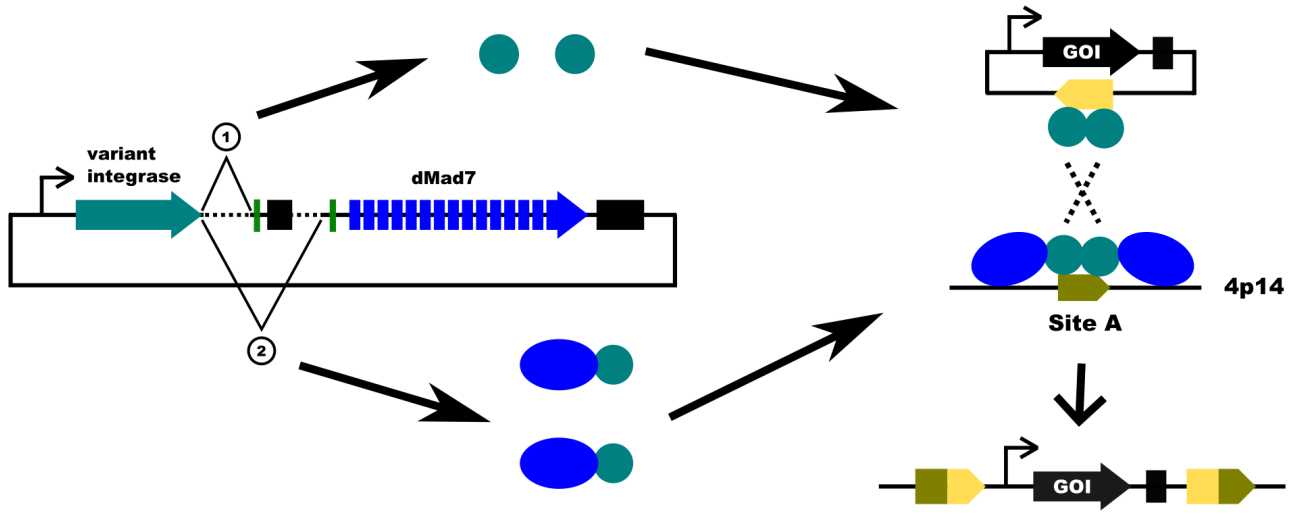
AUTHORS

Alfonso P. Farruggio[1], Lin Jiang[1], Karen Duong[1], Cynthia Nguyen[1], Razan Kaddoura[1] and Ruby Tsai[1,*]

[1] Research and Development, Applied StemCell Inc., Milpitas, CA, 95035, USA

* To whom correspondence should be addressed. Email: Ruby.Tsai@appliedstemcell.com

GRAPHICAL ABSTRACT



ABSTRACT

As a consequence of their sizes, many loss-of-function genetic mutations fall within large genes. A major gene-therapy tool that could be used to solve large swaths of the genetic diseases that result from these inherited mutations is large-fragment knock-in. I.e. instead of attempting to create separate treatments for each and every location that these mutations occur in, large groups of patients could be aided via a single safe-harbor integration of the full-length coding sequence. Towards this goal, we have created a set of early stage gene-editing enzymes that can help mediate large cargo integration at a safe harbor locus in human cells. When expressed in stable lines, our S-SELeCT (Site-Specific Large Cargo Targeting) integrase fusions can facilitate integration of a 10 kb plasmid at frequencies up to 32%, and when delivered transiently via plasmid transfection, we were able to achieve up to 13% knock-in. These are the first serine integrase enzymes that have been evolved fully in human cells, and the first to recognize an endogenous symmetric non-pseudosite – the first true human serine integrase attachment site.

INTRODUCTION

Since the initial reports of eukaryotic gene editing in 2012-2013, the CRISPR/Cas9 system has completely changed the ease and speed at which cell line and animal models are developed (1), and CRISPR-based therapeutics have now been part of hundreds of clinical trials. However, while CRISPR performs exceptionally well at cutting DNA and knocking out genes, Cas9-mediated transgene knock-in often suffers from low efficiencies, due to a reliance upon the endogenous host cell's repair machinery for the introduction of exogenous DNA. In addition, these DNA breaks can lead to undesired outcomes like off-target indels, large deletions, translocations and inversions (2–5). Therefore, there is a need for technologies that can drive site-specific insertion of large transgenes for therapeutic gene editing.

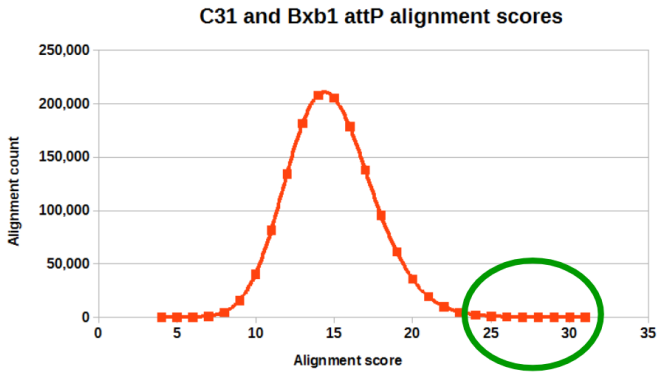
Serine integrases are excellent candidates for this because they facilitate both the cleavage and re-ligation reactions that are needed for transgene knock-in (6), thereby eliminating reliance upon host DNA-repair machinery. These enzymes mediate integration via recombination of two distinct 40-50 bp sequences called the attP and attB sites, which are named after on the locations that they naturally occur in: the phage and bacterial attachment sites, respectively (6). In the absence of the Recombination Directionality Factor (RDF) protein, this reaction is unidirectional, i.e. dimers bound to the resultant attL and attR sites (left and right attachment sites) do not form tetramers, thus the reverse reaction cannot proceed (6, 7). Att-sites have dyad symmetry, i.e. they are semi-identical inverted repeats that bind to homodimers of the integrase protein (6, 7).

While serine integrases have evolved to function in prokaryotes, a variety have been observed to function well in mammalian cells (8, 9). Some of these enzymes can mediate integration into endogenous mammalian sequences called "pseudosites", however most require that their wildtype attP or attB sequence be pre-introduced into the genome, especially for optimal integration efficiency (9, 10). Unlike recombination between their natural sequences, pseudosite integration often involves host-cell DNA-repair activities, which can result in undesired outcomes like small insertions and deletions (indels) near the insert junctions, but also larger deletions, translocations and inversions (10, 11). Pseudosites are often not symmetrical, and usually bare only low-levels of identity to wildtype attP and attB sites (10–12).

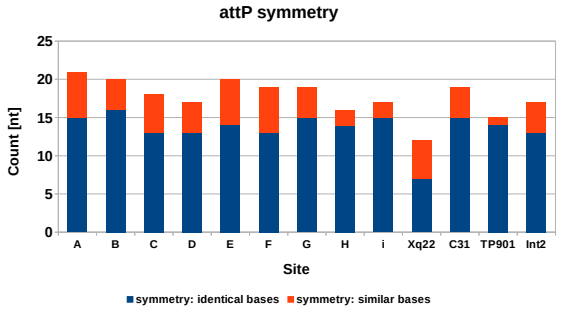
To advance large-insert gene-therapy, we have derived S-SELeCT integrase enzymes, which are variants of phiC31 integrase (C31-int) created via directed evolution. With assistance from fused dMad7, a deactivated Cas nuclease (13, 14), they are capable of mediating plasmid knock-in to a novel human safe-harbor locus on chromosome 4 (4p14) that we refer to as "site A". While many C31-int attP pseudosites are present in the human genome, integration frequencies at these sites is far lower than that for introduced wildtype attP (9, 15, 16). Thus, to more closely mirror the wildtype reaction, which facilitates high-efficiency no-damage recombination (15, 16), we have opted to evolve C31-int variants that target a symmetric non-pseudosite. I.e. site A is not a known pseudosite, and it has dyad symmetry that falls within the range of standard wildtype serine integrase attP sequences (Fig. 1).

Figure 1: Site candidate search: Alignment scores and symmetry

A



B



C

Name	Sequence		
Site A	gagcatccccaacgaagaggactt	TC	aggtctccttatttgggaaaaccc
WT attP	gtagtgccccaactggggtaacct	TT	gagttctctcagttgggggcgtag
Site A	gagcatccccaacgaagaggactt	TC	aggtctccttatttgggaaaaccc
RC WT attP	ctacgcccccaactgagagaactc	AA	aggttaccccagttggggcactac

(**A**) attP-like sites' alignment against C31-int attP. Approximately 75 sites, exhibiting the highest alignment scores, were subjected to further analysis (green circle). (**B**) Site symmetry counts (left vs right half-site alignment results) for candidate-sites A-i, pseudosite Xq22 and three wildtype serine integrases. Blue shows identical bases, and red shows similar bases (i.e., A is similar to G, and C is similar to T). (**C**) Sequence alignments of Site A and both orientations of WT C31-int attP (RC: reverse complement). In the Site A sequence, yellow bases match C31-int attP, while green bases do not.

To our knowledge, while there have been several previous attempts to alter the specificity of C31-int, they have all only yielded variants that performed well in bacteria. I.e. when the top mutants were tested in mammalian cells, their activity was majorly reduced (17) (also personal communications). Thus, to avoid these problems, we have performed all S-SELeCT variant screens in HEK293 cells. Screening in mammalian cells is preferable to bacteria or other organisms since activity in human cells is our primary interest, and it is not unusual for wildtype or variant prokaryotic DNA-modification enzymes to perform significantly worse in mammalian cells (8, 17, 18). Despite having the same sequence, a protein expressed in bacteria and mammalian cells can have significantly distinct activity because of different host-cell organelles (protein localization), post-translational modification systems (phosphorylation, glycosylation, etc.), and even protein-folding chaperones (19–21). Utilization of a mammalian-cell-based evolution system ensures that we are evolving toward a final protein that will function in mammalian cells.

MATERIAL AND METHODS

Selection of Site A

We started the search for potential directed evolution targets by identifying all 50 nucleotide human sequences that have a non-palindromic di-nucleotide core and a dyad symmetry consistent with characterized serine integrases. Next, these sequences were aligned against the forward and reverse-complement phiC31 and Bxb1 attP sites. Candidates with a score in the top 75 were then manually screened for GC content between 20-60%, uniqueness, overlap with known genes, proximity to DNAseI hypersensitivity sites (HSS), and information on the flanking genes. A set of 17 active-integrase attP sites was used to determine the GC content boundaries: phiC31, phiTB1, R4, Bxb1, TP901, TG1, Int2, Int3, Int4, Int5, Int7, Int8, Int9, Int10, Int11, Int12, and Int13 (22–28). BLAT was used to check for uniqueness (29). To investigate gene overlap, DNaseI HSS distance, and neighboring gene information, we used the UCSC Genome Browser (GRCh38/hg38) (30). To assess if the flanking genes were known to be potential direct oncogenes, we performed PubMed searches. For prediction of expression, we consulted GTEx RNA-Seq expression data of the 5′ and 3′ genetic neighbors (31). The DNAseI HSS data that we used was from the ENCODE project (32). To further test the openness of chromatin at the candidate sites, we confirmed that Cas9 was able to mediate knock-in of a 4 kb insert (CMV-puro-TK-GFP) at each locus (pool-based analysis, data not shown).

General plasmid cloning

All enzymes used for cloning PCRs, restriction-digests, ligations and Golden Gate Assembly (GGA) were purchased from NEB. We primarily used the GeneArt GeneOptimizer® tool for codon optimization (33), and in a few instances the Codon Optimization OnLine (COOL) algorithm was used (34). DNA segments were ordered from GeneArt when they contained codon optimizations produced by their tool. In all other cases, dsDNA segments and oligos used for cloning were purchased from IDT. PCR was performed with Q5 Hot Start High-Fidelity 2X Master Mix (NEB M0494L). DNA segments were gel- or column-purified using Zymo Research DNA Clean & Concentrator kits and reagents (e.g. D4004), then quantified using the Qubit 1X dsDNA Broad Range kit (Thermo Q33266) or UV-absorption (BioTek Synergy HTX). For elution from columns, Qiagen Buffer EB (Qiagen 19086) was used (heated to 95°C for segments larger than 2 kb).

When performing cloning via ligation of purified restriction-enzyme-digested segments, we combined 5 femtomoles (fmol) of plasmid backbone with 15 fmol of each insert in a 5 uL reaction. These mixes contained 0.5 uL 10X T4 DNA Ligase Buffer, 0.5 uL 10 mM ATP (NEB P0756S) and 0.5 uL T4 DNA Ligase (400 U/uL; NEB M0202L). When only one insert was present, we thermocycled the ligation reaction as follows: 15 cycles of 10°C for 30 sec, 30°C for 30 sec; 1 cycle of 65°C for 12 min, 4°C hold. For higher insert counts, we increased the cycling: 30 cycles for two inserts, 45 cycles for three inserts, etc.

For GGA, we primarily used the NEB BsaI-HFv2 and BsmBI-v2 Golden Gate enzyme mixes (NEB E1601L and E1602L, respectively). For 1-3 insert GGA, we combined 5 fmol of plasmid backbone with 15 fmol of each insert in a 5 uL reaction. The mixes contained 0.5 uL 10X T4 DNA Ligase Buffer and 0.25 uL of the BsaI or BsmBI GGA enzyme mix. When only one insert was present, we thermocycled BsaI GGA reactions as follows: 15 cycles of 37°C for 30 sec, 16°C for 30 sec; 1 cycle of 65°C for 12 min, 4°C hold.

BsmBI GGA single-insert reactions were cycled between 42°C and 16°C, i.e.: 15 cycles of 42°C for 30 sec, 16°C for 30 sec; 1 cycle of 65°C for 12 min, 4°C hold. For higher insert counts, we increased the cycling: 30 cycles for two inserts, 45 cycles for three inserts.

To perform GGA involving 4 or more inserts, we increased the volumes and DNA amounts as needed, and switched to an equal molecular ratio (i.e. all backbone and inserts present at the same copy number). Final DNA concentrations ranged between 1-6 nM, and cycling conditions were increased up to the manufacturer's recommended maximum (30 cycles of 37°C/42°C for 5 min, 16°C for 5 min); reactions were then heat-killed at 65°C for 12 minutes and held at 4°C. When a different type IIS restriction enzyme was needed, we used the NEBridge Ligase Master Mix (NEB M1100L) and followed the manufacturer's guidance.

After ligation or GGA, we combined 1 uL of the reaction with 20 uL of chemically-competent 10-beta E. coli cells (NEB C3019H) in a 300 uL thin-walled PCR tube, and then incubated the solution at 4°C for 30 minutes. Transformation was performed by heat-shocking the cells for 30 seconds in a thermocycler heat-block set to 42°C. The cells were immediately chilled at 4°C for 5 minutes, then transferred into 400 uL of NEB 10-beta/Stable Outgrowth Medium (NEB B9035S). The cells were grown for 1 hour at 37°C with 250-300 RPM shaking, then spread onto LB Agar plates with the needed antibiotic, and incubated for 12-18 hours at 37°C.

Plasmid DNA amplification, purification and sequencing

For all scales of plasmid amplification, we started by seeding a single colony into 2 mL of Supreme Media (SM) or Terrific Broth (TB; Teknova T7660) supplemented with the needed antibiotic: 50 ug/mL kanamycin (Fisher AAJ67354AD), 100 ug/mL carbenicillin (Teknova C2199), or 34 ug/mL chloramphenicol (Fisher AAJ67273AB). The growth was incubated at 37°C for 16-24 hours with shaking at 250-300 RPM. 500 mL SM was prepared as follows: 472.51 mL NERL High Purity Water (Fisher 23-249-590), 4.13 mL 50% (w/w) Glucose (Teknova G9005), 18.11 g Bacto CD Supreme (Fisher A4973701), 5 mL 99.5% Glycerol (Fisher AC327255000), 0.25 g L-Leucine (MP Biomedicals 02102158.3). The medium was then filter-sterilized (Corning 431097) and stored at 4°C.

Mini-prep scale plasmid purification was performed using the QIAprep Spin Miniprep Kit (Qiagen 27106) and protocol, however with doubled volumes. I.e. 2 mL cell growths were pelleted, resuspended in 500 uL Buffer P1, lysed with 500 uL Buffer P2 for 3-5 minutes, then neutralized with 700 uL Buffer N3. The lysate was centrifuged for 10 minutes at 20,000 RCF, then decanted into a separate 2 mL tube. Clarified lysate was passed through a single QIAprep column via centrifugation (18,000 RCF for 1 min) or vacuum pressure (300 mbar), and the column was then washed with 0.8-1 mL Buffer PE. Drying was performed via centrifugation (18,000 RCF for 1 min) and incubation at 50°C for 5 minutes, both in fresh collection tubes. For elution, we used 30-60 uL of Buffer EB.

For midi-prep scale purification, we seeded 300 uL of the 2 mL growth into 140 mL of antibiotic-supplemented SM or TB in a sterile 250 mL PETG baffled Erlenmeyer flask (Thermo 10-530). If using carbenicillin, before seeding we first washed the cells twice with fresh growth medium (because the resistance enzyme is secreted/present outside the cells). Flasks were then incubated at 37°C for 16-24 hours with 250-300 RPM shaking. Growths were aliquoted into pre-weighed 50 mL tubes, then centrifuged at 3000 RCF for 15 minutes at 4°C. After discarding supernatants, tubes were weighed

again to determine pellet masses. We next used the NucleoBond Xtra Midi EF and Finisher Midi kits (Machery-Nagel 740420.50 and 740439.50) according to the manufacturer's protocol, however with modified wash and elution steps. 10.67 mL of each buffer (RES-EF, LYS-EF, NEU-EF) was used per 1 gram of cell pellet. Both wash steps were repeated an additional time (i.e. two 35 mL ENDO-EF and two 15 mL WASH-EF washes in total). Qiagen Buffer EB heated to 50°C was used to elute from finisher columns, and concentrations were determined via UV-vis absorption (BioTek Synergy HTX). Frequently used plasmids were aliquoted and stored at -20°C in a non-cycling freezer (i.e. not frost-free).

All plasmids were sequenced with Oxford Nanopore devices, and custom scripts were used to analyze the results (alignment to expected sequence, identification of mismatches, coverage level, etc.). Examples of products used include the Rapid Barcoding Kit 24 V14 (Oxford Nanopore Tech. SQK-RBK114.24), R10.4.1 flongle flow cells (FLO-FLG114), and R10.4.1 MinION flow cells (FLO-MIN114).

General library cloning

GGA was used to create all libraries. We most often created a 30 uL mix with 90 or 180 fmol of each segment present, i.e. 3 or 6 nM final concentrations. The solutions contained 3 uL 10X T4 DNA Ligase Buffer and 1.5 uL of the BsaI-HFv2 or BsmBI-v2 Golden Gate enzyme mix. We used a thermocycler to incubate BsaI GGA reactions as follows: 30 cycles of 37°C for 5 min, 16°C for 5 min; 1 cycle of 65°C for 12 min, 4°C hold. BsmBI GGA reactions were incubated at 42°C instead of 37°C: 30 cycles of 42°C for 5 min, 16°C for 5 min; 1 cycle of 65°C for 12 min, 4°C hold. After GGA, the reactions were purified and concentrated via column-purification using the Zymo DNA Clean & Concentrator kit (Zymo D4004). We followed the manufacturer's protocol, however to elute the DNA we used 8 uL of Buffer EB heated to 95°C (Qiagen 19086).

Next, 1-3 uL of the purified GGA was combined with 25 uL of 10-beta E. coli (NEB C3020K) in a 0.1 cm gap cuvette (Bio-Rad 1652083), and the mix was electroporated using a BTX Electro Cell Manipulator 600 at 1.7 kV with 186 Ohm resistance. If successful (i.e. no spark, 7-8 msec discharge time), the cells were then immediately transferred into 1 mL of NEB 10-beta/Stable Outgrowth Medium (NEB B9035S), and incubated for 1 hour at 37°C with 250-300 RPM shaking. Serial dilutions of the outgrowth were plated for library size or coverage estimates. The remainder of the growth (typically 925-950 uL; specific amount measured) was then seeded into 140 mL of antibiotic-supplemented SM or TB in a sterile 250 mL PETG baffled Erlenmeyer flask (Thermo 10-530). The cells were incubated at 37°C for 16-24 hours with shaking at 250-300 RPM. If the library size/coverage was adequate, we then purified the plasmid pool using our described midi-prep protocol.

Intronization

To intronize sequences, we first performed codon optimization using the GeneArt GeneOptimizer tool, with desired restriction sites and strong potential exon splice junctions excluded (CAG/GT, AAG/GT) (35). Next, we created a custom algorithm to introduce exon splice junctions at desired intervals via silent codon changes. We weighted potential splice junctions with guidance from the Ma et al. sequence logos (i.e. high scoring for 5' C/A-A-G, weak preference for 3' G-T-T) (35). Our exons mostly fall in the 200-300 nt range. Finally, we inserted introns at the introduced exon splice junctions. Introns were obtained from endogenous human genes including: ACTB, GAPDH, PGK1, PPIA, RPL13A, RPLP0, TTN, and YWHAZ. They were either naturally less than 201 nt in size, or we truncated them to

meet this requirement (by fusing together the first and last 100 nt of the intron). In certain instances, we modified introns to start with the dominant "GTAAG" motif identified by Ma et al. (35). We have annotated all exons and introns in our deposited sequences.

Deactivated Mad7

Since no precedence for the use of deactivated Mad7 (dMad7) in mammalian cells existed at the time we performed this work, we evaluated the three mutations described by Price et al.: D877A, E962A, and D1213A (14). The Mad7 sequence from Wierson et al. was modified to create four dMad7 candidates: d1a (D877A), d1b (E962A), d2 (D877A, E962A) and d3 (D877A, E962A, D1213A) (13). A custom EGxxGFP reporter assay (e.g. see Mashiko et al.) (36) was used to compare activity of the candidates to active Mad7 (data not shown). d2 and d3 performed the best, so they were next tested in triplicate, where d2 was observed to be optimal (Fig. S1). We thus adopted d2-Mad7 for use in all integrase-dMad7 fusions.

Inversion-assay plasmid assembly

Three-exon GFP recombination-reporter plasmids were cloned using Golden Gate Assembly (GGA) of synthesized DNA segments and PCR products (37). They consist of two expression cassettes oriented in a divergent manner, flanked by UMS terminator sequences (38). The first is driven by a CMV promoter, contains the three-exon GFP segment, and is terminated by the Bovine Growth Hormone (BGH) 3′ UTR/polyA. The two introns in GFP are derived from Li et al. (39) and the second human Hemoglobin Subunit Beta intron (HBB; NCBI Gene 3043). The second expression cassette consists of the human EF1-alpha promoter, a 3′ intein-3′ mCherry2 protein fusion, WPRE (Woodchuck Hepatitis Virus Posttranscriptional Regulatory Element), and an SV40 3′ UTR/polyA (40–42). We have deposited the sequences of a representative plasmid sequence (pKSP3-dAf-8) and the corresponding recombination product (pKSP3-dAf-8-flipped) at Zenodo.

Mini-chromosome assay reporter plasmid cloning

Mini-chromosome assay reporter plasmids were cloned using GGA of synthesized DNA segments and PCR products. They contain a split-GFP integration reporter, a polycistronic trEBNA-1 puroR expression cassette, and the EBV oriP replication origin (43). The expression cassette is driven by the human EF1-alpha promoter, contains a WPRE, and is terminated by the BGH 3′UTR/polyA. trEBNA-1 is a truncated version of the EBNA1 protein (43), and it is separated from the puromycin-resistance protein via P2A-mediated ribosomal skipping (44). The integration reporters consist of an attP-3′ intron-3′GFP-BGH 3′UTR/polyA segment that is flanked by genomic Site A sequence. The attP is either from WT phiC31 or the human Site A sequence. The 3′ intron is derived from the synthetic intron acceptor described by Li et al. (39), and 3′GFP is a C-terminal segment from EGFP4m (mut4EGFP) (45). In all plasmids, upstream and downstream Site A genomic sequences have been cloned 5′ and 3′ of the reporter, respectively. In the MAX and SAR1 vectors, 5 kb and 2.5 kb genomic sequence has been cloned in each region (10 and 5 kb total in each plasmid), respectively. A 1.2 kb chicken heat-shock 4 (cHS4) insulator (46) separates the reporter and trEBNA-1 puroR expression cassette. Sequences of the following reporter plasmids have been deposited at Zenodo: pKEO-SAR1-WT-attP, pKEO-SAR1, pKEO-SAR-MAX-WT-attP, pKEO-SAR-MAX.

Split-GFP assay donor plasmid assembly

Donor plasmids used in both chromosomal and mini-chromosome split-GFP assays were cloned using GGA of synthesized DNA segments and PCR products. They contain a single CMV-driven polycistronic cassette that is terminated by an SV40 3'UTR/polyA. The first protein expressed is a 3' intein-3' mCherry2 protein fusion (40, 41). The next protein is separated via a P2A peptide, and consists of an N-terminal EGFP4m exon without a stop codon. An intron splice donor derived from Li et al. (39) borders the EGFP4m exon, and the following elements are downstream of it: (i) either the wildtype or fully mutated Site A attB, (ii) an intron splice acceptor derived from HBB, (iii) the sLTSV(-) hammerhead ribozyme (HHR) (47, 48). Both plasmid sequences (pDiG-SP3-WT-attB, pDiG-SP3-Af8) have been deposited at Zenodo.

Native Site A donor plasmid cloning and barcoding

The native Site A donor precursor-plasmid was cloned using GGA of synthesized DNA segments and PCR products. It contains a single CAG-driven heavily-intronized polycistronic cassette with a WPRE and BGH 3'UTR/polyA terminator. Three proteins are co-expressed and separated via P2A peptides: EGFP4m, a 3' intein-3' mCherry2 fusion, and a puromycin-resistance enzyme (40, 41, 45). The cassette is flanked by a single full-length cHS4 insulator on each side (46). Upstream of the first insulator, the plasmid also contains a SapI Golden Gate cloning platform, and the fully mutated Site A attB sequence. We have deposited sequences of the precursor (pFAD-HBH) and barcoded (degenerate) donor (pFAD-HBH-K18) at Zenodo.

To create a barcoded plasmid pool, we synthesized a degenerate oligonucleotide mix, extended it with Q5 polymerase to create dsDNA, then performed Golden Gate Assembly with SapI. The degenerate oligonucleotide mix was synthesized as an IDT Ultramer, and had the following sequence: 5'-gaaaaggagaagggaaagagggaggttgttGCTCTTCtAGAa KKKKKKaagaaKKKKKKaggaKKKKKKaGAGtGAA GAGCttgttgTGAGATCTTTTGATCTCAcaacaagctcttcactc-3'. To produce double-stranded DNA (dsDNA), we first created the following 100 uL mix: 30 uL UltraPure H2O (Thermo), 50 uL Q5 HotStart HF 2X Master Mix (NEB M0494L), 20 uL 10 uM oligo. There is no second primer because the oligo forms a 3' hairpin (GAGtGAAGAGCttgttgTGAGATC TTTT GATCTCAcaacaagctcttcactc). Next, the mix was thermocycled as follows: 98°C for 30 sec; 10 cycles of 98°C for 7 sec, 65°C for 10 sec (-0.5°C/cycle), 72°C for 4 sec; 25 cycles of 98°C for 7 sec, 60°C for 10 sec, 72°C for 4 sec; 72°C for 2 min, 4°C hold. The ~94 basepair (bp) product and 10.1 kbp SapI-digested plasmid backbone were then gel-purified, and used in the following 30 uL Golden Gate Assembly (GGA) mix: 11.36 uL UltraPure H2O, 10 uL 3X NEBridge Ligase Master Mix (NEB M1100S), 3.64 uL backbone (49.4 fmol / uL), 1 uL dsDNA barcode pool (180 fmol / uL), 4 uL SapI (NEB R0569S). The mix was thermocycled as follows: 37°C for 5 min; 30 cycles of 16°C for 1 min, 37°C for 1 min; 4°C hold. Next, we column-purified the mix according to the manufacturer's protocol (Zymo D4004), however for elution we used 8 uL of Qiagen Buffer EB heated to 95°C. We electroporated 3 uL of the purified GGA into 25 uL of electrocompetent cells (NEB C3020K) using 0.1 cm gap cuvettes (Bio-Rad 1652083) and a BTX Electro Cell Manipulator 600 (1.7 kV, 186 Ohm), then followed the remainder of our standard plasmid library production protocol. Based on colony-forming-unit (CFU) counts of the post-electroporation 1 hour outgrowths, our barcoded-plasmid pool diversities ranged from 232-259 e6.

Solo-integrase stable-expression plasmid assembly

The precursor plasmid used for non-fused expression of WT C31-int and variants was cloned using GGA of synthesized DNA segments and PCR products. It contains two divergent promoterless expression cassettes centered on a Bxb1 attB site. The first incomplete cassette contains the coding sequence for a blasticidin-resistance protein, and it is terminated by an SV40 3'UTR/polyA. The second partial-cassette has a BsaI GGA platform flanked by WT C31-int coding sequence with a C-terminal mouse TCOF NLS. Downstream of this, the following elements are present: P2A peptide, 5' mCherry2-5' intein fusion protein, BGH 3'UTR/polyA. In addition, two UMS terminators are present in the plasmid, one on each flank of the divergent partial-cassette segment. To clone complete integrase sequences or variant libraries, BsaI GGA was performed using our standard conditions (for plasmids and libraries, respectively). We have deposited the precursor (pHKi-CL-357-60mT), WT C31-int (pHKi-WTC31-60mT), V7 (pHKi-V7-60mT), and V12 (pHKi-V12-60mT) plasmid sequences at Zenodo.

Transient solo-integrase expression plasmid cloning

For transient-expression of variants in a non-fused form, a precursor plasmid was cloned using GGA involving synthesized DNA segments and PCR products. Most plasmid backbone segments were obtained from pCMV-SPORT-βgal (Thermo 10586014) via PCR. The precursor vector has a single expression cassette driven by a CMV promoter, and it contains the following elements: BsaI GGA platform with flanking WT C31-int coding sequence, C-terminal human TCOF NLS, 3' small SV40T intron (from pCMV-SPORT-βgal), SV40 3'UTR/polyA. To clone variants, a central segment encoding the desired mutations was generated via PCR or gene synthesis, then it was cloned into the precursor plasmid backbone (BsaI digested and gel-purified) using our standard BsaI GGA conditions. We have deposited the precursor vector (pDCMV-CL-357-hT60) and solo-V7 expression (pDCMV-V7-hT60) plasmid sequences at Zenodo.

Fused-integrase stable-expression plasmid assembly

The precursor plasmid used for expression of WT C31-int and variants fused to dMad7 was cloned using GGA of synthesized DNA segments and PCR products. It contains two divergent promoterless expression cassettes centered on a Bxb1 attB site. The first incomplete cassette contains the coding sequence for a blasticidin-resistance protein, and it is terminated by a BGH 3'UTR/polyA. The second partial-cassette has a BsaI GGA platform flanked by intronized WT C31-int coding sequence. Downstream of this is the NL1 or NL2 NLS-linker that is used to fuse integrase to intronized dMad7. The NL1 sequence consists of the mouse TCOF NLS and a Gly_4-Ser-Gly-Gly linker, whereas NL2 is a fusion of the human Nucleoplasmin NLS and an Gly_4-Ser-Gly_4-Ser-Gly-Gly-Gly linker. Two SV40 NLS tags are present at the C-terminus of dMad7, and the following elements are located immediately downstream: T2A peptide, intronized 5' mCherry2-5' intein fusion protein, WPRE, SV40 3'UTR/polyA. In addition, two UMS terminators are present in the plasmid, one on each flank of the divergent partial-cassette segment. To clone complete integrase sequences, BsaI GGA was performed using our standard conditions. We have deposited the NL1-precursor (pLKi-CL357-NL1-dM7), NL2-precursor (pLKi-CL357-NL2-dM7), and V30-NL2 (pLKi-V30-NL2-dM7) plasmid sequences at Zenodo.

Alternative-splicing WT C31-int-GFP/mCherry plasmid cloning

The plasmids used to test alternative-splicing introns were assembled using GGA of synthesized DNA segments and PCR products. They consist of two divergent promoterless expression cassettes centered on a Bxb1 attB site. The first incomplete cassette contains the coding sequence for a blasticidin-resistance protein, and it is terminated by a BGH 3'UTR/polyA. The second partial-cassettes contain the following elements: intronized WT C31-int, splice donor, splice-1 acceptor, P2A peptide, EGFP4m, splice-2 acceptor, P2A peptide, mCherry2, WPRE, SV40 3'UTR/polyA. In addition, we cloned a version of all plasmids with a BGH 3'UTR/polyA between EGFP4m and the splice-2 acceptor ("with internal polyA", (Fig. S2). The alternative splice donors and acceptors are all derived from chicken TNNT intron 4 as described by Aebischer-Gumy et al. (49). In addition, two UMS terminators are present in the plasmid, one on each flank of the divergent partial-cassette segment. We have deposited sequences of the I4(7Y)-I4sh constructs with (pLK-int-GmC-4) and without (pLK-int-GmC-4-alt) an internal polyA at Zenodo.

Assembly of stable-expression alternative-splicing solo and fused integrase plasmids

Precursors of vectors used to simultaneously stably-express solo and dMad7-fused integrases were cloned via GGA of synthesized DNA segments and PCR products. They contain two divergent promoterless expression cassettes centered on a Bxb1 attB site. The first incomplete cassette contains the coding sequence for a blasticidin-resistance protein, and it is terminated by a BGH 3'UTR/polyA. The second partial-cassettes contain the following elements: a BsaI GGA platform flanked by intronized WT C31-int coding sequence, splice donor, splice-1 acceptor, human TCOF NLS, BGH 3'UTR/polyA, splice-2 acceptor, NL1 or NL2 NLS-linker, intronized dMad7, two SV40 NLS tags, T2A peptide, intronized 5' mCherry2-5' intein fusion protein, WPRE, SV40 3'UTR/polyA. In addition, two UMS terminators are present in the plasmid, one on each flank of the divergent partial-cassette segment. To clone complete integrase sequences or variant libraries, BsaI GGA was performed using our standard conditions (for plasmids and libraries, respectively). We have deposited the NL1-precursor (pLKa4-NL1-dM7-plat), NL2-precursor (pLKa4-NL2-dM7-plat), and V12-NL1 (pLKa4-V12-NL1-dM7) plasmid sequences at Zenodo.

gRNA identification and expression plasmid cloning

Potential guide RNAs were identified using the CRISPOR tool with the YTTV PAM Mad7 setting (50). The precursor of the vectors used to transiently express guide RNAs was assembled via GGA of synthesized DNA and PCR products. There is a single cassette in the plasmid that contains the human U6 promoter, 'TAATTTCTACTCTTGTAGAT' crRNA scaffold (13), and a BsmBI GGA platform. To clone gRNA-expression plasmids, BsmBI GGA was performed. The g6 and g10 expression vectors were made by cloning the 'TGTGGGAGTTCAGTTCAACAACT' and 'AGCCGCTACCCCGACCACATGAA' spacers, respectively. We deposited the precursor (pKU6-M7-plat), g6 (pKU6-M7-SA-6) and g10 (pKU6-M7-SA-10) plasmid sequences at Zenodo.

Transient fused-integrase expression plasmid assembly

Precursors of plasmids used to express integrase-dMad7 fusions were made using GGA of synthesized DNA and PCR products. They consist of three cassettes that express: (i) g6 gRNA, (ii) g10 gRNA, and (iii) an intronized integrase-dMad7 fusion with or without alternative-splicing. The gRNA-expression cassettes are driven by the human U6 promoter, and integrase-fusions are expressed using a CAG or

mPGK promoter. In some plasmids, a WPRE and rabbit beta-globin 3'UTR/polyA are used, while in others just a BGH 3'UTR/polyA is present. All utilized introns and alternative splicing sequences match what were used in the respective stable-expression vectors. We have deposited the sequences of precursor plasmids (pLK-CAG-357-NL1-5C, pLK-CAG-357-NL2-5C, pLK-mPGK-357-NL1-5C, pLK-mPGK-357-NL2-5C, pLK-CAG-357-NL1a4-5C, pLK-CAG-357-NL2a4-5C, pLK-mPGK-357-NL1a4-5C-W, pLK-mPGK-357-NL2a4-5C-W), and two alternative-splicing vectors (pLK-CAG-V12-NL1a4-5C, pLK-mPGK-V7-NL2a4-5C-W) at Zenodo.

Vectors used to compare transient expression of WT C31-int fusions to dCas9a (51) and dMad7 (Fig. S3) were created via GGA of synthesized DNA segments and PCR products. They consist of a single CMV-driven cassette with a polycistron, 5' HIV1 TAR (52, 53), 3' WPRE, and that is terminated by a BGH 3'UTR/polyA. The polycistrons encode HIV1-Tat (52, 53), a C31-int-dCas protein, and mCherry 2. The three proteins are separated via 2A skipping peptides (P2A and T2A). C31-int, dCas9a and mCherry2 are heavily intronized in all constructs, whereas dMad7 was only cloned this way in one construct (Fig. S3). We have deposited the sequences of these constructs (pLKCTT-i23-L8-miC-W, pLKC-L8-dM7-TTF, pLKC-L8-i20-dM7-TTF) at Zenodo.

Mutagenesis of C31-int to create variant libraries

Integrase-variant libraries were generated using a variety of methods, including regional error-prone PCR (e.g. Yang et al.) (54), multi-codon saturation, and GGA-shuffling (55) of segments containing mutations from enriched pools. For codon saturation involving oligos, we used NNK degeneracy (56). In oligo pools, in addition to NNK, we sometimes synthesized 20 separate oligos (one for each possible amino acid) to perform saturation. Insertion and deletion saturation involving 1-4 codons was also performed using oligo pools. For insertion saturation, a combination of NNK and separate-oligos were used to vary amino acids that were introduced, but also at the flanking positions. When performing deletions, the flanking positions were saturated using NNK and separate-oligos in the pool. All synthetic DNA used for mutagenesis was purchased from IDT (oligos, Ultramers and oPools), and GGA was always used to fuse mutant segment pools with plasmid backbones.
Mutation summaries for V7, V12, V29, V30, V32, V36, V37, and V40 have been provided in supplementary tables S1-S8. Total mutation counts for these variants are listed in Table S9.

HEK293 cell culture

HEK293 cells were maintained in complete DMEM medium consisting of high-glucose DMEM (Fisher 11-960-077) supplemented with 4.5% fetal bovine serum (FBS; Corning 35-075-CV), 2 mM GlutaMAX (Fisher 35-050-061), 1 mM sodium pyruvate (Fisher 11-360-070), 50 mM HEPES (Fisher 15-630-080), 100 ug/mL Normocin (InvivoGen ant-nr-1), 50 units/mL penicillin, and 50 ug/mL streptomycin (Fisher 15-140-148). CellBIND multi-well plates and flasks were used at all times to improve cell attachment (e.g. Corning 3337, 3335, 3290, 3291, etc.). Cells were cultured at 37°C in a humidified incubator with 5% CO_2 (Fisher Scientific, Model 3530). To improve survival after FACS isolation, FBS was increased to 10% final for 1 week, then returned to 4.5%. Cells that contained the H11 landing-pad were sometimes grown with 200 ug/mL hygromycin (InvivoGen ant-hg-1) present. Libraries were maintained via selection with 20 ug/mL blasticidin (InvivoGen ant-bl-1). Cultures of cell lines with the SAR1 or SARMAX mini-chromosome were additionally supplemented with 1-2 ug/mL puromycin (InvivoGen ant-pr-1). When both hygromycin and puromycin were present, we also added 2.5 mM

oxaloacetic acid (Fisher AA1578909) to improve cell growth and health. Cells were passaged at 70-90% confluency using TrypLE Express (Fisher 12-605-028) with 1:20-1:2 split densities.

Transfection

The day before transfection, cells were seeded with the aim of reaching 60-80% confluency in 18-24 hours. In experiments involving transient expression, plain cell culture medium without added drugs was used (high-glucose DMEM, 4.5% FBS, 2 mM GlutaMAX, 1 mM sodium pyruvate, 50 mM HEPES), and the final volume per well of a CellBIND 24 well-plate (24ww; 2 cm^2) was 0.5 mL. When expression occurred from the H11 landing-pad, cells were seeded in plain cell culture medium with 1 ug/mL doxycycline (Sigma D3072-1ML). Plain HEK293 and H11-landing-pad cells were seeded at densities of 175-225k live cells per 24ww. Cells with the SAR1 or SARMAX mini-chromosome were seeded at densities in the 300-350k range per 24ww. Cell counts were performed with a Countess II (Thermo) in the presence of 0.2% trypan blue live/dead stain.

The poly beta-amino ester transfection reagent Xfect (Takara 631318) was used for all plasmid delivery, at a 0.3 uL per 1 ug ratio. At minimum, 3.33 ug of plasmid DNA was mixed with 1 uL Xfect in a final volume of 55.4 uL. The solution was vortexed well immediately after addition of Xfect, then complexes were allowed to form for 10-15 minutes. Before pipetting, the sample was briefly vortexed, then 33.3 uL (2 ug DNA) was delivered into each 24 ww. These plasmid amounts, cell densities and volumes were proportionally scaled up or down when transfecting at different scales (e.g. everything increased 5X when working in 6 well-plates, which have 10 cm^2 growth area).

Creation of the HEK293 H11 landing-pad line

Our landing-pad cell line was created by knocking-in a Bxb1 integrase landing-pad (LP) at the H11 locus in a HEK293 line that stably expresses a doxycycline-activated transcription factor (57). To introduce the LP, we used the same TALEN-stimulated approach described in our DICE publication (15). A plasmid mix that consisted of 833 ng MR015-H11-L2-TALEN (Addgene 51554), 833 ng MR015-H11-R2-TALEN (Addgene 51555), and 1665 ng of the LP plasmid was complexed with 1 uL Xfect, then 2 ug of it was delivered into a 24 ww. After 72 hours, we expanded the cells 1:20 and grew them under 200 ug/mL hygromycin selection for 2 weeks. Single cells from this pool were sorted via FACS, and outgrowths were genotyped for the desired knock-in. We have deposited the Tet promoter-attP-EF1a promoter LP segment sequence depicted in Fig. S4a at Zenodo (H11_LP_core).

Establishment of solo or fused-integrase variant HEK293 libraries

To create pools of HEK293 cells that express a single solo or fused integrase variant from the H11 locus, we co-transfected a Bxb1-integrase expression vector with a stable-expression plasmid library into our LP cell line. Our primary Bxb1 integrase plasmid contains a codon-optimized sequence with a C-terminus X. laevis nucleoplasmin NLS; the vector (pCMV-opt-Bxb1-XeNLS) sequence has been deposited at Zenodo. For transfection into a well of a 6-well plate (6ww), the DNA mix contained 2.75 ug Bxb1 integrase vector, 8.25 ug variant-integrase plasmid pool, and was brought to a total volume of 179.8 uL with Xfect Buffer. After complexing with 3.3 uL Xfect, 10 ug (166.7 uL) was delivered into the 6 ww. In parallel, to estimate library size or coverage, a promoterless attB-mCherry donor was co-transfected with the Bxb1 integrase into separate wells (seeded the same way). Three days after

transfection, library size or coverage was estimated by measuring the percent of mCherry-positive cells above background (integration efficiency) in our attB-mCherry control (e.g. Fig. S4b). I.e. library diversity = integration efficiency x total cell count. If the library size and coverage (= diversity / size) was adequate, the cells were expanded 1:20 and maintained for two weeks in medium supplemented with 20 ug/mL blasticidin. The cells were fed 1-3 times a week, and split as needed to maintain at least 3X library coverage.

Creation of split-GFP WT-attP reporter line

To introduce split-GFP reporter sequences at both Site A loci, we took a four-step approach where full-cassette knock-in was followed by excision of the promoter and 5′ segment of GFP. First, an attL-CMV-5′ GFP-attR-3′ GFP-BGHpA amplicon from a pool was knocked-in to a Site A locus using SpCas9-stimulated NHEJ. PCR from a plasmid template was used to create the insert pool with the following primers: 5′-g*t*tgtt NNNNNN ggTAGGGATAACAGGGTAAT ggctacgcccccaac tgagagaac*t*c-3′, 5′-c*a*acaa NNNNNN AAGCATGCCTGCTATTGTCTTC CCAATC-3′, where '*' is used to indicate a phosphorothioate linkage. The sequence of the PCR amplicon (EXLP1_amplicon) has been deposited at Zenodo. SpCas9 and the guide RNA were delivered in RNP form, and cleavage was targeted at the following spacer: CAAGAGCATCCCCAACGAAG (AGG PAM). After two weeks of outgrowth, GFP-positive cells were sorted into a pool. Second step: after 1 week of growth, these cells were then transfected with pCS-kRI, which co-expresses the phiC31 RDF and integrase (58). Two weeks after transfection, single GFP-negative cells were sorted, outgrown and screened for a desirable attP-3′ GFP-BGHpA knock-in via genotyping.

Our third and fourth steps were to repeat this process at the second locus with a slightly different insert, and an alternate spacer. The sequence of the PCR amplicon (EXLP2_amplicon) has been deposited at Zenodo, and the primers used to generate it were: 5′-g*t*tgtt NNNNNN ctacgcccccaactgagagaac*t*c-3′, 5′-c*a*acaa NNNNNN AAGCATGCCTG CTATTGTCTTCCCAATC-3′. The spacer used was: GACTTTCAGGTCTCCTTATT (TGG PAM). The sequences of the first and second locus in the clone that we chose have been deposited at Zenodo (SG51d6__first_locus, SG51d6__second_locus).

WT-attP Site A integration efficiency assay

When using transiently-expressed C31-int vectors to measure integration efficiency in our Site A WT-attP split-GFP assay, we used a 1:3 integrase:donor plasmid ratio. I.e. in a 24 ww, Xfect transfection was used to deliver 0.5 ug and 1.5 ug of the integrase and donor vectors, respectively. When solo C31-int was expressed from the H11 locus (e.g. in NLS screens), we transfected 2 ug of the donor plasmid per 24 ww, and 10 ug per 6 ww. For stably-expressed C31-int-dCas fusions (Sp dCas9 or dMad7), gRNA-expression, solo-integrase expression, and donor plasmids were present at 6%, 19% and 75% of the DNA mix by mass, respectively. Unless otherwise indicated, percentages of cells positive for both mCherry and GFP were measured 72 hours after transfection via fluorescent flow-cytometry (BD FACSMelody).

Inversion assay

The three-exon GFP inversion assay was only conducted in cells that expressed a solo or fused integrase from the H11 locus. We transfected 2 ug of the reporter plasmid per 24 ww, and 10 ug per 6

ww. Unless otherwise indicated, percentages of cells positive for both mCherry and GFP were measured 72 hours after transfection via fluorescent flow-cytometry (BD FACSMelody).

Mini-chromosome establishment and assay

To create pools of HEK293 H11 LP cells with a mini-chromosome reporter vector (e.g. SAR1, SARMAX), we first used Xfect to transfect 2 ug of the respective plasmid per 24 ww. Three days after transfection, we then split the cells at a 1:2-1:4 ratio, and cultured them in medium supplemented with 1-2 ug/mL puromycin, 200 ug/mL hygromycin, and 2.5 mM oxaloacetic acid for a minimum of two weeks. After this period, splits were performed at a 1:10 ratio. These supplements were only removed when performing integration efficiency tests.

To conduct the reporter assay in lines that express integrase fusions from the H11 locus, we used the same 6%-19%-75% DNA mix described for the WT-attP Site A efficiency tests. When using alternative-splicing to express both solo and fused integrases from H11, the gRNA-expression and donor plasmids constituted 6% and 94% of the DNA mix delivered per well, respectively. When expressing variant-dMad7 fusions transiently, we tested a variety of plasmid ratios. For two-plasmid experiments, the alternative-splicing integrase-dMad7 to donor mass ratios were as follows: E=1:3, F=1:2, G=1:1, H=2:1, i=3:1. In three-plasmid experiments, we tested the following fused : solo : donor ratios: J= 1:1:1, K=1:1:2, L=2:1:2, M=2:1:3, N=2:1:1. Unless otherwise indicated, percentages of cells positive for both mCherry and GFP were measured 72 hours after transfection via fluorescent flow-cytometry (BD FACSMelody).

Native Site A assay

In cells that express integrase fusions from the H11 locus, to perform the native Site A assay, we again used the 6%-19%-75% DNA mix for gRNA-expression (equal mix of g6 and g10), solo integrase-expression and donor plasmids, respectively. When alternative-splicing was used to express solo and fused integrase from H11, the gRNA-expression and donor plasmids were 6% and 94% of the DNA mix delivered per well, respectively. For transient expression tests, we used the same ratios and labels established for the mini-chromosome assay.

Unless otherwise indicated, cells positive for both mCherry and GFP were sorted 72 hours after transfection. In stable-expression experiments, 61.2 – 90k cells were sorted per 24 ww, and they were then outgrown for a week before being pelleted. In transient-expression tests, 5k cells were sorted per 24 ww, outgrown for a week in normal sort medium, then cultured in 1 ug/mL puromycin-supplemented medium for 1-2 weeks until the well was 80-100% confluent, after which they were pelleted.

Genomic DNA (gDNA) was extracted from frozen cell pellets using the Quick-DNA Microprep kit (Zymo Research D3021) according to the manufacturer's instructions, however we eluted using Qiagen Buffer EB. DNA concentrations were measured using the Qubit 1X dsDNA Broad Range (BR) kit. Next, we performed the primary 1199 bp PCR with primers 5'-CGCCTTCTATCGCCTTCTT*G*A-3' and 5'-CCTCACCTCACATGGCTC*T*G-3', where * indicates phosphorothioate linkages. The solution contained 50 ng gDNA, 1.25 uL 10 uM forward primer, 1.25 uL 10 uM reverse primer, 12.5 uL Q5 HotStart 2X Master Mix, and was brought to 25 uL with UltraPure H2O (Fisher 10-977-015). The thermocycling

conditions were as follows: 1 cycle of 98°C for 1 min; 10 cycles of 98°C for 14 sec, 67°C for 20 sec (decrease 0.5°C/cycle), 72°C for 51 sec; 25 cycles of 98°C for 7 sec, 62°C for 10 sec, 72°C for 51 sec; 1 cycle of 72°C for 5 min. If the desired 1199 bp band was visible in an agarose gel run, we then performed nested 1039 bp PCR with primers 5′-ACTTTTCGGGGAAATGTG*C*G-3′ and 5′-GCATTCTTGCAGCAATCTC*G*G-3′. The solution contained 2 uL 1:100 diluted 1199 bp PCR, 2.5 uL 10 uM forward primer, 2.5 uL 10 uM reverse primer, 25 uL Q5 HotStart 2X Master Mix, and 18 uL UltraPure H2O. The cycling conditions were as follows: 1 cycle of 98°C for 1 min; 10 cycles of 98°C for 14 sec, 66°C for 20 sec (decrease 0.5°C/cycle), 72°C for 42 sec; 25 cycles of 98°C for 7 sec, 61°C for 10 sec, 72°C for 42 sec; 1 cycle of 72°C for 5 min. If the desired 1039 bp band was visible in an agarose gel run, we then proceeded with sequencing.

Barcode sequencing and counting

Oxford Nanopore technology was used to perform all barcode sequencing. For stable-line experiments, where 61.2k-90k cells were sorted, we counted barcodes from three variant tests: V12-NL1a4, V30-NL2 and V37-NL1a4. From each of these tests, there was one biological sample, and three technical replicates (1199 primary PCR → 1039 nested PCR). In total, there were nine 1039-PCR pools, each of which was amplified from an independent 1199-PCR. We used the Plasmidsaurus custom sequencing service, at a 1 gigabase (Gbp) scale for each sample (9 Gbp total).

In V7-NL2a4 transient-expression experiments, each biological sample was derived from a 5k cell sort. For experiments with a standard seeding density (175k cells in 24 ww), we sequenced 3 technical replicates derived from one ratio E biological sample, and 2-4 technical replicates derived from two biological replicates of ratio F (2 tech. reps. from first sort, 4 tech. reps. from second sort). When cells were seeded at a higher cell density (HCD; 200k in 24 ww), we sequenced 3-4 technical replicates derived from two biological replicates of ratio E (4 tech. reps. from first, 3 tech. reps. from second), and 3 technical replicates derived from one biological sample of ratio F. Sequencing was performed both in-house using a MinION device, and via the Plasmidsaurus custom sequencing service at a 111 megabase scale per sample.

To count barcodes from the large sorts (61.2k-90k), we used a custom script that looked for sequences matching the designed K6x3 pattern (KKKKKK-AAGAA-KKKKKK-AGGA-KKKKKK, K=G or T) that were present in at least 2 reads in 2 or more technical replicates. Due to the large amount of data (9 Gbp total), only the following error-correction was performed inside of K6 segments: homopolymer over- and under-runs, C-T miscalls, A-G miscalls. In addition, for a barcode to be included in a technical replicate, a perfect 45 nt recombination junction sequence (5′-gagcatcccaagaggactt TC AGGTCTCCTTATTTGGGAAAACCC-3′) had to be the first, second or third most-frequently observed junction sequence of reads with the barcode. We did not count barcodes that were still non-conformant (non-K6x3) after error-correction. The barcodes and counts have been provided in Supplementary Document S1.

To compensate for lower coverage levels in 5k sort sequencing, K6x3 and non-K6x3 barcodes were counted if they were present in at least 1 read in 2 or more technical replicates, and additional error-correction was performed on barcodes before counting them. First, non-K6x3 sequences were merged with K6x3 barcodes if they had 2 or less mismatches in an alignment (1 mismatch is defined as a mismatched base, gap open, gap extension, or a base in unaligned start/end region). VSEARCH (59)

was used to perform alignments with the following arguments: "--allpairs_global", "--userout", "--userfields", "query+target+alnlen+mism+gaps+opens+exts", "--acceptall". Next, remaining non-K6x3 sequences (i.e. those that were not merged into K6x3 barcodes) were clustered with other remaining non-K6x3 sequences using VSEARCH. The following arguments were used: "--cluster_size", "--consout", "--clusterout_sort", "--id 0.94", "--iddef 1", "--sizeorder", "--sizein", "--sizeout", "--uc", "—msaout". When counting barcodes, we looked for exact matches to either type of cluster sequence (K6x3 or non-K6x3) that were present in 2 or more technical replicates. The barcodes and counts have been provided in Supplementary Document S2.

RESULTS

Selection of site A

To search for attP-like sites for our evolution work, we employed the following criteria: (1) putative sites should have high sequence homology to wild-type C31-int attP, (2) they should possess similar degrees of symmetry that are commonly observed in wild-type attP sites, and (3) they must be located in an intergenic region (i.e. we do not want to disrupt a gene). Having a high sequence-similarity to C31-int attP helps to decrease the number of directed-evolution cycles needed, and increases our likelihood of success. Symmetry is important because homodimers (identical integrase molecules bound via their catalytic domains) bind to att-sites. As shown in Fig. 1, attP sites from C31-int, TP901, and Int2 have similar degrees of symmetry, much higher than that of a well-studied pseudo-site, Xq22.1, which has previously performed poorly in directed evolution attempts.

A custom program was written to screen the entire human genome for regions exhibiting the dyad symmetry that is found in wildtype (WT) serine integrase attP sites. Over 1 million sites were returned, and identified genomic sites were plotted as a function of alignment score to C31-int and Bxb1 integrase attP sequences (Fig. 1a). Approximately 75 sites, exhibiting the highest alignment scores (green circle in Fig. 1a), were further analyzed to identify sequences which met the intergenic location requirement outlined above. In addition, the following criteria were applied: (a) sites should have a GC content between 20-60%, which is the range typically observed for wild-type serine integrases, and (b) sites should be unique (e.g. not present in repeat elements).

We focused on C31-int candidates because they were the most promising – putative human attP sites, A through i, are listed in Fig. S5, along with their locus information (chromosome arm and band), the neighboring 5' and 3' genes, and their distance from the site. Further, the symmetry scores for these sites (Fig. 1b) are all similar to three representative wild-type attP sites bound by the phiC31, TP901, and Int2 serine integrases. The sites were prioritized based on proxies for chromatin accessibility: (1) evidence for active expression from neighboring genes, and (2) proximity to DNAseI hypersensitive sites (HSS). Chromatin accessibility was further validated by confirming that Cas9 could mediate dsDNA integration at the loci (data not shown).

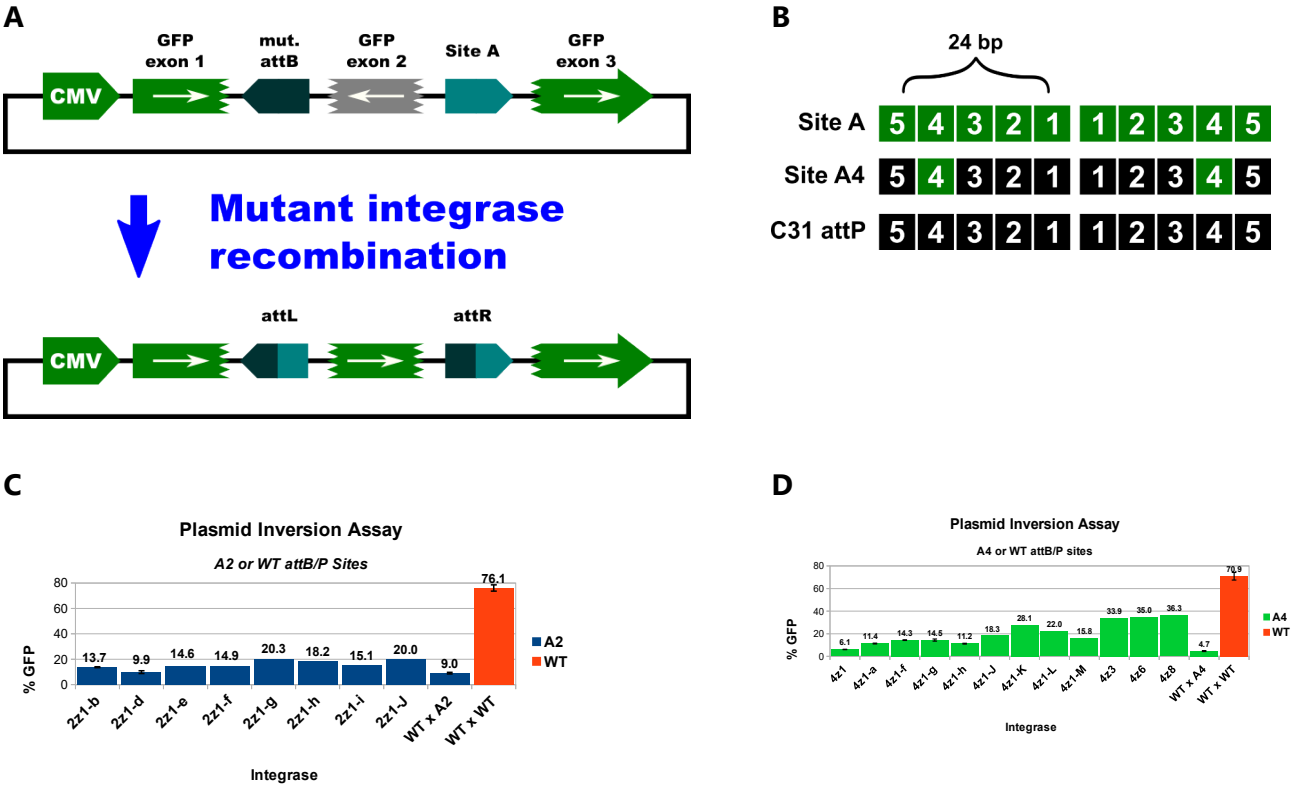
We chose Site A as the directed evolution target because of its homology, symmetry, and location. Compared with WT phiC31 attP, 22 and 29 of the 48 nucleotides are identical in the forward and reverse-complement alignments, respectively (i.e. di-nucleotide core not considered; Fig. 1c). Considering that it would likely be difficult to directly identify a C31-int variant that recognizes Site A in a single round of screening, we instead followed the example of Sarkar et al. (60). I.e., a series of

intermediate attP sequences were generated in order to allow us to gradually evolve C31-int from WT to Site A attP and attB recognition.

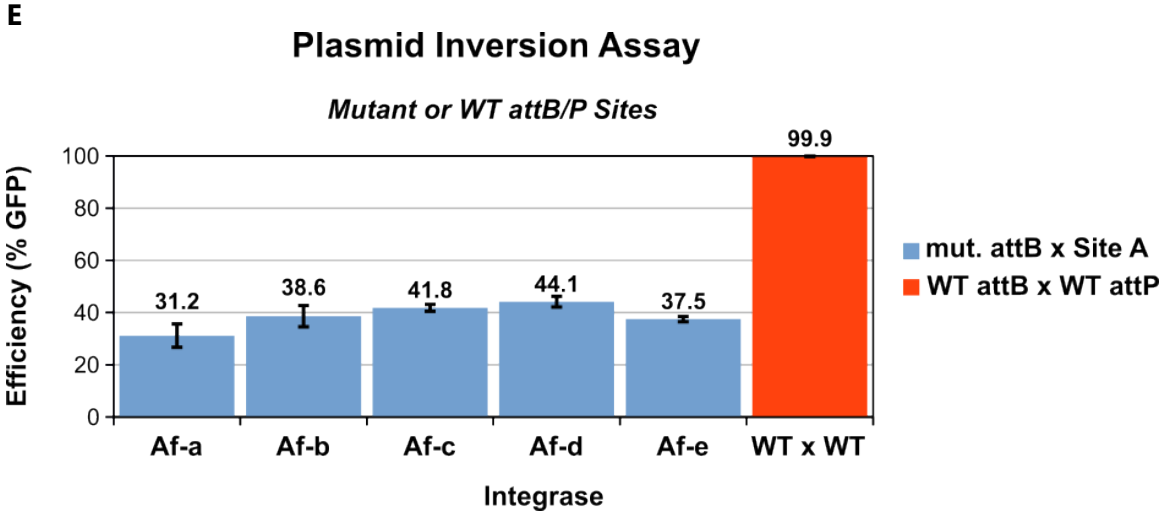
Directed evolution using the inversion assay screen

Site A was divided into 5 segments, A1-A5 (Fig. 2), each of which differs from C31 attP by 3-4 bases in each half site (6-8 bases total) (Fig. S6), and are grouped based on predicted shared proximity to key amino acids in bound C31-int. Intermediate hybrid sequences were utilized in activity screens, and they were named after the corresponding segments that differ from WT attP (e.g. for A2 attP, segment 2 in each attP half-site has been mutated to match the respective segments in Site A; Fig. S6). To predict which residues in C31-int are likely to be in close proximity to DNA, we referred to a pairwise alignment against LI integrase, for which a protein-DNA structure is available (61). Mutations made in the A2 and A4 regions most-severely inhibited wildtype C31-int (Fig. 2c-d), so these intermediates were used to screen our initial variant libraries. Regional mutagenesis was employed (codon saturation combined with insertions and deletions; also segment-specific error-prone PCR) to create amino acid diversity in stretches of the protein that were either predicted to be involved in protein-DNA interactions of the respective segment, or that were experimentally observed to be part of said interactions via alanine-knockout scans (data not shown).

Figure 2: Inversion assay & screen



E



(**A**) To minimize false-positive signals, mutant integrases are subjected to a three-exon GFP plasmid inversion test. If no recombination occurs, the central GFP exon remains in the reverse orientation, which prevents production of green fluorescence above background. In cells with an active variant, the two attachment sites are recombined, which leads to inversion of exon 2 and the production of complete GFP. (**B**) Site A and C31 attP were divided into 5 segments, A1-A5 (1-5). Intermediate sites are named after the segment of C31 attP that has been mutated in both half sites to match the respective Site A sequence. (**C**) Variants with improved activity on A2 intermediate. WTxA2 and WTxWT show the activity of WT C31-int in the A2 and WT attP x attB inversion assays, respectively. Assay performed for 72 hours in HEK293 cells that stably express the respective variant from the H11 locus. (**D**) Variants with improved activity on A4 intermediate. WTxA4 and WTxWT show the activity of WT C31-int in the A4 and WT attP x attB inversion assays, respectively. Assay performed for 72 hours in HEK293 cells that stably express the respective variant from the H11 locus. (**E)** Five variants with the strongest ability to recombine the attachment sites of interest were tested using the plasmid inversion assay over 96 hours in HEK293 cells that stably express the respective variant from the H11 locus. For (C), (D) and (E), a split-intein mCherry system was used to limit analysis to cells that both received the inversion plasmid and that also expressed the variant integrase (single copy expressed from H11 locus). For plots with error-bars (standard error; STDEV/SQRT(3)), N=3 biological replicates.

In mammalian cells, plasmid transfection delivers hundreds-to-thousands of plasmid molecules per cell (62, 63). To maintain a 1-to-1 genotype-to-phenotype connection, we integrated a single variant per cell using an orthogonal serine integrase (Bxb1) at the H11 locus (15, 64) into a previously installed landing-pad (Fig. S4). To further support the expression of only one variant per cell, we also made use of a split-cassette system (Fig. S4), i.e. the integrated variant has a promoter upstream of its coding sequence, while non-integrated variants instead have anti-transcription elements 5′ of their ORF.

A map of the three-exon GFP recombination-reporter plasmids that we used in inversion-assay screens is shown in Fig. 2a. It incorporates an attP and attB site that flank the central inverted GFP exon -- all inversion-reporter constructs had this arrangement, where each contained a specific intermediate (A1-A5) or full Site A attP-attB pair. If no recombination occurs, the central GFP exon remains in the reverse orientation, which prevents production of green fluorescence above background. In cells with an active variant, the two attachment sites are recombined, which leads to inversion of the middle

exon, and thus production of complete GFP (the attL and attR sequences are spliced out with introns). FACS isolation of GFP+ cells was used to enrich active integrase variants.

All screens include an attB sequence that has been mutated in a parallel manner to the Site A attP or intermediate being tested (Fig. S6). For libraries that involve mutagenesis of the zinc-beta-ribbon domain, which is predicted to have a major role in distinguishing attP and attB, we have mutated attB in accordance with the Rutherford et al. model (7, 61). In libraries where the zinc-beta-ribbon domain has not been changed, the attB mutations simply match those made to attP.

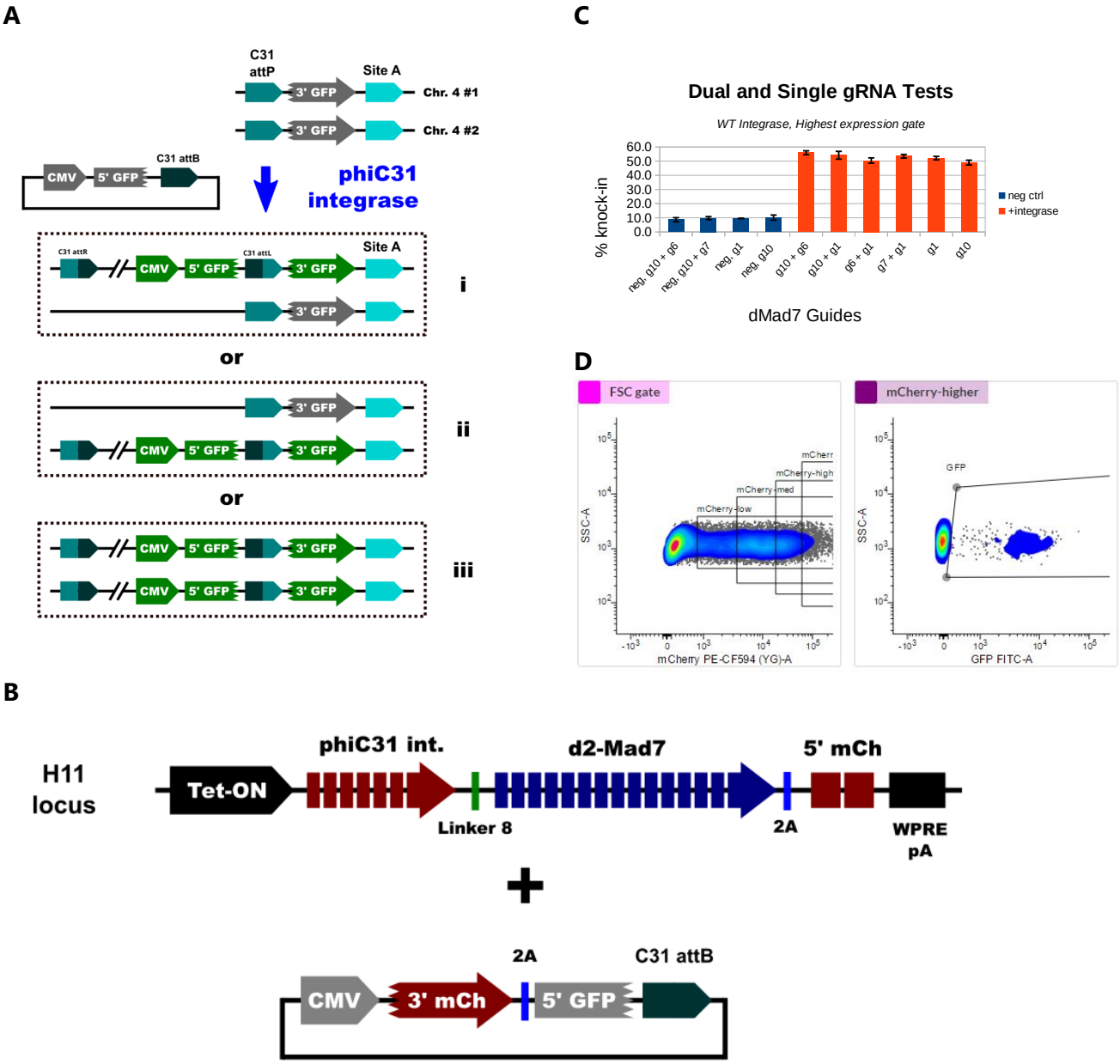
After 2-5 rounds of directed evolution – mutagenesis, activity screens, shuffling of enriched variants – we isolated a variety of A2 and A4 mutants that were able to recombine the respective intermediates at levels higher than wildtype C31-int (Fig. 2c-d). Next, we merged various A2 and A4 mutations to create "Af" (A-full) variants, and some of these were able to recombine the full Site A attP and attB sequences (top five performers shown in Fig. 2e).

Optimization of site A localization

In parallel to our directed evolution of variants, we studied optimization of wildtype (WT) C31-int localization at site A. As we are seeking to limit mutations to regions that impact DNA-interaction specificity, the ability of our variants to reach site A is unlikely to be changed from that of the wildtype enzyme. I.e., based on the LI integrase-DNA structure (61), the regions of the protein responsible for DNA interaction are likely to be separate from those that have an impact on localization. To estimate the upper limit of site A integration efficiency that we will be able to achieve, it is thus appropriate to study how well WT C31-int is able to mediate integration into an attP sequence placed at site A. In addition, any localization improvements we are able to make to WT C31-int – e.g. use of particular nuclear localization sequences (NLS), domain fusions, etc. – are likely to also improve the ability of our variants to reach site A.

To estimate WT C31-int integration efficiency at Site A, we knocked-in a split-GFP recombination reporter into both loci (Fig. 3a) via Cas9-stimulated non-homologous end-joining. This system allows for measurement of integration efficiency at site A within 2-4 days, and it also supports optimization of localization efficiency via pool-based screens.

Figure 3: Localization optimization



(**A**) Split-GFP site A integration efficiency assay. Wildtype C31 attP was placed at both site A loci in HEK293 cells. Downstream of each attP, a splice acceptor, 3′ segment of GFP, and transcription-termination sequence were also introduced. To enable detection of site-specific integration, a donor plasmid was constructed that contains the elements needed to form a complete GFP-expression cassette: CMV promoter, 5′ GFP segment, splice donor, wildtype attB site. After co-transfection of the donor and integrase-expression plasmids, cells where site-specific integration has occurred can be identified by looking for green fluorescence. Integration can happen at one (panels i and ii) or both loci (panel iii). (**B**) WT C31-int dMad7 fusion protein expression cassette used to test impact of different gRNA on site A localization efficiency. A 5′ mCherry segment fused to a trans-splicing intein

domain was co-expressed and separated from int-dMad7 via a 2A-skipping peptide. In the donor plasmid, the remaining 3′ mCherry segment fused to the complementary trans-splicing intein domain was co-expressed with the 5′ GFP segment mentioned in (A), and these two proteins were separated via 2A peptide ribosome skipping. (**C**) Results of site A localization experiments for the indicated guide RNAs and combinations, in transfected cells that express the fusion protein most strongly. Assay was performed for 72 hours in HEK293 cells that stably express the fusion protein from the H11 locus. (**D**) Expression (mCherry) and GFP gating. The rightmost mCherry gate was used to analyze the cells summarized in (C).

The most impactful improvements to localization that we observed were the use of particular NLS tags, and also fusion to a deactivated Cas nuclease at the C-terminus of C31-int. To optimize the NLS, we performed a screen involving 174 sequences from the N- and C-terminal regions of 87 proteins known to localize to various sub-compartments of the nucleus. The 11 most-enriched sequences were individually tested, and the best performance was observed from the C-terminal NLS of the mouse TCOF protein (UniProt O08784; data not shown).

As Cas proteins possess an active DNA-search mechanism that can penetrate chromatin, we tested the ability of C31-dCas9 and C31-dMad7 fusion proteins to improve localization efficiency. Both were able to facilitate the highest levels of integration efficiency in our split-GFP assay, however we found it necessary to heavily intronize their sequences in order to reach reasonable levels of expression (Fig. 3, Fig. S7, Fig. S3 and data not shown). Our intronized C31-dMad7 protein was expressed most strongly (Fig. S3), so we focused on this fusion approach. We tested a variety of guide RNAs that anchor the dMad7 fusion on both flanking regions of site A, and also co-expression of two guides that should anchor the protein upstream and downstream. While other choices would likely be acceptable (i.e. no statistical significance observed between guides), we focused on using the combination of guides g10 and g6 because they performed the best and have the potential to anchor integrase-dMad7 proteins on both sides of site A.

Finally, we also optimized the sequence that connects WT C31-int to dMad7. An 86-member library that mostly consisted of an NLS tag (human nucleoplasmin, mouse TCOF, or human TCOF) connected to a standard peptide linker (22 distinct sequences) was constructed and separately screened using 12 guide RNAs (Fig. S8). Two NLS-linkers dominated: NL1 and NL2 were among the top 3 most-frequently observed for 80% and 60% of the gRNA tests, respectively. We focused on these two NLS-linkers for all of our variant-dMad7 screens and tests.

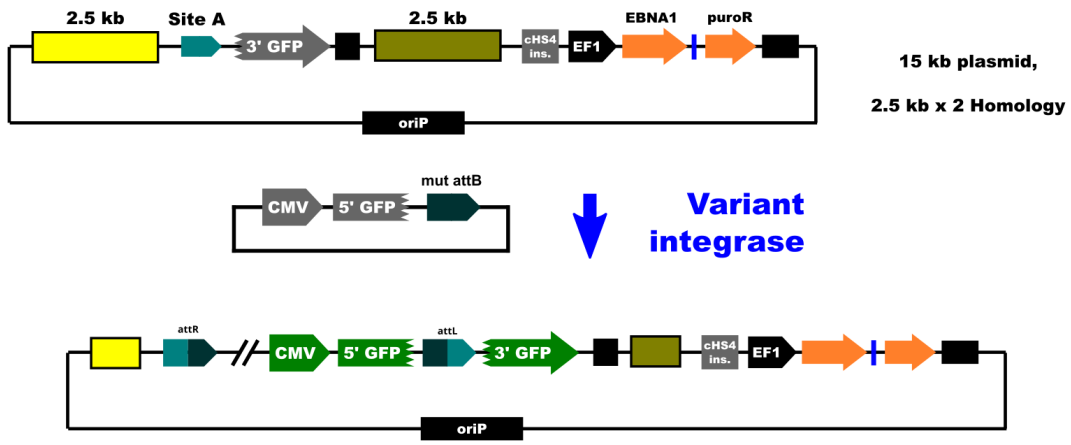
Variant-dMad7 mini-chromosome integration assay and screens

To measure variant integration at Site A in a manner similar to our localization-optimization experiments, we made several attempts to knock-in a split-GFP reporter downstream of Site A. Unfortunately, the knock-in efficiencies were too low, i.e. we could only detect the desired cells in large pools (1,000+ cells). To create an alternative that would be compatible with pool-based screens, we implemented a SAR (Site A Reporter) mini-chromosome assay (Fig. 4a). The needed split-GFP reporter was cloned in a plasmid with 2.5 - 5 kb site A genomic sequence on each side (5-10 kb total), along with other elements needed for EBNA1-mediated extrachromosomal maintenance in human cells (EBNA1 expression, oriP, drug-resistance marker) (65). A wildtype attP version of each plasmid was

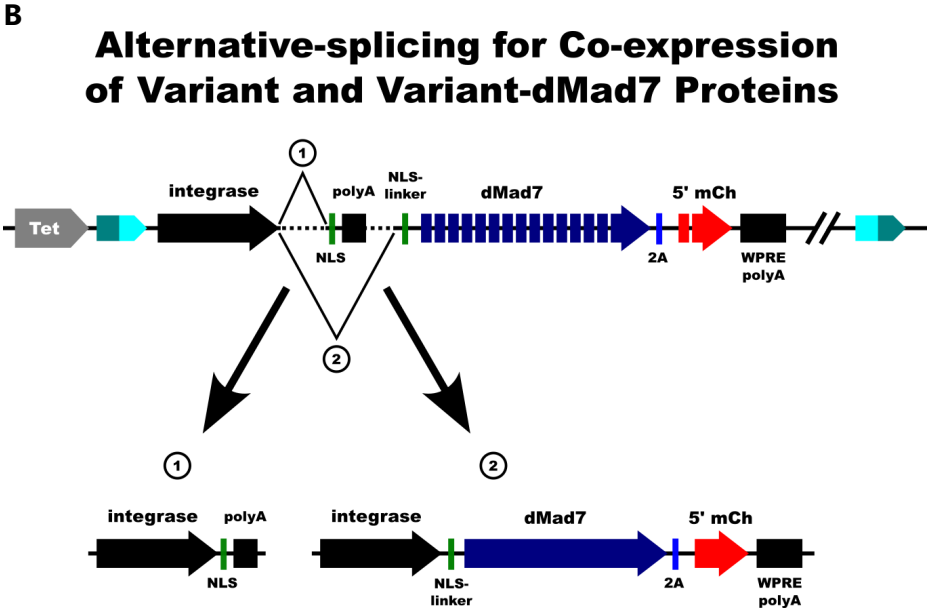
also cloned for comparison to our site A split-GFP assay: we found that WT C31-int performed 5-fold worse in the SAR MAX (10 kb homology) assay, and 3-fold better in the SAR1 (5 kb homology) test (10% vs 2% and 10% vs 30%, respectively; Fig. S8a-b). We chose to focus on the SAR1 assay because with unmodified chromosomal loci present in cells, only a fraction of total Site A integrations will be detected (i.e. the over-performance helps compensate for under-counting).

Figure 4: Minichromosome assay

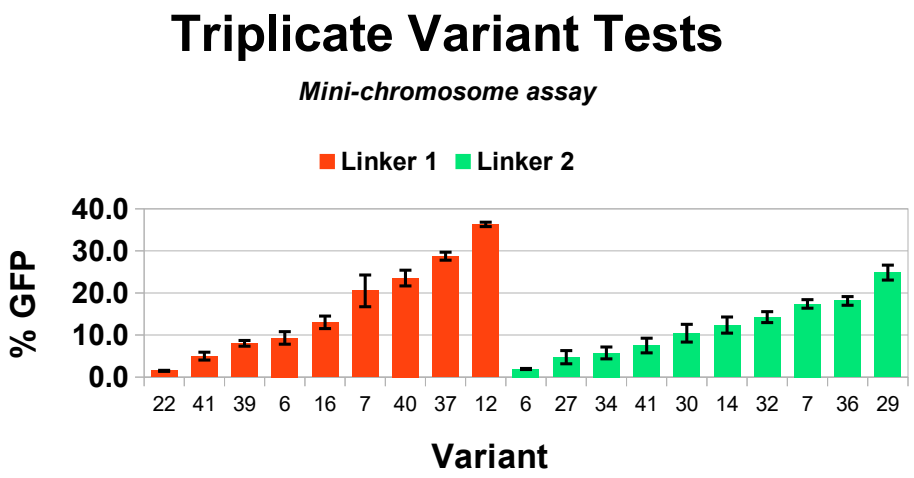
A



B



C



(**A**) Site A-splice acceptor-3′ GFP-polyA was cloned in a plasmid with 2.5 - 5 kb Site A genomic sequence on each side, along with other elements needed for EBNA1-mediated extrachromosomal maintenance in human cells (EBNA1 expression, oriP, puromycin-resistance marker). The 20 kb (SAR MAX, not shown) and 15 kb (SAR 1) plasmids have 5 kb and 2.5 kb genomic Site A sequence on each side (10 and 5 kb total), respectively. (**B**) When the mini-chromosome assay was used in screens, we used the depicted alternative-splicing expression construct to produce the needed solo and fused forms of each variant integrase. (**C**) Results from individual tests of the top-performing variants in the mini-chromosome assay. Assay was performed for 72 hours in HEK293 cells with the Site A SAR1 mini-chromosome that stably expressed the indicated variant using an alternative-splicing cassette from the H11 locus. N=3 biological replicates, error bars are standard error (STDEV/SQRT(3)).

When testing integrase-dMad7 fusion protein activity, we also co-express the relevant integrase protein in "solo" form without the fusion. This is done to ensure that adequate integrase dimers bind to the attB site in our donor plasmids (100s-1,000s present per transfected cell), as the fusion proteins are difficult to express. To accomplish this in pool-based screens where a mutant attB site is used in the donor plasmid, we implemented an alternative-splicing scheme (Fig. 4b), in which each library cell produces both the solo variant protein and the variant-dMad7 fusion. To decide on the specific alternative-splicing introns to use, we tested four combinations described by Aebischer-Gumy et al. (49) in an integrase-2A-GFP/mCherry reporter (Fig. S2a and data not shown). The best performance was observed from "I4(7Y)-I4sh", however we had to include an internal polyA sequence for optimal results (Fig. S2b). This alternative-splicing arrangement is what we implemented to produce solo variants with a C-terminal human TCOF NLS and variant-NL1/NL2-dMad7 fusions (Fig. 4b).

To identify cells that are expressing integrase-dMad7 fusions and that have also been transfected with the donor plasmid, we use a split-mCherry reporter. In the integrase-dMad7 transcript, separated by a 2A skipping peptide, we co-express a 5′ segment of mCherry fused to one half of a trans-splicing intein domain (40, 44). In our donor plasmids, the remaining 3′ segment of mCherry fused to the complementary trans-splicing intein domain is co-expressed with a 5′ segment of GFP, separated by a 2A skipping peptide. Thus, if a cell is expressing integrase-dMad7 and has also received the donor

plasmid, then both halves of the mCherry protein are joined via trans-intein splicing, and red fluorescence can be produced. The 5′ and 3′ segments of GFP are seamlessly connected via RNA cis-splicing after site-specific integration.

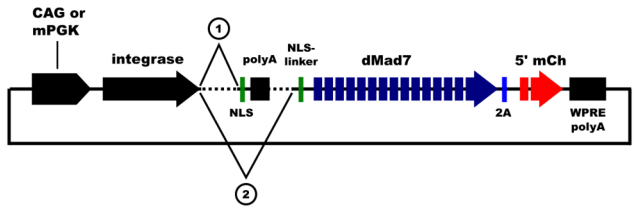
As the mini-chromosome assay is far more challenging than plasmid inversion, we used it to screen for higher-performance variants. Two additional cycles of directed evolution were carried out using the mini-chromosome assay, and we were able to isolate a variety of mutants with improved activity from each generation (Fig. S9). We tested 9 variant-NL1-dMad7 and 10 variant-NL2-dMad7 fusions in triplicate using stable-line expression, and they had a wide range of activities (2-36%; Fig. 4c). In general, more of the top-performers came from the NL1 group (e.g. 29% for V37, and 36% for V12), however there was significant overlap (2-36% for NL1 vs 2-25% for NL2).

Transient expression mini-chromosome assays

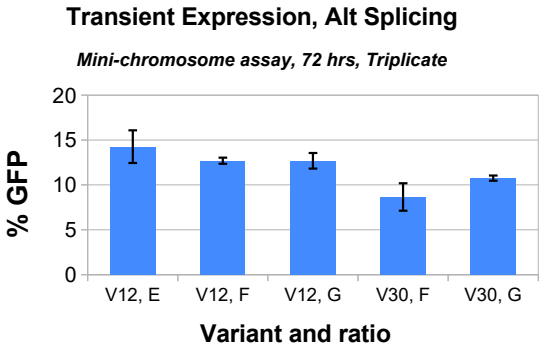
To move closer to real-world applications, we next tested plasmid-based transient-expression of the solo and dMad7 fusion variants. For co-expression of the two variant forms, we compared plasmids with a single alternative-splicing vector, to various ratios of vectors that separately express each protein (solo and fusion). For most of the variants tested, there was not a major difference between the two approaches (data not shown), so for the majority of tests we have used a single alternative-splicing plasmid to co-express solo and variant-dMad7 fusions (Fig. 5a). Overall, as expected due to the difficulty of expressing the large fusion proteins, efficiencies dropped compared to stable-line expression experiments (e.g. V12), however some variants had comparable performance (e.g. V30; Fig. 5b). Transient plasmid-based expression has the potential to temporarily express variants at higher levels than is possible from single-copy stable lines, so we also tested weaker expression of certain variants (i.e. mPGK promoter instead of CAG; Fig. 5c). For certain variants, this led to efficiencies that were competitive with our top strong-promoter reference (CAG-V12, ratio E), so clearly there is no single optimal expression level for our variants.

Figure 5: Transient variant expression, mini-chromosome assay

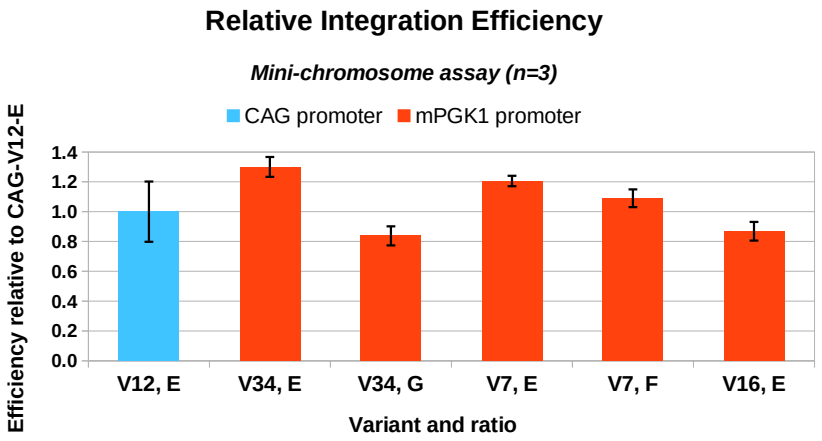
A



B



C



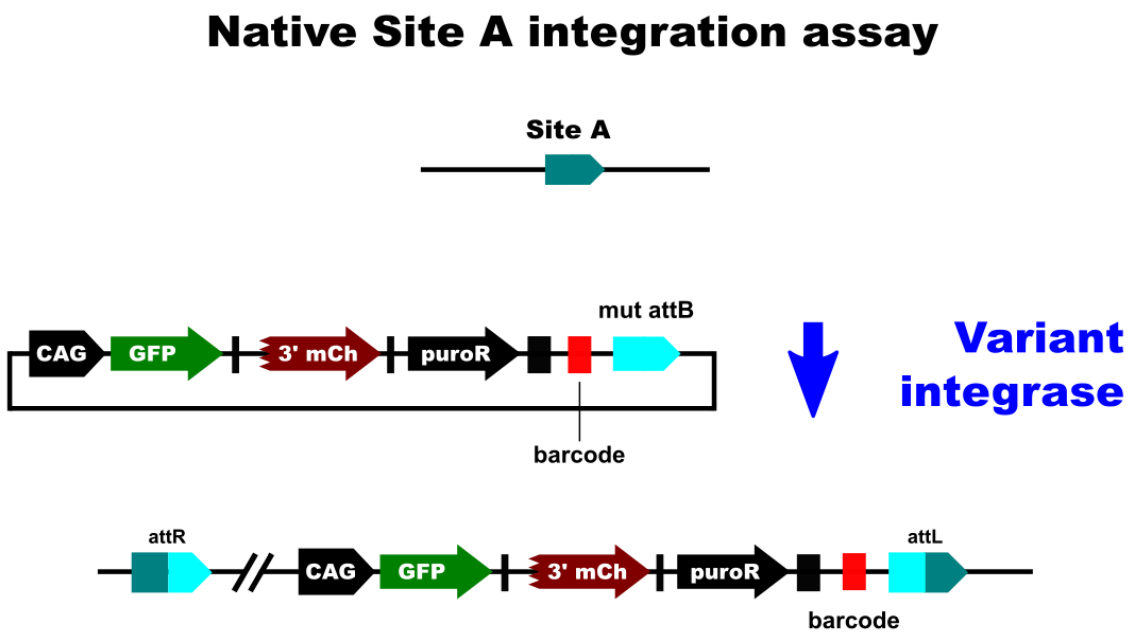
(**A**) Drawing of the plasmid used to express solo and fused forms of variants via alternative-splicing in transient-delivery assays. hU6-driven expression cassettes for guide RNAs g10 and g6 were also present in these vectors (not shown). (**B**) Results from V12 (ratios E, F, G) and V30 (ratios F, G) expressed via the CAG-promoter in the transient expression mini-chromosome assay. (**C**) Results for mPGK-promoter driven expression of V34 (ratios E, G), V7 (ratios E, F) and V16 (ratio E), normalized to CAG-driven V12 ratio E. All experiments were performed over 72 hours in HEK293 cells that contained the mini-chromosome (Site A SAR1). N=3 biological replicates and error-bars are standard error.

Stable line and transient expression native site A integration assays

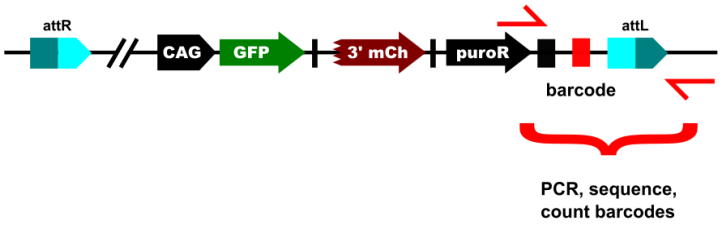
Our final tests estimate the integration efficiency of variants at unmodified (native) site A. The 10.2 kb donor plasmid co-expresses full GFP, 3' mCherry fused to a trans-splicing intein domain, and a puromycin-resistance marker, all separated by 2A skipping peptides (Fig. 6a). For integration-efficiency measurement, a barcode pool (18 nucleotide G/T; K18) was cloned near the mutant attB site. Specific amounts of transfected cells were isolated via FACS three days after plasmid delivery, and they were then outgrown in medium supplemented with puromycin (to remove cells that haven't stably integrated the plasmid). To estimate integration efficiencies, a region encompassing the barcode segment and integration junction was amplified via nested genomic PCR and then sequenced (Fig. 6b). When this assay was performed with HEK293 cells stably-expressing V12-NL1a4-dMad7, V37-NL1a4-dMad7 or V30-NL2-dMad7, we observed 32.4%, 6.94% and 15.5% percent integration efficiencies, respectively (Fig. 6c). For transient plasmid-based expression, only the mPGK-V7-NL2a4-dMad7 vector yielded detectable integration efficiencies, and the values ranged from 2.1% - 13.3% (Fig. 6c). Ratios E and F yielded integration efficiencies close to what was observed for V7 in the mini-chromosome assay, however the conditions were quite unforgiving, i.e. small changes in confluency at the time of transfection majorly impacted the outcomes (e.g. for ratio E, 2.8% vs 6.2-13.3% for a Higher Cell Density (HCD); visa-versa for ratio F).

Figure 6: Native site A assay

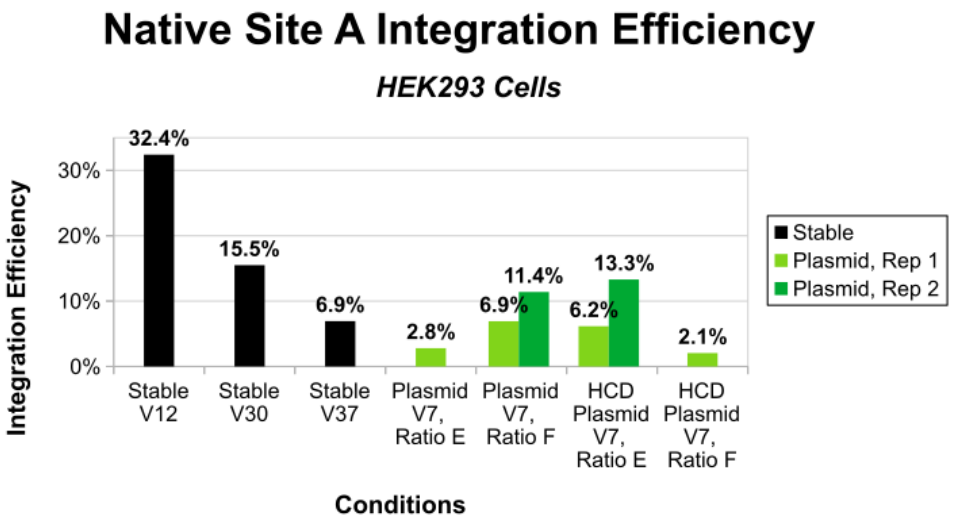
A



B



C



(**A**) The 10.2 kb donor plasmid co-expresses full GFP, 3′ mCherry fused to a trans-splicing intein domain, and a puromycin-resistance marker, all separated by 2A skipping peptides. A barcode pool (18 nt G/T; K18) was cloned near the mutant attB site. (**B**) Specific amounts of transfected cells were isolated via FACS three days after plasmid delivery, and they were then outgrown in medium supplemented with puromycin. To estimate integration efficiencies, the depicted region encompassing the barcode segment and integration junction was amplified via nested genomic PCR and sequenced, then barcodes were counted. (**C**) Barcode-counting results for stable and transient-expression Native Site A assay. For Stable V12, V30 and V37, the expression constructs were V12-NL1a4-dMad7, V30-NL2-dMad7 and V37-NL1a4-dMad7, respectively. NL1a4 refers to alternative splicing with NLS-linker-1. V30-NL2-dMad7 was not expressed using alternative splicing, so a separate solo V30 expression plasmid was co-transfected with the donor. For transient plasmid-based expression, only the mPGK-V7-NL2a4-dMad7 vector (ratios E and F) yielded detectable integration efficiencies. HCD: Higher Cell Density.

DISCUSSION

In this study we have described the development and initial characterization of S-SELeCT integrases that have been evolved to target a novel safe-harbor site in human locus 4p14 (Site A). To our knowledge, this is the first time that a serine integrase has been modified to act on an endogenous symmetrical attachment site (att-site) in the "landing-pad" mode, i.e. where integrase homodimers carry-out the recombination reaction to completion. This is in contrast to pseudosite integration (using homodimers), where assistance from host DNA-repair machinery is often needed, and can result in small indels, large deletions and inversions.

It is also our understanding that we are the first to conduct pool-based activity screens in mammalian cells as part of the directed evolution process for altering serine integrase specificity. This is slower than screening in prokaryotes, however it was necessary in our case because C31-int-family variant activity observed in bacteria has often not translated into mammalian cells (17, 66). In addition to minimizing prokaryotic dead-end-variants, pool-based mammalian screens uniquely enable high-throughput optimization of expression, nuclear localization, and fusion-protein linkers. In addition, mammalian screens implicitly select against sequences that receive inhibitory post-translational modifications and that have problematic interactions with host cell proteins. Mammalian libraries have been more difficult and time consuming to establish because unlike bacteria, plasmid delivery (transfection) is not clonal (instead of single plasmid delivery, 100s-1000s are introduced per cell). Serine integrases and pool-based screening methods like those we have described here offer the most efficient solution to these problems, i.e. they are the best at performing large-insert site-specific single-copy integration (67), and this enables the reasonable establishment of libraries with diversities in the 1e6-1e7 range and higher (e.g. 1e8-1e9 with large-volume or flow electroporators).

To optimize the ability of variants to reach Site A, we created a reporter line where an attP-3′GFP-polyA segment was integrated into both copies of the locus (Fig. 3a). This line was straightforward to create and proved to be quite useful, so we made several attempts to create a version that could be used to detect site A integration. Unfortunately, SpCas9 and Mad7 targets in desirable regions outside of site A were far less efficient at stimulating knock-in of our 3′GFP-pA insert, so we opted to instead use an EBNA1-oriP extra-chromosomal plasmid (mini-chromosome) where the reporter segment was flanked

by site A genomic sequence (5.0 kb total, 2.5 kb on each side; Fig. 4a). This system was used successfully to (i) optimize NLS-tag-linker sequences, (ii) select helpful dMad7 guide RNAs, and (iii) isolate variants that can mediate integration into Site A. The mini-chromosomal system was not always predictive of native Site A integration efficiency, however they are conducted on different time scales (3 days vs 1-2 weeks), so phenomena that take longer to manifest could be at least partially to blame. E.g. efficiencies might drop for variants that have toxic side-effects, and conversely efficiencies could increase for variants that act more slowly (we have observed the latter for certain variants, e.g. Fig. S10).

As we anticipated, transient expression via plasmid delivery majorly reduced integration efficiency (Fig. 4c vs Fig. 5b-c). We were still able to measure integration efficiency for V7, however only when we used a weaker promoter (mPGK) than in the initial tests (CAG), and there was major sensitivity with regards to the confluency of the cells on the day of transfection (which is difficult to control and measure; Fig. 6c). Screens were performed using single-copy expression, so for more optimal results, it might be necessary to implement more sophisticated expression control (to ensure an even and low-level in more cells). It is likely that RNP and/or mRNA would improve results and consistency even further, as they have for a variety of Cas-based gene-editing approaches. However, these are not as trivial to produce given the prokaryotic disconnect observed for C31-int-family proteins (e.g. protein might not fold the same) and the large sequences involved (e.g. truncated mRNA could poison integrase tetramers). Smaller fusion partners would make plasmid, RNP and mRNA delivery more tractable, so they are a worthy future direction to consider.

For native Site A efficiency measurement, we focused on barcode counting to avoid potential pitfalls that can be encountered with ddPCR and other quantitative PCR methods. For example, Pandey et al. (68) recently pointed out a major flaw in the primers used to measure PASTE efficiencies, which may have at least partially contributed to their over-estimations. ddPCR can be used to accurately estimate efficiencies, but calibration with appropriate controls is critical for accuracy (and barcoding can help with that; (68)).

Barcode counting is not free of problems, and we encountered several, including major PCR skew in barcode counts (Fig. S11), PCR noise, and sequencing miscalls. In the future, to reduce skew, different donor backbones and primers should be tested to improve amplification efficiency. If this is successful (e.g. nested PCR no longer needed), then emulsion or droplet PCR could be considered for further reduction of over-representation (i.e. by barcodes that get amplified early in the reaction) (69).

With regards to alternative gene-editing approaches, we consider ee/evoPASSIGE and MINT to be the most similar to ours (12, 68). While the PASSIGE methods allow for rapid development of targeting systems, they are all based on Prime editing, which involves a series of enzymatic steps (nicking of multiple DNA strands, reverse transcription, flap resolution) that can generate undesired outcomes like indels at off-target sites (68, 70, 71). For example, in a recent analysis, Witte et al. point out that eePASSIGE targeted at AAVS1 was reported to have 1.6X more off-targets than on-targets (see Table S6 of Science publication, not PMC version), which was similar to their modified transposon system (68, 72).

Sangamo's MINT system utilizes serine integrases with modified specificity like we do, however they are focused on integrating into endogenous asymmetric att-sites (12). While carrying out directed

evolution screens, their process is conducted in bacteria: full and intermediate symmetric att-sites based on each half of the asymmetric endogenous site are used. Then, to test integration in human cells, they combine a top-performer from each half-site screen with the intention of forming heterodimers that will mediate integration into the endogenous asymmetric site.

Integrating into asymmetric sites that are not necessarily pseudosites vastly increases the number of potential genomic locations that can be targeted (i.e. the only sequence-based requirements would be uniqueness and a non-palindromic di-nucleotide core). However, forcing heterodimerization of serine recombinase catalytic domains would not be trivial and has the potential to create undesirable DNA damage, so this approach is likely to always have a more limited integration-efficiency ceiling compared to symmetric-att-site integration. This is because each integrase variant can form homodimers, which limits the amount of desired tetramers in cells. MINT also uses the WT integrase, which further limits integration efficiencies. This is because WT Bxb1 int can heterodimerize with each variant, and none of those are productive for the desired reaction. In the longer-term, tyrosine integrases would be a better choice for asymmetric integration because the reaction mechanism would not be in conflict with forced heterodimerization (and there is precedence for this with Cre (73)), however the problem currently is that they are not nearly as efficient as high-performance serine integrases (e.g. Voziyanova et al.) (74).

For S-SELeCT, additional important future directions include (i) more rounds of directed evolution, (ii) development of variants that target different safe-harbor sites, (iii) integration efficiency measurements in other cell types such as primary cells and iPSC, (iv) off-target analysis.

Additional rounds of directed evolution are warranted because we observed improvements in efficiency for previous rounds (Fig. S9), and we have not yet reached the maximum that was observed in our WT C31-int experiments (~50%; Fig. 3c) -- i.e. there is clearly room to grow.

Variants that target different locations should be developed because Site A is a novel safe-harbor locus that might not prove ideal for certain therapeutic applications. E.g. integrations into this locus could trigger unwanted impacts on the expression of various proteins/ncRNA, or expression might be silenced after differentiation into particular lineages, or the locus could be unavailable for integration in certain cell types (either directly because of heterochromatin or indirectly because of inhibitory proteins).

Integration efficiency measurements and off-target analysis should be performed in potential primary cells of interest because these outcomes cannot be predicted, and immortal polyploid lines like HEK293 are not representative. Still, we did not make use of HEK293 cells that express the SV40 T antigens, so in combination with RNP-delivery, 10-30% integration efficiencies in permissive primary cells (e.g. iPSCs) would not be unreasonable. We're also doubtful that off-target integrations are a major occurrence for the current generation of S-SELeCT integrases fused to dMad7, however they very likely will be in future derivative variants that have higher efficiencies, as is often the case for a variety of serine integrases and other gene-editing systems. We have this opinion for four major reasons: (i) parts of the integrase that are likely to create hyperactivity were not mutated (e.g. as was done for ee/evoPASSIGE (68)), (ii) our assays only yield signal when sequence-specific recombination has occurred, (iii) the nature of the mutations we've observed to date do not point to specificity-reduction (e.g. very few conversions to alanine, glycine, etc.; Tables S1-S8), and (iv) no results from

inversion-assay experiments have supported specificity loss. If we were weakening specificity, then variants would have predominantly interacted with DNA elsewhere in the cell instead of our plasmids, thus decreasing GFP signal in the assays. This is a phenomenon that was observed when studying the phiC31 RDF, i.e. the pseudosite-prone pCSI integrase expression vector mediated less attP x attB recombination than the pCS-kRI plasmid (which co-expresses the inhibitory RDF), so pCS-kI was created (pCS-kRI without RDF) and it was observed to perform far better at WT recombination (15, 58) (and data not shown). In addition, reduced specificity would not have produced outcomes like we observed for variants 4z-1 and 4z-2, which have majorly lost the ability to recombine WT att-sites, but handle the A4 intermediates better (4z-1) or at a similar low rate (4z-2) (Fig. S12).

AUTHOR CONTRIBUTIONS

Alfonso Farruggio: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. Lin Jiang: Data curation, Investigation, Supervision, Validation, Writing – review & editing. Karen Duong: Data curation, Investigation, Supervision, Writing – review & editing. Cynthia Nguyen: Investigation. Razan Kaddoura: Data curation, Investigation, Supervision. Ruby Tsai: Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review & editing.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

CONFLICT OF INTEREST

The authors are inventors on patents and patent applications relating to C31-int and variants described in the article.

FUNDING

This work was supported by the National Institutes of Health: [grant numbers 1R44GM136045-01, 4R44GM136045-02, 5R44GM136045-03]. Funding for open access charge: National Institutes of Health.

DATA AVAILABILITY

The data underlying this article are available in the article, in its online supplementary material, and in Zenodo at https://doi.org/10.5281/zenodo.18356960.

REFERENCES

1. Baker,M. (2014) Gene editing at CRISPR speed. *Nat Biotechnol*, **32**, 309–312.

2. Kosicki,M., Tomberg,K. and Bradley,A. (2018) Repair of double-strand breaks induced by CRISPR–Cas9 leads to large deletions and complex rearrangements. *Nat Biotechnol*, **36**, 765–771.

3. Alanis-Lobato,G., Zohren,J., McCarthy,A., Fogarty,N.M.E., Kubikova,N., Hardman,E., Greco,M., Wells,D., Turner,J.M.A. and Niakan,K.K. (2021) Frequent loss of heterozygosity in CRISPR-Cas9–edited early human embryos. *Proc. Natl. Acad. Sci. U.S.A.*, **118**, e2004832117.

4. Song,Y., Liu,Z., Zhang,Y., Chen,M., Sui,T., Lai,L. and Li,Z. (2020) Large-Fragment Deletions Induced by Cas9 Cleavage while Not in the BEs System. *Molecular Therapy Nucleic Acids*, **21**, 523–526.

5. Höijer,I., Emmanouilidou,A., Östlund,R., Van Schendel,R., Bozorgpana,S., Tijsterman,M., Feuk,L., Gyllensten,U., Den Hoed,M. and Ameur,A. (2022) CRISPR-Cas9 induces large structural variants at on-target and off-target sites in vivo that segregate across generations. *Nat Commun*, **13**, 627.

6. Stark,W.M. (2017) Making serine integrases work for us. *Current Opinion in Microbiology*, **38**, 130–136.

7. Rutherford,K. and Van Duyne,G.D. (2014) The ins and outs of serine integrase site-specific recombination. *Current Opinion in Structural Biology*, **24**, 125–131.

8. Xu,Z., Thomas,L., Davies,B., Chalmers,R., Smith,M. and Brown,W. (2013) Accuracy and efficiency define Bxb1 integrase as the best of fifteen candidate serine recombinases for the integration of DNA into the human genome. *BMC Biotechnol*, **13**, 87.

9. Durrant,M.G., Fanton,A., Tycko,J., Hinks,M., Chandrasekaran,S.S., Perry,N.T., Schaepe,J., Du,P.P., Lotfy,P., Bassik,M.C., *et al.* (2023) Systematic discovery of recombinases for efficient integration of large DNA sequences into the human genome. *Nat Biotechnol*, **41**, 488–499.

10. Keravala,A., Groth,A.C., Jarrahian,S., Thyagarajan,B., Hoyt,J.J., Kirby,P.J. and Calos,M.P. (2006) A diversity of serine phage integrases mediate site-specific recombination in mammalian cells. *Mol Genet Genomics*, **276**, 135.

11. Chalberg,T.W., Portlock,J.L., Olivares,E.C., Thyagarajan,B., Kirby,P.J., Hillman,R.T., Hoelters,J. and Calos,M.P. (2006) Integration Specificity of Phage ϕC31 Integrase in the Human Genome. *Journal of Molecular Biology*, **357**, 28–48.

12. Fauser,F., Arangundy-Franklin,S., Davis,J.E., Liu,L., Schmidt,N.J., Rodriguez,L., Xia,D.F., Nguyen,N., Scarlott,N.A., Mureli,R., *et al.* (2024) Reprogrammed Serine Integrases Enable Precise Integration of Synthetic DNA. 10.1101/2024.05.09.593242.

13. Wierson,W.A., Simone,B.W., WareJoncas,Z., Mann,C., Welker,J.M., Kar,B., Emch,M.J., Friedberg,I., Gendron,W.A.C., Barry,M.A., *et al.* (2019) Expanding the CRISPR Toolbox with ErCas12a in Zebrafish and Human Cells. *The CRISPR Journal*, **2**, 417–433.

14. Price,M.A., Cruz,R., Bryson,J., Escalettes,F. and Rosser,S.J. (2020) Expanding and understanding the CRISPR toolbox for *Bacillus subtilis* with MAD7 and dMAD7. *Biotech & Bioengineering*, **117**, 1805–1816.

15. Zhu,F., Gamboa,M., Farruggio,A.P., Hippenmeyer,S., Tasic,B., Schüle,B., Chen-Tsai,Y. and Calos,M.P. (2014) DICE, an efficient system for iterative genomic editing in human pluripotent stem cells. *Nucleic Acids Research*, **42**, e34–e34.

16. Zhang,M., Ehmann,M.E., Matukumalli,S., Boob,A.G., Gilbert,D.M. and Zhao,H. (2023) SHIELD: a platform for high-throughput screening of barrier-type DNA elements in human cells. *Nat Commun*, **14**, 5616.

17. Farruggio,A.P. and Calos,M.P. (2014) Serine integrase chimeras with activity in *E. coli* and HeLa cells. *Biology Open*, **3**, 895–903.

18. Zetsche,B., Gootenberg,J.S., Abudayyeh,O.O., Slaymaker,I.M., Makarova,K.S., Essletzbichler,P., Volz,S.E., Joung,J., van der Oost,J., Regev,A., *et al.* (2015) Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. *Cell*, **163**, 759–771.

19. Pourmir,A. and Johannes,T.W. (2012) DIRECTED EVOLUTION: SELECTION OF THE HOST ORGANISM. *Computational and Structural Biotechnology Journal*, **2**, e201209012.

20. DeSantis,M.E. and Shorter,J. (2012) The elusive middle domain of Hsp104 and ClpB: Location and function. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, **1823**, 29–39.

21. Murray,A.N. and Kelly,J.W. (2012) Hsp104 Gives Clients the Individual Attention They Need. *Cell*, **151**, 695–697.

22. Groth,A.C., Olivares,E.C., Thyagarajan,B. and Calos,M.P. (2000) A phage integrase directs efficient site-specific integration in human cells. *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 5995–6000.

23. Gregory,M.A., Till,R. and Smith,M.C.M. (2003) Integration Site for *Streptomyces* Phage φBT1 and Development of Site-Specific Integrating Vectors. *J Bacteriol*, **185**, 5320–5323.

24. Olivares,E.C., Hollis,R.P. and Calos,M.P. (2001) Phage R4 integrase mediates site-specific integration in human cells. *Gene*, **278**, 167–176.

25. Kim,A.I., Ghosh,P., Aaron,M.A., Bibb,L.A., Jain,S. and Hatfull,G.F. (2003) Mycobacteriophage Bxb1 integrates into the *Mycobacterium smegmatis groEL1* gene. *Molecular Microbiology*, **50**, 463–473.

26. Stoll,S.M., Ginsburg,D.S. and Calos,M.P. (2002) Phage TP901-1 Site-Specific Integrase Functions in Human Cells. *J Bacteriol*, **184**, 3657–3663.

27. Morita,K., Yamamoto,T., Fusada,N., Komatsu,M., Ikeda,H., Hirano,N. and Takahashi,H. (2009) The site-specific recombination system of actinophage TG1. *FEMS Microbiology Letters*, **297**, 234–240.

28. Yang,L., Nielsen,A.A.K., Fernandez-Rodriguez,J., McClune,C.J., Laub,M.T., Lu,T.K. and Voigt,C.A. (2014) Permanent genetic memory with >1-byte capacity. *Nat Methods*, **11**, 1261–1266.

29. Kent,W.J. (2002) BLAT —The BLAST -Like Alignment Tool. *Genome Res.*, **12**, 656–664.

30. Casper,J., Speir,M.L., Raney,B.J., Perez,G., Nassar,L.R., Lee,C.M., Hinrichs,A.S., Gonzalez,J.N., Fischer,C., Diekhans,M., *et al.* (2026) The UCSC Genome Browser database: 2026 update. *Nucleic Acids Research*, **54**, D1331–D1335.

31. Keen,J. and Moore,H. (2015) The Genotype-Tissue Expression (GTEx) Project: Linking Clinical Data with Molecular Analysis to Advance Personalized Medicine. *JPM*, **5**, 22–29.

32. Thurman,R.E., Rynes,E., Humbert,R., Vierstra,J., Maurano,M.T., Haugen,E., Sheffield,N.C., Stergachis,A.B., Wang,H., Vernot,B., *et al.* (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75–82.

33. Fath,S., Bauer,A.P., Liss,M., Spriestersbach,A., Maertens,B., Hahn,P., Ludwig,C., Schäfer,F., Graf,M. and Wagner,R. (2011) Multiparameter RNA and Codon Optimization: A Standardized Tool to Assess and Enhance Autologous Mammalian Gene Expression. *PLoS ONE*, **6**, e17596.

34. Chin,J.X., Chung,B.K.-S. and Lee,D.-Y. (2014) Codon Optimization OnLine (COOL): a web-based multi-objective optimization platform for synthetic gene design. *Bioinformatics*, **30**, 2210–2212.

35. Ma,S.L., Vega-Warner,V., Gillies,C., Sampson,M.G., Kher,V., Sethi,S.K. and Otto,E.A. (2015) Whole Exome Sequencing Reveals Novel PHEX Splice Site Mutations in Patients with Hypophosphatemic Rickets. *PLoS ONE*, **10**, e0130729.

36. Mashiko,D., Fujihara,Y., Satouh,Y., Miyata,H., Isotani,A. and Ikawa,M. (2013) Generation of mutant mice by pronuclear injection of circular plasmid expressing Cas9 and single guided RNA. *Sci Rep*, **3**, 3355.

37. Engler,C., Kandzia,R. and Marillonnet,S. (2008) A One Pot, One Step, Precision Cloning Method with High Throughput Capability. *PLoS ONE*, **3**, e3647.

38. Mcgeady,M.L., Wood,T.G., Maizel,J.V. and Woude,G.F.V. (1986) Sequences Upstream from the Mouse *c-mos* Oncogene May Function as a Transcription Termination Signal. *DNA*, **5**, 289–298.

39. Li,X., Liu,S., Wang,D., Chen,H. and Xia,H. (2011) Adenoviral delivered eGFP-intron splicing system for multiple gene RNAi. *Biotechnol Lett*, **33**, 1723–1728.

40. Pinto,F., Thornton,E.L. and Wang,B. (2020) An expanded library of orthogonal split inteins enables modular multi-peptide assemblies. *Nat Commun*, **11**, 1529.

41. Shen,Y., Chen,Y., Wu,J., Shaner,N.C. and Campbell,R.E. (2017) Engineering of mCherry variants with long Stokes shift, red-shifted fluorescence, and low cytotoxicity. *PLoS ONE*, **12**, e0171257.

42. Donello,J.E., Loeb,J.E. and Hope,T.J. (1998) Woodchuck Hepatitis Virus Contains a Tripartite Posttranscriptional Regulatory Element. *J Virol*, **72**, 5085–5092.

43. Yates,J.L., Warren,N. and Sugden,B. (1985) Stable replication of plasmids derived from Epstein–Barr virus in various mammalian cells. *Nature*, **313**, 812–815.

44. Szymczak,A.L., Workman,C.J., Wang,Y., Vignali,K.M., Dilioglou,S., Vanin,E.F. and Vignali,D.A.A. (2004) Correction of multi-gene deficiency in vivo using a single 'self-cleaving' 2A peptide–based retroviral vector. *Nat Biotechnol*, **22**, 589–594.

45. Tasic,B., Miyamichi,K., Hippenmeyer,S., Dani,V.S., Zeng,H., Joo,W., Zong,H., Chen-Tsai,Y. and Luo,L. (2012) Extensions of MADM (Mosaic Analysis with Double Markers) in Mice. *PLoS ONE*, **7**, e33332.

46. Chung,J.H., Whiteley,M. and Felsenfeld,G. (1993) A 5′ element of the chicken β-globin domain serves as an insulator in human erythroid cells and protects against position effect in Drosophila. *Cell*, **74**, 505–514.

47. Forster,A.C. and Symons,R.H. (1987) Self-cleavage of plus and minus RNAs of a virusoid and a structural model for the active sites. *Cell*, **49**, 211–220.

48. Wurmthaler,L.A., Klauser,B. and Hartig,J.S. (2018) Highly motif- and organism-dependent effects of naturally occurring hammerhead ribozyme sequences on gene expression. *RNA Biology*, **15**, 231–241.

49. Aebischer-Gumy,C., Moretti,P., Brunstein Laplace,T., Frank,J., Grand,Y., Mosbaoui,F., Hily,E., Galea,A., Peltret,M., Estoppey,C., *et al.* (2024) Alternative splicing for tuneable expression of protein subunits at desired ratios. *mAbs*, **16**, 2342243.

50. Concordet,J.-P. and Haeussler,M. (2018) CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Research*, **46**, W242–W245.

51. Maeder,M.L., Linder,S.J., Cascio,V.M., Fu,Y., Ho,Q.H. and Joung,J.K. (2013) CRISPR RNA–guided activation of endogenous human genes. *Nat Methods*, **10**, 977–979.

52. Laspia,M.F., Rice,A.P. and Mathews,M.B. (1989) HIV-1 Tat protein increases transcriptional initiation and stabilizes elongation. *Cell*, **59**, 283–292.

53. Cho,M.-S., Yee,H., Brown,C., Jeang,K.-T. and Chan,S. (2001) An oriP expression vector containing the HIV-1 Tat/TAR transactivation axis produces high levels of protein expression in mammalian cells. *Cytotechnology*, **37**, 23–30.

54. Yang,J., Ruff,A.J., Arlt,M. and Schwaneberg,U. (2017) Casting epPCR (cepPCR): A simple random mutagenesis method to generate high quality mutant libraries. *Biotech & Bioengineering*, **114**, 1921–1927.

55. Engler,C., Gruetzner,R., Kandzia,R. and Marillonnet,S. (2009) Golden Gate Shuffling: A One-Pot DNA Shuffling Method Based on Type IIs Restriction Enzymes. *PLoS ONE*, **4**, e5553.

56. Acevedo-Rocha,C.G., Reetz,M.T. and Nov,Y. (2015) Economical analysis of saturation mutagenesis experiments. *Sci Rep*, **5**, 10654.

57. Zhou,X., Vink,M., Klaver,B., Berkhout,B. and Das,A.T. (2006) Optimization of the Tet-On system for regulated gene expression through viral evolution. *Gene Ther*, **13**, 1382–1390.

58. Farruggio,A.P., Chavez,C.L., Mikell,C.L. and Calos,M.P. (2012) Efficient reversal of phiC31 integrase recombination in mammalian cells. *Biotechnology Journal*, **7**, 1332–1336.

59. Rognes,T., Flouri,T., Nichols,B., Quince,C. and Mahé,F. (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, **4**, e2584.

60. Sarkar,I., Hauber,I., Hauber,J. and Buchholz,F. (2007) HIV-1 Proviral DNA Excision Using an Evolved Recombinase. *Science*, **316**, 1912–1915.

61. Rutherford,K., Yuan,P., Perry,K., Sharp,R. and Van Duyne,G.D. (2013) Attachment site recognition and regulation of directionality by the serine integrases. *Nucleic Acids Research*, **41**, 8341–8356.

62. Carpentier,E., Paris,S., Kamen,A. and Durocher,Y. (2007) Limiting factors governing protein expression following polyethylenimine-mediated gene transfer in HEK293-EBNA1 cells. *Journal of Biotechnology*, **128**, 268–280.

63. Cohen,R.N., Van Der Aa,M.A.E.M., Macaraeg,N., Lee,A.P. and Szoka,F.C. (2009) Quantification of plasmid DNA copies in the nucleus after lipoplex and polyplex transfection. *Journal of Controlled Release*, **135**, 166–174.

64. Hippenmeyer,S., Youn,Y.H., Moon,H.M., Miyamichi,K., Zong,H., Wynshaw-Boris,A. and Luo,L. (2010) Genetic Mosaic Dissection of Lis1 and Ndel1 in Neuronal Migration. *Neuron*, **68**, 695–709.

65. Lindner,S.E. and Sugden,B. (2007) The plasmid replicon of Epstein–Barr virus: Mechanistic insights into efficient, licensed, extrachromosomal replication in human cells. *Plasmid*, **58**, 1–12.

66. Sclimenti,C.R. (2001) Directed evolution of a recombinase for improved genomic integration at a native human sequence. *Nucleic Acids Research*, **29**, 5044–5051.

67. Fong,J.H.C. and Ceroni,F. (2025) Transgene integration in mammalian cells: The tools, the challenges, and the future. *Cell Systems*, **16**, 101426.

68. Pandey,S., Gao,X.D., Krasnow,N.A., McElroy,A., Tao,Y.A., Duby,J.E., Steinbeck,B.J., McCreary,J., Pierce,S.E., Tolar,J., *et al.* (2024) Efficient site-specific integration of large genes in mammalian cells via continuously evolved recombinases and prime editing. *Nat. Biomed. Eng*, **9**, 22–39.

69. Terekhov,S.S., Eliseev,I.E., Ovchinnikova,L.A., Kabilov,M.R., Prjibelski,A.D., Tupikin,A.E., Smirnov,I.V., Belogurov,A.A., Severinov,K.V., Lomakin,Y.A., *et al.* (2020) Liquid drop of DNA libraries reveals total genome information. *Proc. Natl. Acad. Sci. U.S.A.*, **117**, 27300–27306.

70. Anzalone,A.V., Gao,X.D., Podracky,C.J., Nelson,A.T., Koblan,L.W., Raguram,A., Levy,J.M., Mercer,J.A.M. and Liu,D.R. (2022) Programmable deletion, replacement, integration and inversion of large DNA sequences with twin prime editing. *Nat Biotechnol*, **40**, 731–740.

71. Yarnall,M.T.N., Ioannidi,E.I., Schmitt-Ulms,C., Krajeski,R.N., Lim,J., Villiger,L., Zhou,W., Jiang,K., Garushyants,S.K., Roberts,N., *et al.* (2023) Drag-and-drop genome insertion of large sequences without double-strand DNA cleavage using CRISPR-directed integrases. *Nat Biotechnol*, **41**, 500–512.

72. Witte,I.P., Lampe,G.D., Eitzinger,S., Miller,S.M., Berríos,K.N., McElroy,A.N., King,R.T., Stringham,O.G., Gelsinger,D.R., Vo,P.L.H., *et al.* (2025) Programmable gene insertion in human cells with a laboratory-evolved CRISPR-associated transposase. *Science*, **388**, eadt5199.

73. Zhang,C., Myers,C.A., Qi,Z., Mitra,R.D., Corbo,J.C. and Havranek,J.J. (2015) Redesign of the monomer–monomer interface of Cre recombinase yields an obligate heterotetrameric complex. *Nucleic Acids Res*, **43**, 9076–9085.

74. Voziyanova,E., Malchin,N., Anderson,R.P., Yagil,E., Kolot,M. and Voziyanov,Y. (2013) Efficient Flp-Int HK022 dual RMCE in mammalian cells. *Nucleic Acids Research*, **41**, e125–e125.