Systematic Development of Reprogrammed Modular Integrases Enables Precise Genomic Integration of Large DNA Sequences

Friedrich Fauser[1,2,3], Sebastian Arangundy-Franklin[1,2], Jessica E Davis[1], Lifeng Liu[1], Nicola J Schmidt[1], Luis Rodriguez[1], Nicholas A Scarlott[1], Rakshaa Mureli[1], Danny F Xia[1], Sarah J Hinkley[1], Bhakti N Kadam[1], Nga Nguyen[1], Stephen Lam[1], Bryan Bourgeois[1], Emily Tait[1], Mohammad Qasim[1], Vishvesha Vaidya[1], Adeline Chen[1], Andrew Nguyen[1], Patrick Li[1], David E Paschon[1], Gregory D Davis[1] and Jeffrey C Miller[1,3]

[1]Sangamo Therapeutics, Inc., Richmond, CA, USA
[2]These authors contributed equally
[3]E-mail: ffauser@sangamo.com and jmiller@sangamo.com

Summary

Despite recent progress in the ability to manipulate the genomes of eukaryotic cells[1-3], there is still no effective and practical method to precisely integrate large synthetic DNA constructs into desired chromosomal sites using a programmable integrase. Serine integrases can perform the necessary molecular steps[4], but only if their natural target site is first installed into the recipient genome by other methods. A more elegant approach would be to directly reprogram the serine integrase itself to target a desired site endogenous to the genome that is different from the natural recognition site of the integrase[5]. Here, we describe the development of a platform of Modular Integrases (the MINT platform), a versatile protein-guided genome editing method that can facilitate site-directed targeted integration of synthetic DNA into chromosomal sites. Through a combination of structural modeling, directed evolution, and screening in human cells we have reprogrammed the specificity of the serine integrase Bxb1. We then utilized these reprogrammed Bxb1 variants to enable precise integration of kilobase-sized constructs into multiple endogenous locations within the human genome with up to 35% efficiency and promising genome-wide specificity. We demonstrate the therapeutic potential of the MINT platform by retargeting Bxb1 to the human TRAC and AAVS1 loci where wild-type Bxb1 has no detectable activity.

Introduction

The ability to integrate synthetic DNA constructs into desired chromosomal locations in eukaryotic genomes would have broad implications for the development of genomic medicines, synthetic biology, agriculture, and basic research. Initially, there was excitement that homology-directed-repair stimulated by engineered nucleases would be able to accomplish this goal, but numerous limitations were encountered when this was assessed[6-9]. Furthermore, despite the significant progress in programmable DNA binding using zinc fingers (ZFs)[10-12], transcription activator-like effectors (TALEs)[13,14], and RNA-guided proteins such as Cas9[15], simply tethering such engineered DNA binding domains to the catalytic domains of recombinases or transposons utilizing a cut-and-paste mechanism has not yielded reagents capable of driving high levels of integration in human cells[16-23]. We initially explored such systems and observed high levels of indels (Extended Data Fig. 1a) which may partially explain the difficulties others have encountered.

Greater success was achieved by inserting ZFs into the tyrosine recombinase CRE to perform inversions[24], however fusions to other types of integrases have not proved nearly as successful[25-29]. There has been some recent progress using CRISPR-associated transposases[30,31] and other nucleic acid-guided systems[32], but such systems require simultaneous delivery of multiple components to mammalian cells and thus far have only achieved modest levels of activity[33].

Large serine recombinases (LSRs - also known as serine integrases) have long been proposed as the ideal tool for genome engineering[34,35] due to their unique properties among other integration systems (**Figure 1a**). In their natural context, an integrase dimer binds to a phage attP site, while another dimer binds an attB site in the bacterial host's genome. The assembled and fully active tetramer then forms a covalent bond to the DNA via its active-site serine residues producing a temporary two base-pair overhang. Subunit rotation facilitates strand exchange and ligation, without leaving free DNA ends or nicks, thus obviating the need to engage the host cell's DNA repair mechanisms. Importantly, in the absence of a phage-encoded reversibility factor, integration proceeds unidirectionally and irreversibly. Nevertheless, the use of LSRs as genome editing reagents has been hampered by the difficulty in their engineering and deployment towards non-cognate targets.

The best results to date have been achieved with systems that use a phage-derived serine integrase such as Bxb1 combined with prime editing to simultaneously introduce the natural attB target site into mammalian cells[36,37]. However, the requirement to first use prime editing to integrate the target site makes the process more complex for therapeutic development, requiring the simultaneous delivery of numerous genetic components to the cell. A more straightforward approach would involve directly reprogramming the target preference of the serine integrase itself while retaining the desirable properties of natural LSRs.

Despite the utility of reprogramming an LSR such as Bxb1[38], there are several technical challenges to overcome. First, while TALEs and CRISPR systems can easily be targeted to desired DNA sequences using simple targeting rules, no one has been able to divine such a "simple code" that governs Bxb1 retargeting. Directed evolution systems have engineered proteins such as ZFs that lack a simple DNA targeting code, but the modularity and compactness of the ZF repeat enables strategies that engineer one zinc finger at a time[39,40]. Directed evolution can also be used on non-modular proteins such as Cre, but this required over a hundred sequential rounds of directed evolution for a given target site[41,42]. A further challenge for reprogramming Bxb1 is that no structure of Bxb1 bound to its target site is currently available.

In this study, we have used a combination of structural modeling and experimental characterization to map critical protein-DNA interactions between Bxb1 and its target site. The only known structures of an LSR suggest that the zinc ribbon domain (ZD) and the recombinase domain (RD) are separated by a flexible polypeptide linker and recognize their portions of the attB and attP targets in a modular fashion[43]. On the assumption that Bxb1 recognizes its target site in a similar way, we developed a strategy for LSR engineering that uses directed evolution to reprogram the key specificity-determining amino acid residues of the RD and ZD domains in parallel and then combine the successful variants into fully reprogrammed LSR variants. We used this strategy to successfully target multiple sites endogenous to the human genome with up to 35% efficiency and to target two therapeutically relevant loci with up to 1% efficiency.

DNA-Bxb1 interaction mapping

To compensate for the lack of a known structure of Bxb1 bound to its DNA target site, we performed an initial experiment designed to map interactions between residues in the Bxb1 RD and ZD domains and key regions of its natural DNA target site. Based on a sequence alignment of

Bxb1 and the regions of the LI integrase known to interact with its DNA target[43,44], we identified 89 residues in Bxb1 that are likely to interact with DNA (Extended Data Fig. 2). We performed scanning mutagenesis on these 89 residues, combined those variants with plasmids bearing symmetric changes to a single position of each half-site of both the attB and attP sequences that are specified by wild-type Bxb1[35] and delivered them to human K562 cells. This experiment identified two clusters of residues which, when mutated, could alter specificity at portions of the DNA target site. A first cluster of residues at positions 231, 233, and 237 was able to alter specificity at positions -10 and -9 of the DNA target site. A second cluster of residues at positions 314, 315, 316, 318, 323, and 325 was able to alter specificity at positions -18 through -13 of the DNA target site. The residue at position 158 was also able to alter specificity at positions -7 and -6 of the DNA target site. Saturation mutagenesis of these residues was able to achieve further specificity shifts. Mapping the Bxb1 mutations that shifted targeting specificity onto a Rosettafold[45] structural model of Bxb1 indicated two regions in the RD domain and one region in the ZD domain where focused protein engineering should be able to alter the DNA targeting specificity (**Figure 1b**). We refer to these regions as the helix, loop, and hairpin respectively based on the structure of each region in our Bxb1 structural model (**Figure 1c,d**). Notably, the central dinucleotide of the attachment sites is not directly recognized by Bxb1 but needs to match between the attP and attB site for integration to occur.

Directed evolution of Bxb1 DNA interaction domains

In order to fully reprogram the specificity of Bxb1, screening of a large combinatorial library of Bxb1 variants is needed. To achieve this, we developed a directed evolution system based on a previously described method[46] whereby a "stuffer" sequence flanked by divergent attB and attP target sequences is placed within the coding sequence of an antibiotic resistance gene such that recombination of these test sequences by a Bxb1 variant expressed in the same bacterial cell restores the open reading frame of the antibiotic resistance gene and allows the bacterial cell to survive an antibiotic challenge (**Figure 2a**). In our system we placed the Bxb1 variant library and the DNA target sites on separate plasmids to enable more rapid testing of the same library of Bxb1 variants against different DNA target sites (**Figure 2b**, Extended Data Fig. 3). This system also provides a convenient means of assessing the DNA targeting specificity of selected Bxb1 variants by using a single Bxb1 variant tested against a library of different target sites (**Figure 2c**).

To inform the design of our Bxb1 variant libraries, we analyzed the structure of the LI integrase and observed that residues in the structure that correspond to residues 231-237 in Bxb1 form an alpha helix that docks with its target DNA in a manner reminiscent of the interaction between a zinc finger (ZF) and its target trinucleotide[47]. Thus, we also adopted the same residue randomization scheme used with ZFs[40] and kept residue 235 fixed as a leucine since this seems analogous to the leucine that is often at +4 of a ZF recognition helix. This resulted in a library of Bxb1 helix variants with residues 231-234, and 236-237 randomized **(Figure 2b)**.

Since ZFs can target 3 bp DNA sequences we hypothesized that the Bxb1 helix might also be able to specify the DNA bases at positions -11, -10, and -9 of the target site. Thus, we performed 64 separate selections using our Bxb1 helix variant library against all 64 possible DNA triplets at positions -11, -10 and -9. Upon deep sequencing the plasmids contained in bacteria that survived a single round of antibiotic challenge, we observed enrichment of four-residue peptide motifs for 37 out of 64 selections. For these successful selections we identified the individual helix sequences that best represented the enriched four-residue motifs and characterized the DNA targeting preferences of Bxb1 variants bearing each selected helix individually. Many selected helices demonstrated dramatic changes in target specificity at positions -11, -10, and -9 of the DNA target site. A comparison of the target preference for the wild-type helix SATALKR and 19 selected helix sequences is shown in **Figure 2d**.

Next, we used our directed evolution system to select Bxb1 variants with mutations at the hairpin region (residues between 314 and 325). Since the entire hairpin region would not be efficiently randomized by a typical library generated in E coli (in the range of 10e9 variants), we adopted a structure-based scheme, where we fully randomized residues 314, 316, 318, 321, 323, and 325 and partially randomized residue 322 **(Figure 2b)**. The residues we fully randomized face the DNA major groove in our structural model while position 322 is likely to influence hairpin structure if it is a proline. We used this Bxb1 variant library in selections with target sites where positions –19 to –12 matched potential Bxb1 target sites in the human genome. Several of our selected hairpins have obvious target preference differences in comparison to the wild-type hairpin, and the positions where the selected hairpins show the strongest sequence preferences differ across selected hairpins (**Figure 2e**).

Finally, we performed selections using a library of Bxb1 variants where positions 154-159 (the loop region) were randomized **(Figure 2b)**. We performed 16 separate selections using this library against all possible DNA dimer sequences at positions -7 and -6 and found that, surprisingly, a single residue change of S157G could shift the sequence preferences of wild-type Bxb1 towards nearly every single dinucleotide at positions -7 and -6 except for CG (**Figure 2f**). Other selected loop variants were able to show improved target specificity relative to the wild-type loop, but the sequences that can be targeted specifically in the loop variants we have characterized tend to be limited to sequences with A or T at position -7 and C or T at position -6 (**Figure 2f**).

Identification of Bxb1 pseudo-sites in the human genome

We next wanted to test the activity of our engineered Bxb1 variants at chromosomal target sites in human cells. To avoid the likely issue of having no detectable integration with our initial fully reprogrammed Bxb1 variants, we decided to first target sites in the human genome where wild-type Bxb1 had detectable activity. In order to identify such Bxb1 pseudo-sites in human K562 cells, we performed both a computational search based on published Bxb1 specificity data[35] and an experimental assessment utilizing anchored multiplex PCR[48] to map the location of integrated donor constructs. To identify as many active sites as possible, we designed the computational search to identify potential pseudo-sites with any sequence at the central dinucleotide and we designed the experimental approach to use a pool of 16 different donor constructs each bearing a different central dinucleotide (Extended Data Fig. 4a). Combining the results of both approaches, we were able to identify 23 sites with at least 0.1% integration in human K562 cells and we selected five sites with between 39% and 61% homology to the natural Bxb1 attB site where wild-type Bxb1 achieves between ~0.20% and ~2.45% integration as test targets for our engineered Bxb1 variants (Extended Data Fig. 4b). Notably, with both computational and experimental approaches we were only able to identify active attB pseudo-sites, and no attP pseudo-sites, mirroring the direction of natural Bxb1-directed integration into its host genome. This is consistent with our finding that "landing pad" cell lines with attB sites pre-integrated into the genome support higher levels of integration than landing pad cells lines with attP sites pre-integrated (Extended Data Fig. 1b).

Directed evolution yields Bxb1 variants with increased activity at human pseudo-sites

As an initial performance validation in human cells of Bxb1 variants derived from our directed evolution system, we tested selected Bxb1 helix variants for integration at a human pseudo-site on chromosome 3 (Extended Data Fig. 5a). First, we performed a helix selection against one half-site of this target site and tested selected variants as mixtures with wild-type Bxb1 since integration also requires binding to the other half-site and the attP site on the donor. This selection yielded multiple families of helix sequences and we tested a total of eight selected helices for their ability to carry out site-specific integration in K562 cells. Helix sequences with an S or T at position 234 were active at the chromosome 3 target even in the absence of wild-type Bxb1. Similar sequences were enriched in selections with other target sites, so we concluded that these represented helix sequences with poor specificity, and we computationally removed similar sequences from other selections. In contrast, selected Bxb1 variants with an N at position 234 such as AGGNLKR were only active at the chromosome 3 target when mixed with wild-type Bxb1 and the DNA targeting specificity characterization of this helix indicated a dramatic shift in specificity towards a T at position -10 (**Figure 2d**) consistent with the GTC sequence at positions -11 to -9 of the half-site we targeted with this helix selection (Extended Data Fig. 5a).

Next, we wanted to investigate whether helix and hairpin variants can be combined to further engineer Bxb1. Initially we performed a selection using our Bxb1 variant hairpin library against the same half-site of the chromosome 3 target that was used for the helix selections. We then combined the most promising selected hairpin LARGRRKWARYR with both the AGGNLKR helix and the more active, but less specific WSSSLKR helix. Both combinations retained full activity at the chromosome 3 site when combined with wild-type Bxb1 but show a substantial decrease without wild-type Bxb1. Thus, WSSSLKR in combination with LARGRRKWARYR now requires wild-type Bxb1 to target the donor and/or other half-site with full activity indicating improved specificity vs. the WSSSLKR helix variant alone. Since both AGGNLKR and LARGRRKWARYR were able to target the intended chromosome 3 half-site more specifically than the wild-type Bxb1 helix and hairpin, we reasoned the combination of this helix and hairpin should be even more specific for the chromosome 3 site and characterized this combination of helix and hairpin using an unbiased genome-wide specificity assay. This fully engineered Bxb1 variant appeared to have fewer integration sites within the human genome than wild-type Bxb1 and we observed that the intended site on chromosome 3 had higher levels of integrations than any other site in the human genome (Extended Data Fig. 5b, Extended Data Fig. 6). Furthermore, if

we had performed the experiment using only a donor with a central dinucleotide that matched the intended target, then only the intended target would have been identified in this experiment.

We then chose two additional endogenous targets identified by the computational genome scan and two endogenous targets identified by the experimental genome-wide screen to serve as DNA target sites for our directed evolution system. We used these targets as an initial test of a screening system that utilizes a large plasmid library of artificial target sequences that will be required to target sites where wild-type Bxb1 has no detectable activity (**Figure 3a**). This system allowed us to rapidly identify evolved helix and hairpin variants that, when combined, are active against their desired half-sites (**Figure 3b,c**, Extended Data Fig. 5d-e), demonstrating the modularity of our MINT platform. We were then able to achieve substantial improvements in activity relative to wild-type Bxb1 at 6 half-sites from these four human pseudo-sites (Extended Data Fig. 5f). Four of these half-sites comprise full pseudo-sites in the *MACO1* and *GYS1* genes and activity at these sites could be increased further by pairing the two variants that worked best at the relevant left and right half-sites, achieving ~35% targeted integration at both loci (**Figure 3d**). For *MACO1* we observed 0.85% TI with wild-type Bxb1 and achieved a 41x increase in activity with engineered Bxb1 variants, while we detected 0.23% TI with wild-type Bxb1 at *GYS1* and achieved a 163x increase with engineered Bxb1 variants.

Retargeting Bxb1 to clinically relevant sites

Encouraged by our success improving Bxb1 performance at chromosomal pseudo-sites, we proceeded to the more difficult challenge of targeting clinically relevant regions of the human genome, such as the well-established safe harbor AAVS1 locus[49] and the T-cell receptor α constant (TRAC) locus (**Figure 4a**), in which wild-type Bxb1 has no detectable activity. We first screened Bxb1 variants against a library of potential target sites within these regions (Extended Data Fig. 7). This resulted in pairs of fully engineered Bxb1 variants (**Figure 4b**) for the TRAC and AAVS1 loci shown in **Figure 4a**. We achieved up to ~1% TI at the TRAC locus using our standard PCR-based NGS assay (**Figure 4c**), with no detectable indels (Extended Data Figure 8). Similarly, we achieved up to ~1% TI at the AAVS1 locus using a digital PCR-based assay (**Figure 4d**, Extended Data Figure 9) and no detectable indels by standard PCR-based NGS assay. As in earlier experiments, we utilized DNA donors with wild-type attP target sites with central dinucleotides matched to the intended genomic target site and included wild-type Bxb1 together with the Bxb1

variants selected for the left and right half-sites of the genomic target site. Notably, we did not observe any integration activity above background levels when we omitted the Bxb1 variant targeted to either the left or right half-site, indicating full retargeting of Bxb1 to TRAC and AAVS1.

Alternative delivery strategies and performance in human T cells

We established our MINT platform in human K562 cells using routine plasmid delivery protocols for the donor molecule. However, plasmid DNA is not an ideal delivery modality for most relevant cell types. We envision that for therapeutic applications Bxb1 will be delivered as mRNA while the donor may be delivered via AAV, minicircle, or comparable methods. To demonstrate that donor delivery is compatible with clinically relevant delivery strategies, we explored alternative delivery modalities for the genetic cargo. We tested both single-stranded AAV (ssAAV) and partially double-stranded AAV (dsAAV) donors and observed that integration was more efficient in our K562 landing pad cell line when making the attP site double-stranded by co-delivering a complementary oligonucleotide (**Figure 5a**). We also used a self-complementary AAV (scAAV) which resulted in up to ~25% TI (**Figure 5b**).

Discussion and future directions

In this work we have demonstrated the direct reprogramming of a site-specific LSR which has been a long-standing challenge for genome engineering[16]. We achieved this through a combination of structural modeling, directed evolution, and screening in human cells. We were able to integrate up to ~3 kb of synthetic DNA into endogenous loci in the human genome with efficiencies of up to 35%. We further demonstrate the therapeutic potential of our MINT platform by retargeting Bxb1 to the human TRAC and AAVS1 loci where we achieved up to 1% integration. In the future, we intend to deploy similar engineering techniques as outlined in this study to further increase the activity of these reagents. We envision that our MINT platform will also accelerate efforts in other research areas where Bxb1 was successfully used, such as for metabolite pathway assembly[50] and various agrobiotechnology applications in model plants[51] and crops[52,53]. We expect the modularity and efficiency of our approach will make it an ideal choice for applications where error-free integration of large synthetic DNA constructs is required.

Techniques that use a cut-and-paste mechanism instead of a DNA replication dependent method have no inherent limit to the size of the construct that can be integrated. This is not only

beneficial for the integration of synthetic DNA into a safe harbor site, as demonstrated in this study, but also for targeting the first intron of a mutated gene to integrate the correct copy of the corresponding coding sequence linked to a splice acceptor. Furthermore, targeting endogenous attP-like sites in addition to attB-like sites would support other genomic rearrangements such as deletions and inversions. Additionally, the ability to target two different endogenous loci in close proximity would enable recombinase mediated cassette exchange (RCME) at endogenous loci which could replace an entire genomic locus with a linear synthetic DNA donor construct. This would enable unprecedented flexibility in types of genomic alterations that can be generated[54]. We were able to successfully target the ~4 kb human AAVS1 locus and a ~2 kb portion of the human TRAC locus which demonstrates a high targeting density that would support RCME at an endogenous locus.

In this study we established a method for retargeting Bxb1 that will likely enable the reprogramming of other serine integrases that have recently been described[37,55]. Since the RD and ZD domains of Bxb1 behave in a modular fashion it is likely that RD and ZD domains from different LSRs can be combined to further expand the ability to target desired genomic loci. Ultimately, our goal is to create an archive of pre-characterized modular protein-guided integrases. Such a pre-characterized archive of integrase variants would allow the *in silico* design of variants to target any sequence of interest without the need for custom directed evolution selections. The LSR engineering approach we demonstrated in this study serves as a blueprint for retargeting LSRs to other gene-sized loci for therapeutic applications and beyond.

Data availability

Amino acid sequences and DNA sequences of constructs used in this study are provided upon request. Illumina sequencing data underlying all experiments will be deposited in the NCBI Sequence Read Archive before peer reviewed publication. Source data are provided with this paper.

Code availability

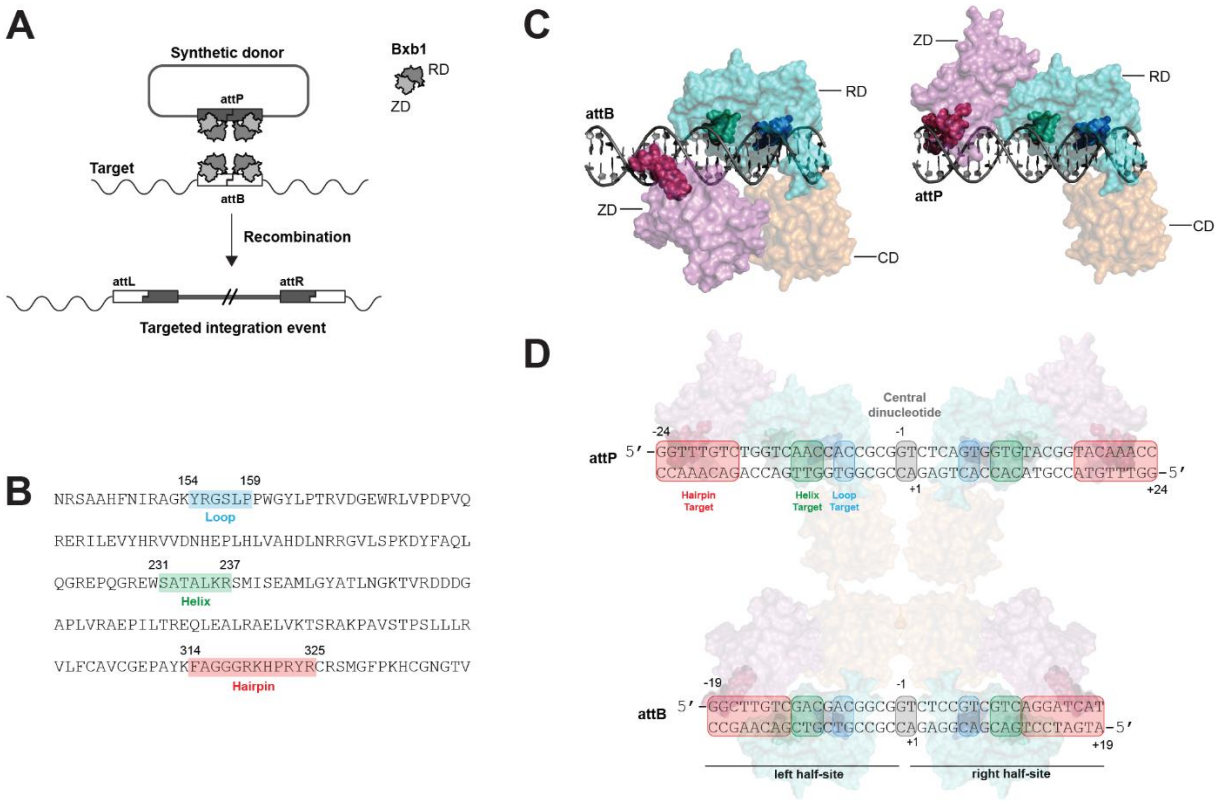
Custom computer code is available upon request although comparable analysis can be performed with publicly available software.

Acknowledgements

We thank Gustavo de Alencastro for assistance with AAV construct design, as well as Sumita Bhardwaj, Tammy Chen, Alicia Goodwin, Emma Petrouski, Sanjna Sridhar, and Hung Tran for assistance with AAV production. We also thank Ed Rebar, Jon Melnick, Deepak Patil, and Charles Paine for contributions to the early development of ZF-targeted Serine Recombinases. We also thank Sandy Macrae, Adrian Woolfson, and Jason Fontenot for their relentless support of this project.

Competing interests

All authors contributed to this work as full-time employees of Sangamo Therapeutics. Sangamo Therapeutics has filed patent applications regarding Integrase systems described in this study, listing F.F., S.A.-F., N.A.S., and J.C.M. as inventors.



306

307 **Figure 1. Bxb1-DNA interaction mapping. a.** Schematic of Bxb1-induced targeted integration
308 of a synthetic circular donor into a genomic target. Attachment sites are bound by Bxb1 dimers
309 and non-reversible recombination is facilitated by a Bxb1 tetramer. **b.** Amino acid sequence of
310 wild-type Bxb1 with residues investigated in this study highlighted. **c.** Structural model of Bxb1
311 bound to attP and attB attachment sites. Bxb1-DNA interaction mapping identified three
312 specificity-determining regions of Bxb1 that can be reprogrammed: the loop and helix regions
313 (residues 154-159 and 231-237 shown in blue and green respectively) in the RD domain of Bxb1
314 and the hairpin region (residues 314-325 shown in red) within the ZD domain of Bxb1. **d**. Sequence
315 of the natural Bxb1 attP and attB attachment sites with the portions recognized by the hairpin,
316 helix, and loop regions of Bxb1 highlighted. The central dinucleotide of the attachment sites is not
317 directly recognized by Bxb1 but needs to match between the attP and attB site for integration to
318 occur.

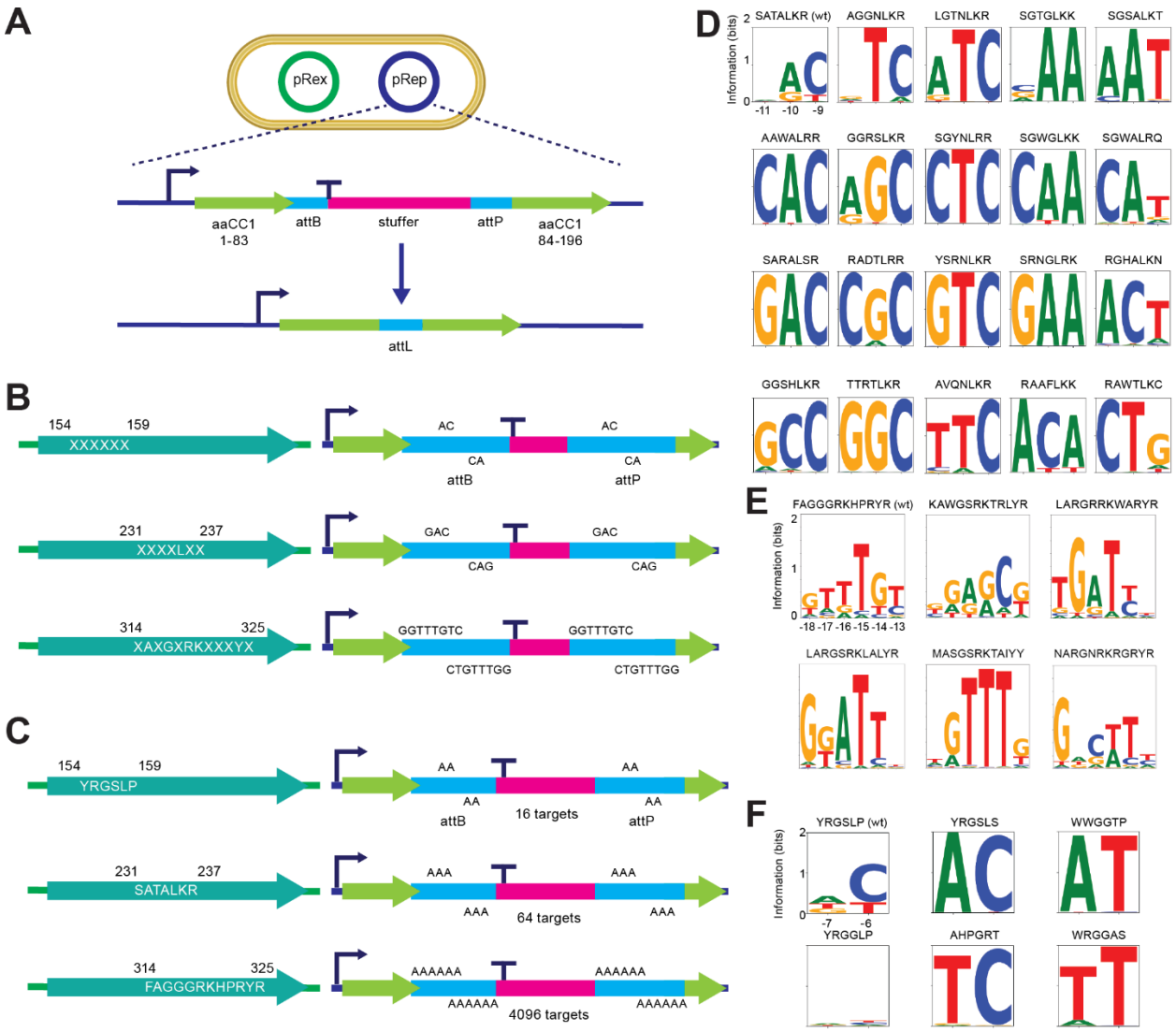
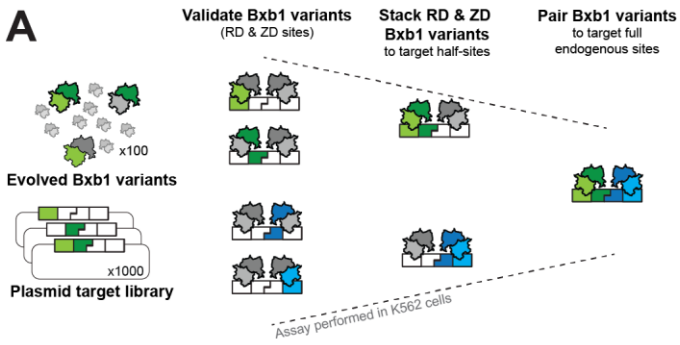
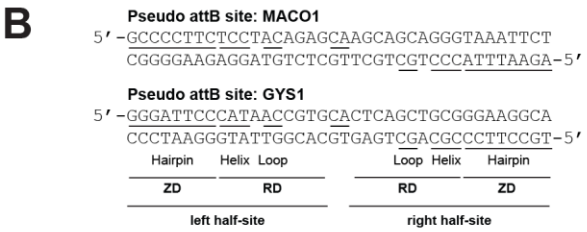


Figure 2. Bxb1 domain engineering using directed evolution. a. Schematic of our two-plasmid directed evolution system; pRex encodes an integrase gene while pRep contains an antibiotic marker disrupted by a stuffer sequence flanked by modified attB and attP sequences such that recombination by an active integrase excises the stuffer sequence and restores the open reading frame of the antibiotic resistance marker. This system enables either screening of libraries of integrase variants against a single DNA target, or screening of a library of DNA targets against a single integrase. **b.** Libraries where the loop, helix or hairpin submotif has been randomized are transformed along with pRep plasmids where the corresponding positions of the attB and attP target sites have been changed **c.** Specificity assay for individual selected integrase variants, where DNA target libraries are tested against a single integrase variant, each target site in the library contains the same modification at both half-sites of the attB and attP sites. **d.** Example DNA specificity plots of selected helix variants. **e.** Example DNA specificity plots of selected hairpin variants. **f.** Example DNA specificity plots of selected loop variants.





C

	Bxb1 peptide sequences derived from directed evolution	
	Helix	Hairpin
Wild-type Bxb1	SATALKR	FAGGGRKHPRYR
MACO1-L	SATALKR	MAGGHRKQALYR
MACO1-R	RAWSLKR	MAGGPRKKGRYR
GYS1-L	HGWSLKV	LARGSRKLALYR
GYS1-R	HGCTLKR	NARGNRKRGRYR

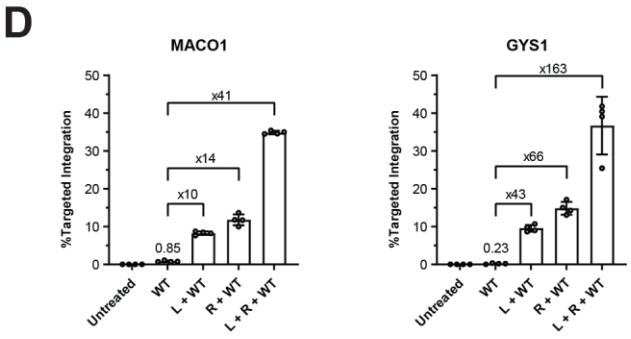


Figure 3. Performance of engineered Bxb1 variants at pseudo-sites in the human genome. a. Schematic of plasmid-based system for testing evolved Bxb1 variants against artificial DNA targets. Our LSR engineering strategy divides each endogenous target site into four "quarter-sites" where the left and right half-sites are each further divided into the portion recognized by the RD domain and the portion recognized by the ZD domain of Bxb1. Individual selected Bxb1 variants are screened against a library of plasmid targets that includes their intended quarter-sites. Successful RD and ZD variants are then combined and tested against the same library of plasmid targets that also includes the relevant half-sites. A single plasmid target library can contain quarter-site and half-site targets for numerous full endogenous target sites. Left and right site candidates derived from this assay can then be tested as pairs against the endogenous target site. See Extended Data Fig. 7b,c for additional details. **b.** Sequence of two Bxb1 pseudo-sites in the human genome. Both sites were identified experimentally using wild-type Bxb1. **c.** Bxb1 peptide sequences of evolved Bxb1 variants that showed improved performance against their corresponding half-sites. **d.** Results from a PCR-based assay demonstrating improved performance of evolved Bxb1 variants against their chromosomal endogenous targets in K562 cells. The presence of a wild-type Bxb1 expression construct is necessary to bind the wild-type attP sequence on the donor plasmid.

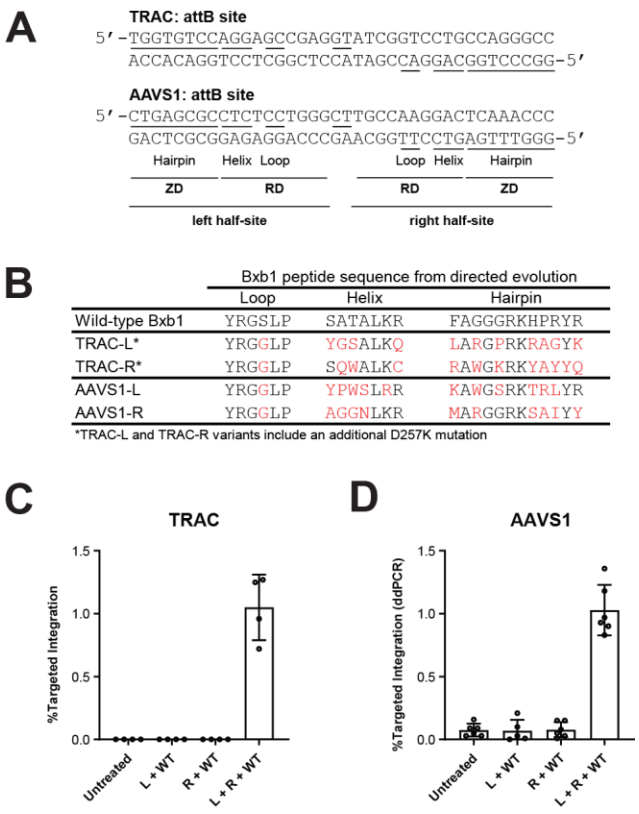


Figure 4. Retargeting Bxb1 to the human TRAC and AAVS1 loci. a. Sequence of the pseudo-sites recognized by the TRAC and AAVS1 Bxb1 variants shown in panel b. **b.** Bxb1 peptide sequences of evolved Bxb1 variants for each of the two TRAC and AAVS1 half-sites generated by the strategy shown in **Figure 3a**. **c.** Results from our standard PCR-based NGS assay demonstrating targeted integration at the depicted TRAC site. See Extended Data Figure 8 for additional details. **d.** Results from a digital PCR-based assay demonstrating targeted integration at the depicted AAVS1 site. See Extended Data Fig. 9 for additional details.

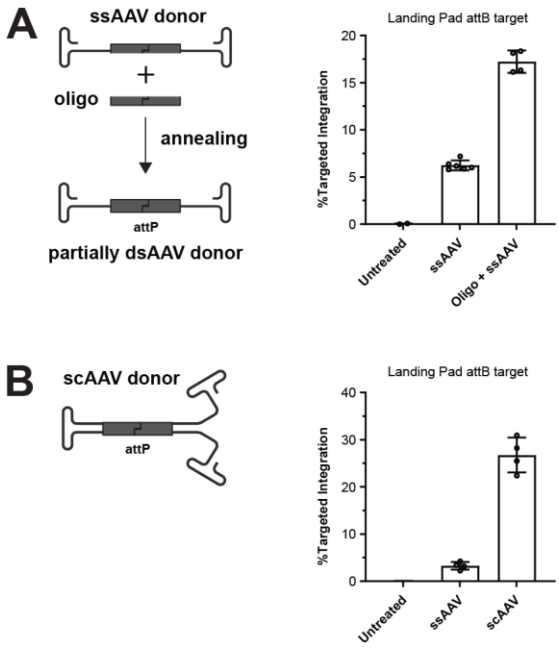


Figure 5. Alternative donor delivery strategies. a. Utilization of a single-stranded AAV (ssAAV) donor for Bxb1-mediated targeted integration. We tested ssAAV as a donor against an attB landing pad in K562 cells and noticed measurable integration levels that can be increased through co-delivery of an oligonucleotide that is complementary to the attP sequence, therefore making the ssAAV donor partially double-stranded. **b.** Utilization of a self-complementary AAV (scAAV) donor for Bxb1-mediated targeted integration.

Methods

Cloning of expression constructs and donors used in mammalian cells

Most constructs were cloned using NEBuilder® HiFi DNA Assembly (NEB, Catalog #E2621X) or Q5® Site-Directed Mutagenesis (NEB, Catalog #E0554S), or synthesized and cloned by Twist Biosciences utilizing their clonal gene services. DNA sequences for all constructs can be found in **supporting sequence information**. All constructs were sequence confirmed using Sanger sequencing services provided by Elim Biopharm Inc., and whole plasmid sequences were verified using either the Nextera XT DNA library prep kit (Illumina, Catalog #FC-131-1096) or whole plasmid sequencing services (Plasmidsaurus Inc.; Elim Biopharma Inc.).

High throughput Bxb1 variants assembly

The Bxb1 variant gene fragments (bases 460 – 1087, corresponding amino acids 144 – 362 with loop, helix and hairpin regions included) were synthesized as eBlocks (Integrated DNA Technologies) and assembled to full length fusion expression cassette through 2-step PCR. In the first step PCR, the eBlocks were amplified with overlapping gene fragments DF148 and DF164 with AccuPrime *Pfx* SuperMix (Invitrogen, Catalog #12344040) and the following thermocycler conditions: initial melt of 95 °C for 3 min; 15 cycles of 95 °C for 30 s, 68 ° for 30 s, and 68 °C for 2 min 30 s; followed by a final extension at 68 °C for 5 min; hold at 4 °C. The full-length expression cassette PCR products were then amplified with forward primer 5'-GCAGAGCTCTCTGGCTAACTAGAG-3' and reverse primer 5'-TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTGAGGGGTCACAGGGATGCCACCCGTAGATC-3' and the following thermocycler conditions: initial melt of 95 °C for 3 min; 30 cycles of 95 °C for 30 s, 68 ° for 30 s, and 68 °C for 2 min 40 s; followed by a final extension at 68 °C for 5 min; hold at 4 °C. The final PCR products were examined on a gel and purified with AMPure XP beads (Bechman# A63881). The purified PCR products were used for mRNA production.

	sequence (5'->3')
DF148	gcagagctctctggctaactagagaacccactgcttactggcttatcgaaattaatacgactcactatagggagagccaa gctgactagcgtttaaacttaagctgatccactagtccagtgtggtggaattcgccaccatgcgggccctggttgttataag gcttagtcgagtgacggatgccacgactagccctgaacgccaactggaaagttgccagcagctctgcgcacagcgcg ggtgggatgtggtgggtgtggctgaagacttggacgtatcaggcgcagtggaccccctttgacaggaaacggcggccca acctcgctcgctggctcgctttcgaggaacagccgtttgacgtgattgtagcgtaccgggtggatcgacttacaagatcaa ttcggcacttgcagcagcttgttcattgggccgaggaccataagaagttggtggtctcagcgacggaagcacattttgac

	accacaacaccattcgctgcagtagtaatcgcgctgatggggactgtcgcacaaatggagttggaagctatcaaagaac ggaatcgcagtgcagcgcacttcaatatcagggctgggaag
DF164	cgccttctgtgaagagcaagttctcgatttgctcggtgatgcagaacggcttgaaaaggtttgggtggccggttctgacag tgctgtggagctggcggaggttaatgcagagcttgtagatctcacgtctctgatcggtagtccagcgtaccgcgcaggat ctccccaacgagaagctcttgacgctagaatagcagctttggcggccagacaggaagaattggagggcttggaagcac ggcctagcgggtgggagtggcgggaaacaggtcaacggtttggtgattggtggcgcgaacaagacaccgcagcaaa aaatacgtggctcagaagcatgaacgtacgccttacttttgacgtgcgaggcggactgacacgcaccatagatttcggag atctccaagagtacgagcagcacttgcggttgggatccgttgttgagcgcctgcatacaggtatgtccggtagcggttctg gctcccatcatcatcaccatcacggctctggccccaagaagaagaggaaggtctgataactcgagctctagaaatcaac ctctggattacaaaatttgtgaaagattgactggtattcttaactatgttgctccttttacgctatgtggatacgctgctttaatgc ctttgtatcatgctattgcttcccgtatggctttcattttctcctccttgtataaatcctggttgctgtctctttatgaggagttgtgg cccgttgtcaggcaacgtggcgtggtgtgcactgtgtttgctgacgcaacccccactggttggggcattgccaccacctg tcagctcctttccgggactttcgctttcccctccctattgccacggcggaactcatcgccgcctgccttgcccgctgctgg acaggggctcggctgttgggcactgacaattccgtggtgttgtcggggaagctgacgtcctttccatggctgctcgcctgt gttgccacctggattctgcgcgggacgtccttctgctacgtcccttcggccctcaatccagcggaccttccttcccgcggc ctgctgccggctctgcggcctcttccgcgtcttcgccttcgccctcagacgagtcggatctccctttgggccgcctccccg cctgtctagatggcccgtttaaacccgctgatcagcctcgactgtgccttctagttgccag

409

410 **mRNA production**

411 The mRNA was prepared with purified PCR products via the mMESSAGE mMACHINE
412 T7 Ultra kit (ThermoFisher, Catalog #AM1345) following the manufacturer's instructions.
413 Synthesized mRNA was purified with RNA cleanXP beads (Beckman, Catalog #A63987) and
414 quantified with Quant-IT kit (Invitrogen, Catalog #Q10213).

415 **AAV production**

416 AAV genome plasmids containing the relevant attP sequences for Bxb1-mediated
417 recombination, flanked by inverted repeat terminal sequences (ITR) for AAV packaging, were
418 cloned using NEBuilder® HiFi DNA Assembly (NEB, Catalog #E2621X). ITRs were mutated to
419 produce scAAV[56]. Bovine growth hormone (bGH) polyA termination sequence was inserted
420 upstream of the 3' ITR to aid transcription termination as well as AAV titration. All constructs were
421 sequence confirmed using Sanger sequencing services provided by Sequetech Corporation. The
422 whole plasmid sequences were verified using Nextera XT DNA library prep kit (Illumina, Catalog
423 #FC-131-1096) or whole plasmid sequencing services provided by Plasmidsaurus Inc. to ensure
424 ITRs are intact prior to AAV production.

The recombinant AAV vectors (rAAV) were produced by triple transfection of suspension human embryonic kidney (HEK) cells in 400 ml flasks, then purified by cesium chloride density gradient centrifugation and dialysis. The rAAV genome was packaged into AAV-DJ (synthetic chimeric capsid of AAV 2/8/9) and titrated using bGH PolyA qPCR assay on the Quantstudio™ 3 Real-time PCR system.

Directed evolution system

Preparation of Integrase Mutant libraries. As a starting point for engineering of Bxb1 variants in E. coli, we cloned a codon optimized ORF into plasmid pRex, which contains a pBR322 origin of replication as well as the L-rhamnose inducible pRhaBAD promoter. Libraries were prepared using inverse PCR and primers encoding degenerate bases using an NNK degeneracy scheme. In order to diversify the "loop" submotif, primers were designed to target residues 154-159. In order to diversify the "helix" submotif, primers were designed to target residues 231-234, and residues 236-237. In order to diversify the "Beta-hairpin" submotif, primers were designed to target residues 314, 316, 318, 321, 323, and 325 using an NNK randomization scheme, while residue 322 was randomized using an SSK randomization scheme. Briefly, iPCR reactions were set up in 50 uL volumes, using 1x KOD ONE Master mix, 0.5uM forward primer, 0.5 uM reverse primer, and 10 ng of template plasmid DNA. Following purification of the PCR products via silica column (Qiagen PCR cleanup kit), and elution in 50 uL of buffer EB, compatible overhangs for ligation were generated by digestion of PCR products using 5 units each of BsaI (NEB, R3733S) and DpnI (NEB, R0176S) in 60 uL of 1x CutSmart buffer. The resulting digest products were then ligated at 10 ng/uL in 1x T4 Ligase buffer, at 8 U/uL T4 Ligase, for 1 hour at room temperature. Following ligation, DNA products were purified using the Qiagen PCR cleanup kit and eluted in 50 uL of buffer EB. The resulting ligated products were used to transform 50 uL of electrocompetent cells (NEB 10Beta, C3020K; or ThermoFisher OneShot Top10, C404052) using a BMX electroporator and a 96-well cuvette (BTX, 45-0450-M) using the manufacturer's protocol and allowed to recover for 1 hour at 37 °C in 1ml total volume of SOC media. The resulting transformations yielded libraries in the order of 4e8 to 5e9 CFUs. After recovery, cells were transferred to 1 L of LB media containing 34 ug/mL chloramphenicol and grown overnight with shaking at 37 °C. The resulting culture was harvested, and plasmids were purified using a Qiagen Plasmid Plus Giga Kit.

Preparation of selection cassette plasmids. Selection plasmid pRep was based on a p15A origin of replication and contained a Spectinomycin resistance cassette. The selection cassette contained a Tet promoter upstream of an open reading frame encoding a gentamicin resistance gene (aacC1) disrupted by a "stuffer sequence". The stuffer sequence was flanked with modified attB sequence at its 5' end, and with a modified attP sequence at its 3' end. The attB and attP sequences were modified such that recombination by an active integrase would leave behind an attL sequence encoding an in-frame peptide insertion, giving rise to an active aacC1. Each selection cassette was generated by the cloning of two ~1050 bp DNA fragments into a linearized pRep backbone plasmid using the NEB HiFi assembly master mix, according to the manufacturer's instructions. Sequence-confirmed plasmids were either prepared by Mini- or Midi- Kits (Qiagen) according to the manufacturer's instructions.

Selection of Bxb1 variants by aacC1 reassembly. Typical selection reactions used 1ug of Bxb1 library, and 1ug selection cassette plasmid. OneShot Top10 electrocompetent cells were mixed with library and selection plasmids before being transferred to a 96-well cuvette, and electroporated as above. After recovery, cells were transferred to 4 mL of LB containing 34 ug/mL chloramphenicol, 50 ug/mL spectinomycin and 0.2% L-rhamnose, then the cultures were shaken at 37 °C for 16 h to allow recombination to take place. After recombination, cultures were harvested, media removed, and the cells were resuspended and transferred to 50 mL Super Broth containing 34 ug/mL chloramphenicol and 20 ug/mL gentamicin. After overnight culture, 1 mL of each culture was used to extract plasmid DNA using the Qiagen Miniprep kit. For selections using agar plates, cells were resuspended in 1 mL of Superbroth and plated on 25x25 cm LB agar plates containing 34 ug/mL chloramphenicol and 20 ug/mL gentamicin. After overnight incubation at 37 °C, colonies were scraped from plates, resuspended in 10 mL LB broth, and a 500 uL volume was removed and used in plasmid purification using Qiagen Miniprep kits.

NGS sequencing of Bxb1 variant submotifs post-selection. Samples were prepared for sequencing using Illumina MiSeq or NextSeq using an integrase-adaptor hybrid primer pair (PCR1) followed by an Illumina-adaptor specific primer pair (PCR2). Briefly, 50uL reactions were prepared using 1x Q5 Hotstart Master mix (NEB M0494S), 0.5uM Forward primer, 0.5uM Reverse primer, and 50ng of plasmid DNA. PCR was performed using the following cycling conditions, 98 °C 30 s, then 15 cycles of 98 °C 5 s, 68 °C 7 s, 72 °C 10 s, followed by a final extension step

at 72 °C for 30s. The resulting PCR product was used as template for PCR2. Briefly, 50 uL reactions were prepared using 1x KOD ONE polymerase master mix (Millipore-Sigma), 0.5 uM Adapter primer 1, 0.5 uM Adapter primer 2, and 1 uL of PCR1 as template. PCR was performed using the following cycling conditions 98 °C 30 s, then 13 cycles of 98 °C 5s, 60 °C 7 s, 72 °C 5 s followed by a final extension step at 72 °C for 30 s. The resulting PCR products were column purified using a Qiagen PCR cleanup kit, and samples were sequenced using standard Illumina Kits for MiSeq or NextSeq.

M**olecular specificity assays.** We built a plasmid reporter to assay the molecular specificity of novel integrase submotifs. Three versions of the pRex plasmids were used as starting point to avoid background activity of the wild-type integrase during cloning. By placing an extra adenine residue within the loop, helix, or hairpin submotif, we created pRex variants where Bxb1 was inactivated by a frameshift-inducing mutation but could be rescued by subsequent mutagenesis. Recombination cassettes from the pRep selection plasmids were amplified by PCR and cloned into the pRex plasmids by Gibson assembly upstream of the Bxb1 ORF. To assay the specificity of selected loops, a single reverse primer and different forward primers which encoded the new loop residues were used in iPCR using the small library of loop target plasmids (16 targets). A similar procedure was used to generate helix mutants using the small library of helix target plasmids (64 targets).

To generate a library of ZD hairpin targets, we amplified and cloned an oligo pool (Twist Biosciences) which comprised all possible attB and attP sites where positions -18 to -13 have been randomized into our pRex plasmid, downstream of the integrase ORF. Such a library ensures that the ZD hairpin targets on attBL and attPL are identical, and the targets on attBR and attPR are inverted repeats of the left side targets. Mutant hairpins are generated by two different iPCR primers, which together encode the novel hairpin sequence.

iPCR reactions were carried out in 50 uL volumes, with 1 x KOD ONE master mix, 0.5uM forward primer, 0.5uM reverse primer, and 20 ng of a plasmid pool containing the relevant recombination cassettes. Thermocycling was carried out using an initial denaturation step of 98 °C for 30s, followed by 35 cycles of 98 °C 10s, 60 °C 5s, 68 °C 30s. PCR amplicons were purified using AMPure XP beads (Beckman) and eluted in 40 uL EB buffer (Qiagen). 30 uL of the purified PCR amplicons were then digested and ligated in a 50 uL one-pot reaction containing 1x T4 ligase

buffer (NEB), 20U DpnI (NEB), 20 U BsaI-HFv2 (NEB), and 400 U of Salt-T4 DNA Ligase (NEB). All reactions were then incubated at 37 °C 30 minutes, 20 °C 30 minutes, 37 °C for 30 minutes. 20 uL of thawed chemically competent NEB 5alpha cells were added 2 uL of the digested/ligated amplicons, and cells were transformed according to the manufacturer's instructions. After recovery, cells were added to 800 uL of media containing 34ug/mL chloramphenicol and 0.2% L-rhamnose (w/v). The resulting culture was incubated for 16 hours at 37 °C with shaking in 96-well deep-well plates, then harvested by centrifugation. Plasmid DNA was extracted using a Qiaprep 96 Turbo Miniprep kit.

To assess the specificity of each clone, primers which flank the resulting attL sequence and contain Illumina adaptor sequences were used to amplify the products of recombination. Briefly, PCR reactions were carried out in 20uL volumes, using 1 x Hotstart Taq master mix (NEB, M0496S), attB_MiSeq forward primer, and attP_Miseq reverse primer, using 1uL of purified plasmid as template. Thermocycling was carried out as follows 98 °C 30s; 25 cycles of [98 °C 5s, 53 °C 10s, 72 °C 5s] Final extension 72 °C 30seconds. A second PCR to install sequencing barcodes was carried out using specific barcoding primers, with the recombination products from each clone being represented by a unique combination of forward and reverse barcoding primers. PCR reactions were carried out in 20 uL volumes, using 1 x KOD ONE master mix, 0.5 uM forward primer, 0.5 uM reverse primer, and 1 uL of the previous PCR product. Thermocycling was carried out as follows 98C 30s; 12 cycles of [98 °C 5s, 60 °C 5s, 68 °C 3s]. The resulting PCR products were column purified using Qiagen PCR cleanup kits, then sequenced using standard Illumina kits for NextSeq or MiSeq.

General mammalian cell culture condition

K562 cells (ATCC, CCL243) were cultured using RPMI-1640 growth medium supplemented with 10% FBS (Fetal Bovine Serum) and 1x PSG (Penicillin-Streptomycin-Gentamycin, Gibco, 10378-016) and maintained at 37 °C with 5% CO2.

K562 tissue culture nucleofection protocol and genomic DNA preparation

Expression constructs were routinely dosed as plasmid DNA (pDNA) in K562 cells. K562 cells were electroporated with pDNA using the SF cell line 96-well Nucleofector kit (Lonza, Catalog#V4SC-2960) or SF Cell Line 384-well Nucleofector Kit (Lonza, Catalog#V5SC-2010),

using manufacturer's protocol. Prior to electroporation, K562 cells were centrifuged at ~300 x g for 5 min, and washed with 1x DPBS (Corning, Catalog#21-031-CV). For 96-well nucleofection, cells were resuspended at 2e5 cells per 12 µl of supplemented SF cell line 96-well Nucleofector solution. 12 µl of cells were mixed with 8 µl of pDNA and transferred to the Lonza Nucleocuvette plate. Nucleofector program 96-FF-120 was used to electroporate K562 cells with the pDNA mix on the Amaxa Nucleofector 96-well Shuttle System (Lonza). After electroporation, cells were incubated for 10 min at room temperature and transferred to a 96-well tissue culture plate containing 180 µl of complete medium (prewarmed to 37 °C).

For 384-well nucleofection, cells were resuspended at 1e5 cells per 14 µl of supplemented SF cell line 384-well Nucleofector solution. 14 µl of cells were mixed with 6 µl of pDNA and transferred to the Lonza Nucleocuvette plate. Nucleofector program FF/120/DA was used to electroporate K562 cells with the pDNA mix on the Amaxa HT Nucleofector System (Lonza, AAU-1001). After electroporation, cells were incubated for 10 min at room temperature and transferred to a 384-well tissue culture plate containing 60 µl of complete medium (prewarmed to 37 °C). K562 cells were incubated for ~72 h and then harvested for quantification of editing events.

Digital Droplet PCR quantification of targeted integration at AAVS1 target site

For AAVS1 target integration ddPCR quantification, 2 targeted integration probes (targeting attL and attR) and 2 reference probes were designed. The target integration probes were designed targeting either attL or attR region which would not be detected in non-transfected cells or plasmid donors. The reference probes were designed within a 10 kb region of the AAVS1 attB target site to mitigate the risk of copy number difference due to abnormal karyotype in K562 cells. The probes and primers were designed with PrimerPlus3 (https://www.primer3plus.com/) based on the recommendation by BioRad manual and were synthesized at IDT (Integrated DNA Technologies, Inc.). The primers and probes were tested in duplex format at different annealing temperatures and with synthetic eblock mixes, non-transfected cell lysate and NTC. The attL and reference2 probes and primers and 57.1°C were selected for AAVS1 target site ddPCR quantification.

attL probe: /56-FAM/CGCCTCTCC/ZEN/TGGGCTCTCAGTGGTGTACG/3IABkFQ/

attL-for: gcatgagatggtggacgag

attL-rev: ggccggtgacatattcctc

Reference2 probe: /5HEX/CGGATCCCG/ZEN/CGCCCAACTCAAGATTGG/3IABkFQ/

Ref-for: agcacaccttgatcttcacc

Ref-rev: agtctctgtcccgattttgg

DNA was extracted with QuickExtract DNA Extraction Solution (Lucigen#QE09050). For each reaction, 50 µl of QuickExtract DNA solution was added to approximately 0.5-1 million pelleted cells, followed by mixing and incubation at 65 °C for 15 min and heat inactivation at 98 °C for 5 min. The cell lysates were mixed by vortexing for 15 seconds before ddPCR. Each ddPCR reaction was prepared and analyzed with a QX200 ddPCR system (Bio-Rad) and ddPCR Supermix for Probes (No dUTP) (Bio-Rad, Catalog #1863024) per Bio-Rad's standard recommendations. All reactions were mixed to 22 µl including 10 U of HindIII-HF (NEB, Catalog #R3104L) and up to 2 µl of QuickExtract lysates. Forward primer, reverse primer and probe were at a 3.6:3.6:1 ratio. Droplets were generated in the droplet generator per Bio-Rad's protocol. Thermocycler conditions: 95 °C for 10 min; 40 cycles of 95 °C for 30 s and 57.1 °C for 60 s; 98 °C for 10 min; and hold at 8 °C. QX Manager Software 2.1 Standard Edition (Bio-Rad) was used for QC and the analysis. The thresholds were set manually at 3000 for channel1/FAM and 1000 for channel2/HEX. All final data was exported into Microsoft Excel for further analysis. The target integration ratio was calculated by the equation: Targeted Integration (%) = 100*CattL/Cref2 (C: volumetric concentration (copies/µl)).

AAV transduction

For evaluating rAAVs. K562 cells were electroporated with 800ng of Bxb1 pDNAusing the SF cell line 96-well Nucleofector kit (Lonza, Catalog#V4SC-2960), using manufacturer's protocol. K562 cells were centrifuged at ~300 x g for 5 min, and washed with 1X PBS (Corning, Catalog#21-031-CV). For 96-well nucleofection, cells were resuspended at 2E5 cells per 12 µl of supplemented SF cell line 96-well Nucleofector solution. 12 µl of cells were mixed with 8 µl of Bxb1 pDNA and transferred to the Lonza Nucleocuvette plate. Nucleofector program 96-FF-120 was used to electroporate K562 cells with the pDNA mix on the Amaxa Nucleofector 96-well Shuttle System (Lonza). After electroporation, cells were incubated for 10 min at room temperature

and transferred to a 96-well tissue culture plate containing 180 µl of complete medium (prewarmed to 37 °C). 30mins post-electroporation, rAAV constructs were dosed at MOI (multiplicity of infections) at 500,000 vg/cell. rAAV donor only control wells were included in parallel. K562 cells were incubated for ~72 h and then harvested for quantification of editing and circularization events. A PCR-based NGS assay was used to measure targeted integration events.

K562 landing pad cell line generation

K562 cells were electroporated with a pair of zinc-finger nuclease (ZFN) mRNA and an ultramer with the attB sequence using the SF cell line 96-well Nucleofector kit (Lonza, Catalog#V4SC-2960), using manufacturer's protocol. The ZFNs generate a double-strand break in the genome that facilitates integration of the ultramer via homology directed repair through the corresponding homologous ends. K562 cells were incubated for ~72 h at 37 °C with 5% CO2. One third of the cells were harvested for quantification of bulk integration of the ultramer using a PCR-based NGS assay. 2/3rd of the cells were maintained for diluting to singles. Samples showing ~10% integration were selected and diluted to singles in a 96-well plate and incubated for ~1.5 weeks. At the end of ~1.5 weeks, the plates were examined under a microscope for the growth of single clones. Cells from wells showing single clones were transferred to a 24-well tissue culture plate and transferred to 37 °C with 5% CO2 for 3 days or until 75-80% confluency is reached. 50% of the cells were harvested for quantification of ultramer integration using the PCR-based NGS assay. Cells were maintained in fresh medium until the NGS assay was completed. Since AAVS1 has 3 alleles, a sample showing 34.39% integration at a single allele was selected for further expansion. The other two wild-type alleles showed a 6bp deletion, but this did not disrupt the performance of the cell line.

Ultramer: 5'-AGGAGACTAGGAAGGAGGAGGCCTAAGGATGGGGCTTTTCGGCC
GGCTTGTCGACGACGGCGGTCTCCGTCGTCAGGATCATCCGGCAGATAAAAGTACCC
AGAACCAGAGCCACATTAACCGGCC-3'

PCR-based NGS assay for targeted integration and indel quantification

3 days post transfection, cells were spun down at ~500 x g for 5 min. Supernatant was discarded, cells were washed in PBS (Corning, Catalog #21-031-CV), and cells were resuspended in 50 µl of QuickExtract DNA Extraction Solution (Lucigen, Catalog #QE09050). Genomic DNA

was extracted by treating the cells to the following protocol: 65 °C for 15 min, 98 °C for 8 min. Target sites were amplified from the genomic DNA using Accuprime HiFi reagents (Invitrogen, Catalog #12346094) and the following PCR conditions: initial melt of 95 °C for 5 min; 30 cycles of 95 °C for 30 s, 55 °C for 30 s and 68 °C for 40 s; and a final extension at 68 °C for 10 min. Primers containing adapters (forward primer adapter: ACACGACGCTCTTCCGATCT; reverse primer adapter: GACGTGTGCTCTTCCGATCT), targeting specific target sites were used at a final concentration of 0.1 μM. Sequences for the primers used can be found in **supporting sequence information**. The PCR productions obtained were then subjected to a second PCR to add Illumina barcodes to the PCR fragments generated in the first PCR. We used Phusion High-Fidelity PCR MasterMix with HF Buffer (NEB, Catalog #M0531L) for the second PCR and used the following PCR conditions, initial melt of 98 °C for 30 s; 12 cycles of 98 °C for 10 s, 60 °C for 30 s and 72 °C for 40 s; and a final extension at 72 °C for 10 min. PCR libraries generated from the second PCR were pooled and purified using QIAquick PCR purification kit (Qiagen, Catalog #28106). Samples were diluted to a final concentration of ~2 nM after they were quantified using the Qubit dsDNA HS Assay kit (Invitrogen, Catalog #Q33231). The libraries were then run on either an Illumina MiSeq using a standard 300-cycle kit or an Illumina NextSeq 500 or an Illumina NextSeq 2000 using a mid-output 300-cycle kit using standard protocol.

Pooled screening of attB sites in K562 cells

Extended Data Fig. 7a outlines the experimental screening of Bxb1 variants against plasmid libraries of artificial target sequences in human K562 cells. Target libraries were designed as shown in Extended Data Fig. 7b,c and cloned using oligo pools (Twist Bioscience). The resulting target libraries were then co-transfected with a universal donor and individual Bxb1 variants. Activity was measured using PCR-based NGS assay.

Genome-wide specificity assay

Our genome-wide specificity assay was adapted from the GUIDE-seq protocol[57], with modifications designed to measure Bxb1-induced off-target editing events. K562 cells were transfected using conditions similar as described above. Transfections involving integrations with multiple donors with different core dinucleotides were performed separately per donor plasmid and then cells were pooled and expanded for 1 week before being spun down for genomic DNA

extraction. Without the DpnI site addition to donor plasmids, the cells were grown out for 3-4 weeks prior to DNA extraction and the DpnI digestion step below was not followed.

Genomic DNA was extracted from K562 cells using Qiagen DNeasy Blood & Tissue kits (Catalog #69504) following the Purification of Total DNA from Animal Blood or Cells Spin-Column Protocol for cultured cells. The optional RNase A incubation was followed for all samples. DNA was eluted in 60 µL Elution Buffer and quantified using the Qubit fluorometer and the Qubit dsDNA HS Assay Kit (Invitrogen, Catalog #Q33231) following the recommended protocol.

Next, 10 uM adapters were prepared by annealing the MiSeq common oligo to each GS_i5 oligo in a 96-well plate format to make a barcoded Y adapter plate. Annealing was performed with 1X oligo annealing buffer (10 mM Tris HCL pH 7.5, 50 mM NaCl, and 0.1 mM EDTA) by following the below thermocycling method: initial melt of 95 °C for 2 min, step-down from 80 °C to 4 °C with -1 °C per cycle and 1 min incubation at each temperature, hold at 4 °C until further use. Adapters were stored at -20 °C, and before use were thawed on ice400 ng (133,000 haploid human genomes) genomic DNA was brought up to 50 uL using 1X IDTE pH 7.5 (IDT, Catalog #11-05-01-05) in each tube of a Covaris 8 microTUBE-130 AFA Fiber H Slit Strip V2 (Covaris, Catalog #520239). Samples were sonicated on a Covaris ME220 using the following settings on a ME220 Rack 8 AFA-TUBE TPX Strip (Covaris, Catalog #PN500609) using the waveguide (Covaris PN 500526): Power 0.0 W, Temperature 19.7 °C, Duration(s) 65.0, Peak Power 40.0, Duty %Factor 10.0, Cycles/Burst 10000, Avg. Power 4.0.

Sheared DNA was purified using 1 volume Ampure XP beads (Beckman Coulter, Catalog #A63880). After beads were added, the solution was mixed and incubated for 5 minutes at room temperature. The mixture was then incubated on a magnet for 5 minutes before the supernatant was removed. 150 uL freshly made 70% ethanol was then used to wash the beads twice, allowing the solution and beads to sit for 30 seconds each time. After the second wash, the beads were dried for 6 minutes before adding 15 uL IDTE pH 7.5 and mixing off the magnet. After 2 minutes the mixture was placed on a magnet and incubated for another 2 minutes. 14.5 uL of the supernatant was collected for the next step.

The reaction was brought up to 50 uL with the addition of CutSmart (final concentration 1X) and 1 uL DpnI (NEB, Catalog #R0176S) and incubated for 1 hour at 37 °C. DNA was purified using 0.8x Ampure XP beads using the same bead clean-up protocol as before. Next, the following

End repair and A-tailing mixture was added to each 14.5 uL DNA mixture, while the reaction was kept on ice: 0.5 uL 10 mM dNTP mix (Invitrogen, Catalog #18427013), 2.5 uL 10X T4 DNA Ligase Buffer (Enzymatics, Catalog #B6030), 2 uL End-repair mix (Enzymatics, Catalog #Y9140-LC-L), 2 uL 10X Platinum Taq Polymerase PCR Rxn Buffer (-Mg2 free) (Invitrogen, Catalog #10966034), 0.5 uL Taq DNA Polymerase Recombinant (5u/ uL) (Invitrogen, Catalog #10342020), and 0.5 uL dsH2O to a total of 22.5 uL. The solution was mixed and incubated on a thermocycler with the following program: 12 °C for 15 min, 37 °C for 15 min, 72 °C for 15 min, hold at 4 °C until further use.

Next, 2 uL T4 DNA Ligase (Enzymatics, Catalog #L6030-LC-L) and 1 uL of one of the 10 uM annealed Y adapters (chosen from a 96-well plate of GS_i5 adapters) was added per end-repaired and A-tailed reaction. Each reaction was mixed and incubated on a thermocycler with the following program: 16 °C for 30 min, 22 °C for 30 min, hold at 4 °C until further use.DNA was then purified using 0.9 volumes Ampure XP beads using the same bead clean-up protocol but using 23 uL IDTE pH 7.5 to resuspend the DNA-bead mixture for elution and collecting 22 uL supernatant after incubation on the magnet.

Next, ligated and sheared DNA fragments were amplified using a primer specific to all adapters (P5_1) and a primer specific to the sequence of interest (GSP1 +/-). To each tube on ice, the following reagents were added: 22 uL DNA from previous step, 3 uL 10X Platinum Taq Polymerase PCR Rxn Buffer (-Mg2 free) (Invitrogen, Catalog #10966034), 1.2 uL 50 mM MgCl2 (Invitrogen, Catalog# 10966034), 0.6 uL 10 mM dNTP mix (Invitrogen, Catalog #18427013), 0.5 uL 10 µM P5_1 primer, 1 uL 10 µM GSP1+/-, 1.5 uL 0.5 M TMAC (Sigma Aldrich, Catalog #T3411), and 0.3 uL Platinum Taq DNA polymerase (5 U/µL) (Invitrogen, Catalog #10966034) to a total of 30.1 uL. Each reaction was mixed and incubated with the following thermocycler conditions: initial melt of 95 °C for 2 min; 15 cycles of 95 °C for 30 s, 70 °C (-1 °C/cycle) for 2 min, and 72 °C for 30 s; followed my 10 cycles of 95 °C for 30 s, 55 °C for 1 min, and 72 °C for 30 s; followed by a final extension at 72 °C for 5 min; hold at 4 °C until further use.

Amplified DNA was purified using the bead clean-up protocol but with 1.2 volumes of Ampure XP beads and using 21 uL IDTE pH 7.5 to resuspend the DNA-bead mixture for elution and collecting 20.4 uL supernatant after incubation on the magnet.

Next, the second round of PCR amplification was performed to add plate barcodes (GS_i7 sequences). To each tube on ice, the following reagents were added: 1.5 ul 10 µM Plate adapter (GS_i7), 20.4 ul DNA from previous step, 3 ul 10X Platinum Taq Polymerase PCR Rxn Buffer (-Mg2 free) (Invitrogen, Catalog #10966034), 1.2 ul 50 mM MgCl2 (Invitrogen, Catalog #10966034), 0.6 ul 10 mM dNTP mix (Invitrogen, Catalog #18427013), 0.5 ul 10 µM P5_2 primer, 1 ul 10 µM GSP2+/-, 1.5 ul 0.5 M TMAC (Sigma Aldrich, Catalog #T3411), and 0.3 ul Platinum Taq DNA polymerase (5 U/µL) (Invitrogen, Catalog #10966034) to a total of 30 ul. Each reaction was mixed and incubated with the following thermocycler conditions: initial melt of 95 °C for 5 min; 15 cycles of 95 °C for 30 s, 70 °C (-1 °C/cycle) for 2 min, and 72 °C for 30 s; followed by 10 cycles of 95 °C for 30 s, 55 °C for 1 min, and 72 °C for 30 s; followed by a final extension at 72 °C for 5 min; hold at 4 °C until further use.

Afterwards, 25 uL of each barcoded reaction was combined into one pool and 0.7 volumes of Ampure XP beads were added. Samples were mixed 10 times and incubated for 5 minutes at room temperature. They were then added to the magnet for 5 minutes then the supernatant was discarded. 2 1x volumes of freshly made 70% ethanol were added, incubating for 30 seconds each time before discarding the supernatant. After the last wash, the beads were air-dried for 6-8 minutes. 75 uL IDTE pH 7.5 was added and the tubes were removed from the magnet and mixed, then incubated for 2 minutes. The reaction was separated on the magnet for 2-4 minutes and then the supernatant was collected into a new tube for NGS library sample submission. Pooled eluate was quantified using the Qubit and the Qubit dsDNA HS kit following the recommended protocol.

Final products were sequenced on a MiSeq or NextSeq 2000 with paired-end 150 bp reads with the cycle settings: 148-10-22-148 for MiSeq or 151-10-22-151 for NextSeq. Samples were sequenced to obtain at least 3,000,000-fold coverage per sample. For MiSeq reactions, 3 uL of 100 uM custom sequencing primer Index1 (5'-3': ATCACCGACTGCCCATAGAGAGGACTCCAGTCAC) was added to MiSeq Reagent cartridge position 13 and 3 uL of 100 uM custom sequencing primer Read2 (5'-3': GTGACTGGAGTCCTCTCTATGGGCAGTCGGTGAT) was added to MiSeq Reagent cartridge position 14.

For NextSeq reactions, 1.98 ul of 100 uM custom sequencing primer Read2 was added to 600 ul Illumina HP21 primer mix for a 0.3 uM final concentration and 3.98 ul of 100 uM custom

sequencing primer Index1 was added to 600 ul BP14 primer mix for a 0.6 uM final concentration. Then 550 µl of custom primer mix was added to custom 1 well or custom 2 well on the reagent cartridge separately.

Computational analysis of directed evolution experiments

Data processing was performed with custom python scripts and logo plots were generated using Logomaker[58]. Sequence reads that don't match sequences encoded by the directed evolution library were filtered out as were sequence reads that were likely artifacts due to being rare sequence reads with only one or two differences from much more common sequence reads. The remaining sequence reads were translated into peptide sequence and the rest of the analysis only considered the peptide sequence of the randomized region. For hairpin selections, the partially randomized position 322 could be either a G, A, R, or P and the motif analysis was conducted separately based on the identity of position 322. The clearest signal was 4 residue motifs within the 6 fully randomized positions. For example, the helix library with a XXXXLXX randomization scheme had a strongly enriched motif of XGGNLXR where X is any amino acid. To account for potential biases in the starting library, motifs were scored based on the FDR corrected p-value of a given motif occurring by chance given the amino acid frequency at each position when considered independently. For helix selections, enriched 4-residue motifs that were observed in selections with a wide variety of different DNA target sequences were considered to be non-specific and were filtered out. Compatible four residue motifs (e.g. XGXNLKR and XXGNLKR) were then combined to create motifs. Several peptides that best matched the most enriched 4 residue motifs in each combined motif were then chosen for additional characterization.

Computational analysis of pooled screening of attB sites in K562 cells

A custom python script was used to count sequence tags corresponding to different recombined targets. Counts were normalized to the total number of sequence reads obtained for a given sample. Normalized sequence tag counts for a given Bxb1 variant + wild-type Bxb1 at a given site were then compared to the normalized sequence tag counts for wild-type Bxb1 alone at for the same sequence tag.

Computational analysis of chromosomal targeted integration events

Sequence data was first processed using our indel analysis software pipeline. The output from this analysis was further processed using a custom python script that identified aligned sequence reads that contained the sequence from the right attP half-site from the donor. Sequence reads that contained this integration tag,but were not the expected length were scored as "TI + indel". Sequence reads that contained this integration tag and were the expected length were scored as "perfect TI". Sequence reads that did not contain the integration sequence tag were scored as either wild-type amplicon or non-TI indel based on the output of the indel analysis software.

Computational identification of Bxb1 pseudo-sites

Raw data from Bessen et al. was processed to produce a position weight matrix for the Bxb1 attB and attP target sites. A custom python script scanned the human genome for potential target sites that matched the strongly preferred G nucleotide at position -4 and at position +4. Sequences that met these criteria were then scored against the position weight matrix. The left and right half-sites of the natural Bxb1 attB site are quite different from each other and likely represent different binding modes of the Bxb1 hairpin region. Thus, for attB sites, each potential site was scored against a position weight matrix representing an inverted repeat of the left half-site of the natural attB sequence, and inverted repeat of the right half-site of the attB sequence, the composite attB left and right half-sites on the top strand of DNA, or the composite attB left and right half-sites on the bottom strand of DNA. The top 12 scores for each category were experimentally characterized for a total of 48 potential attB pseudo-sites in the human genome. The natural attB site is much more symmetric so the position weight matrix for the left and right half-sites was averaged together and this averaged position weight matrix was used to score both sites of potential attP pseudo-sites in the human genome. The top 48 scoring potential attP pseudo-sites were also characterized experimentally.

Computational analysis of genome-wide specificity assay

NGS reads were demultiplexed, adapter trimmed, and filtered for a minimum quality threshold of 14 over all bases. Samples then underwent analysis for plasmid integration site detection. NGS samples that were analyzed for plasmid integration site detection were processed to remove remaining contaminant unintegrated plasmid reads due to incomplete DpnI digestion or fragment removal, aligned to the hg38 genome, and potential integration sites were summarized. First, reads that contained both attP sequence 5' and 3' of the dinucleotide were removed from

analysis, corresponding to unintegrated donor plasmid reads. Then all sequence up to the start of the dinucleotide (up to and including the 5' attP sequence) was removed, leaving the remaining sequence to align to the hg38 genome using Bowtie2. Alignments with a MAPQ less than 23 were removed from the analysis. Next, common read1 start locations, which correspond to unique genomic shear locations and ligation events generated in the protocol, were used to deduplicate common reads. Unique read1 start and dinucleotide positions were then summed per dinucleotide position to generate a list of deduplicated reads per potential integration event. Next, all reads per AMP-seq reaction were summed per alignment position in the genome and per alignment orientation (top and bottom strands). Then, positions were combined into a single potential integration location per AMP-seq reaction and alignment orientation if they fell within a 50 bp window of one another, all reads per this grouping were summed and the coordinate with the most reads was kept per group. Lastly, potential integration locations across top and bottom strand alignments and between both AMP-seq "plus" and "minus" reactions per original transfected sample were combined. Common potential integration locations were merged into one potential integration location if within 50 bp of one another (this would encompass alignments separated by a dinucleotide as a result of sequencing upstream and downstream of integration in both AMP-seq reactions). The final list of potential integration loci were inspected for expected integration genotypes (2 merged locations, in opposite alignment orientation, separated by a dinucleotide that corresponds to the donor plasmid dinucleotide used in the assay).

References

1 Anzalone, A. V. *et al.* Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149-157 (2019). https://doi.org:10.1038/s41586-019-1711-4

2 Gaudelli, N. M. *et al.* Programmable base editing of A*T to G*C in genomic DNA without DNA cleavage. *Nature* **551**, 464-471 (2017). https://doi.org:10.1038/nature24644

3 Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420-424 (2016). https://doi.org:10.1038/nature17946

4 Stark, W. M. Making serine integrases work for us. *Curr Opin Microbiol* **38**, 130-136 (2017). https://doi.org:10.1016/j.mib.2017.04.006

5 Forum: CRISPR roundtable with Doudna and Liu. *Nat Biotechnol* **38**, 943 (2020). https://doi.org:10.1038/s41587-020-0619-8

6 Urnov, F. D. *et al.* Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature* **435**, 646-651 (2005). https://doi.org:10.1038/nature03556

7 Perez, C. *et al.* Factors affecting double-strand break-induced homologous recombination in mammalian cells. *Biotechniques* **39**, 109-115 (2005). https://doi.org:10.2144/05391GT01

8 Kosicki, M., Tomberg, K. & Bradley, A. Repair of double-strand breaks induced by CRISPR-Cas9 leads to large deletions and complex rearrangements. *Nat Biotechnol* **36**, 765-771 (2018). https://doi.org:10.1038/nbt.4192

9 Haapaniemi, E., Botla, S., Persson, J., Schmierer, B. & Taipale, J. CRISPR-Cas9 genome editing induces a p53-mediated DNA damage response. *Nat Med* **24**, 927-930 (2018). https://doi.org:10.1038/s41591-018-0049-z

10 Ichikawa, D. M. *et al.* A universal deep-learning model for zinc finger design enables transcription factor reprogramming. *Nat Biotechnol* **41**, 1117-1129 (2023). https://doi.org:10.1038/s41587-022-01624-4

11 Miller, J. C. *et al.* Enhancing gene editing specificity by attenuating DNA cleavage kinetics. *Nat Biotechnol* **37**, 945-952 (2019). https://doi.org:10.1038/s41587-019-0186-z

12 Paschon, D. E. *et al.* Diversifying the structure of zinc finger nucleases for high-precision genome editing. *Nat Commun* **10**, 1133 (2019). https://doi.org:10.1038/s41467-019-08867-x

13 Boch, J. *et al.* Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* **326**, 1509-1512 (2009). https://doi.org:10.1126/science.1178811

14 Miller, J. C. *et al.* A TALE nuclease architecture for efficient genome editing. *Nat Biotechnol* **29**, 143-148 (2011). https://doi.org:10.1038/nbt.1755

15 Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821 (2012). https://doi.org:10.1126/science.1225829

16 Akopian, A., He, J., Boocock, M. R. & Stark, W. M. Chimeric recombinases with designed DNA sequence recognition. *Proc Natl Acad Sci U S A* **100**, 8688-8691 (2003). https://doi.org:10.1073/pnas.1533177100

17 Chaikind, B., Bessen, J. L., Thompson, D. B., Hu, J. H. & Liu, D. R. A programmable Cas9-serine recombinase fusion protein that operates on DNA sequences in mammalian cells. *Nucleic Acids Res* **44**, 9758-9770 (2016). https://doi.org:10.1093/nar/gkw707

18 Gaj, T., Mercer, A. C., Sirk, S. J., Smith, H. L. & Barbas, C. F., 3rd. A comprehensive approach to zinc-finger recombinase customization enables genomic targeting in human cells. *Nucleic Acids Res* **41**, 3937-3946 (2013). https://doi.org:10.1093/nar/gkt071

19 Gordley, R. M., Gersbach, C. A. & Barbas, C. F., 3rd. Synthesis of programmable integrases. *Proc Natl Acad Sci U S A* **106**, 5053-5058 (2009). https://doi.org:10.1073/pnas.0812502106

20 Mercer, A. C., Gaj, T., Fuller, R. P. & Barbas, C. F., 3rd. Chimeric TALE recombinases with programmable DNA sequence specificity. *Nucleic Acids Res* **40**, 11163-11172 (2012). https://doi.org:10.1093/nar/gks875

21 Prorocic, M. M. *et al.* Zinc-finger recombinase activities in vitro. *Nucleic Acids Res* **39**, 9316-9328 (2011). https://doi.org:10.1093/nar/gkr652

22 Standage-Beier, K. *et al.* RNA-Guided Recombinase-Cas9 Fusion Targets Genomic DNA Deletion and Integration. *CRISPR J* **2**, 209-222 (2019). https://doi.org:10.1089/crispr.2019.0013

23 Voziyanova, E., Li, F., Shah, R. & Voziyanov, Y. Genome targeting by hybrid Flp-TAL recombinases. *Sci Rep* **10**, 17479 (2020). https://doi.org:10.1038/s41598-020-74474-2

24 Mukhametzyanova, L. *et al.* Activation of recombinases at specific DNA loci by zinc-finger domain insertions. *Nat Biotechnol* (2024). https://doi.org:10.1038/s41587-023-02121-y

25 Feng, X., Bednarz, A. L. & Colloms, S. D. Precise targeted integration by a chimaeric transposase zinc-finger fusion protein. *Nucleic Acids Res* **38**, 1204-1216 (2010). https://doi.org:10.1093/nar/gkp1068

26 Owens, J. B. *et al.* Transcription activator like effector (TALE)-directed piggyBac transposition in human cells. *Nucleic Acids Res* **41**, 9197-9207 (2013). https://doi.org:10.1093/nar/gkt677

27 Owens, J. B. *et al.* Chimeric piggyBac transposases for genomic targeting in human cells. *Nucleic Acids Res* **40**, 6978-6991 (2012). https://doi.org:10.1093/nar/gks309

28 Ye, L. *et al.* TAL effectors mediate high-efficiency transposition of the piggyBac transposon in silkworm Bombyx mori L. *Sci Rep* **5**, 17172 (2015). https://doi.org:10.1038/srep17172

29 Yant, S. R., Huang, Y., Akache, B. & Kay, M. A. Site-directed transposon integration in human cells. *Nucleic Acids Res* **35**, e50 (2007). https://doi.org:10.1093/nar/gkm089

30 Klompe, S. E., Vo, P. L. H., Halpin-Healy, T. S. & Sternberg, S. H. Transposon-encoded CRISPR-Cas systems direct RNA-guided DNA integration. *Nature* **571**, 219-225 (2019). https://doi.org:10.1038/s41586-019-1323-z

31 Strecker, J. *et al.* RNA-guided DNA insertion with CRISPR-associated transposases. *Science* **365**, 48-53 (2019). https://doi.org:10.1126/science.aax9181

32 Durrant, M. G. *et al.* Bridge RNAs direct modular and programmable recombination of target and donor DNA. *bioRxiv*, 2024.2001.2024.577089 (2024). https://doi.org:10.1101/2024.01.24.577089

33 Lampe, G. D. *et al.* Targeted DNA integration in human cells without double-strand breaks using CRISPR-associated transposases. *Nat Biotechnol* **42**, 87-98 (2024). https://doi.org:10.1038/s41587-023-01748-1

34 Merrick, C. A., Zhao, J. & Rosser, S. J. Serine Integrases: Advancing Synthetic Biology. *ACS Synth Biol* **7**, 299-310 (2018). https://doi.org:10.1021/acssynbio.7b00308

35 Bessen, J. L. *et al.* High-resolution specificity profiling and off-target prediction for site-specific DNA recombinases. *Nat Commun* **10**, 1937 (2019). https://doi.org:10.1038/s41467-019-09987-0

36 Anzalone, A. V. *et al.* Programmable deletion, replacement, integration and inversion of large DNA sequences with twin prime editing. *Nat Biotechnol* **40**, 731-740 (2022). https://doi.org:10.1038/s41587-021-01133-w

37 Yarnall, M. T. N. *et al.* Drag-and-drop genome insertion of large sequences without double-strand DNA cleavage using CRISPR-directed integrases. *Nat Biotechnol* **41**, 500-512 (2023). https://doi.org:10.1038/s41587-022-01527-4

38 Xu, Z. *et al.* 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 (2013). https://doi.org:10.1186/1472-6750-13-87

39 Rebar, E. J. & Pabo, C. O. Zinc finger phage: affinity selection of fingers with new DNA-binding specificities. *Science* **263**, 671-673 (1994). https://doi.org:10.1126/science.8303274

40 Greisman, H. A. & Pabo, C. O. A general strategy for selecting high-affinity zinc finger proteins for diverse DNA target sites. *Science* **275**, 657-661 (1997). https://doi.org:10.1126/science.275.5300.657

41 Karpinski, J. *et al.* Directed evolution of a recombinase that excises the provirus of most HIV-1 primary isolates with high specificity. *Nat Biotechnol* **34**, 401-409 (2016). https://doi.org:10.1038/nbt.3467

42 Lansing, F. *et al.* Correction of a Factor VIII genomic inversion with designer-recombinases. *Nat Commun* **13**, 422 (2022). https://doi.org:10.1038/s41467-022-28080-7

43 Rutherford, K., Yuan, P., Perry, K., Sharp, R. & Van Duyne, G. D. Attachment site recognition and regulation of directionality by the serine integrases. *Nucleic Acids Res* **41**, 8341-8356 (2013). https://doi.org:10.1093/nar/gkt580

44 Li, H., Sharp, R., Rutherford, K., Gupta, K. & Van Duyne, G. D. Serine Integrase attP Binding and Specificity. *J Mol Biol* **430**, 4401-4418 (2018). https://doi.org:10.1016/j.jmb.2018.09.007

45 Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871-876 (2021). https://doi.org:10.1126/science.abj8754

46 Gersbach, C. A., Gaj, T., Gordley, R. M. & Barbas, C. F., 3rd. Directed evolution of recombinase specificity by split gene reassembly. *Nucleic Acids Res* **38**, 4198-4206 (2010). https://doi.org:10.1093/nar/gkq125

47 Pavletich, N. P. & Pabo, C. O. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science* **252**, 809-817 (1991). https://doi.org:10.1126/science.2028256

48 Zheng, Z. *et al.* Anchored multiplex PCR for targeted next-generation sequencing. *Nat Med* **20**, 1479-1484 (2014). https://doi.org:10.1038/nm.3729

49 Kotin, R. M., Linden, R. M. & Berns, K. I. Characterization of a preferred site on human chromosome 19q for integration of adeno-associated virus DNA by non-homologous recombination. *EMBO J* **11**, 5071-5078 (1992). https://doi.org:10.1002/j.1460-2075.1992.tb05614.x

50 Wang, X. *et al.* Bxb1 integrase serves as a highly efficient DNA recombinase in rapid metabolite pathway assembly. *Acta Biochim Biophys Sin (Shanghai)* **49**, 44-50 (2017). https://doi.org:10.1093/abbs/gmw115

51 Thomson, J. G. *et al.* The Bxb1 recombination system demonstrates heritable transmission of site-specific excision in Arabidopsis. *BMC Biotechnol* **12**, 9 (2012). https://doi.org:10.1186/1472-6750-12-9

52 Blechl, A., Lin, J., Shao, M., Thilmony, R. & Thomson, J. The Bxb1 Recombinase Mediates Site-Specific Deletion in Transgenic Wheat. *Plant Molecular Biology Reporter* **30**, 1357-1366 (2012). https://doi.org:10.1007/s11105-012-0454-2

53 Jiang, L. *et al.* Target lines for recombinase-mediated gene stacking in soybean. *Theor Appl Genet* **135**, 1163-1175 (2022). https://doi.org:10.1007/s00122-021-04015-6

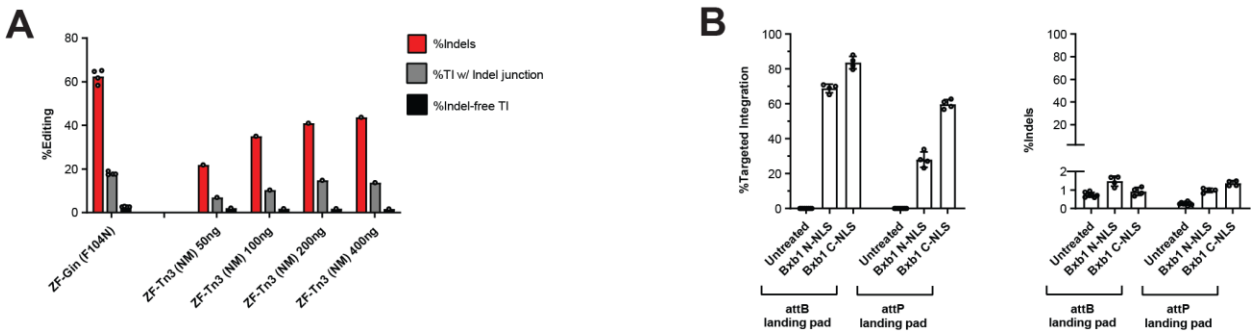
54 Inniss, M. C. *et al.* A novel Bxb1 integrase RMCE system for high fidelity site-specific integration of mAb expression cassette in CHO Cells. *Biotechnol Bioeng* **114**, 1837-1846 (2017). https://doi.org:10.1002/bit.26268

55 Durrant, M. G. *et al.* Systematic discovery of recombinases for efficient integration of large DNA sequences into the human genome. *Nat Biotechnol* **41**, 488-499 (2023). https://doi.org:10.1038/s41587-022-01494-w

56 McCarty, D. M. *et al.* Adeno-associated virus terminal repeat (TR) mutant generates self-complementary vectors to overcome the rate-limiting step to transduction in vivo. *Gene Ther* **10**, 2112-2118 (2003). https://doi.org:10.1038/sj.gt.3302134

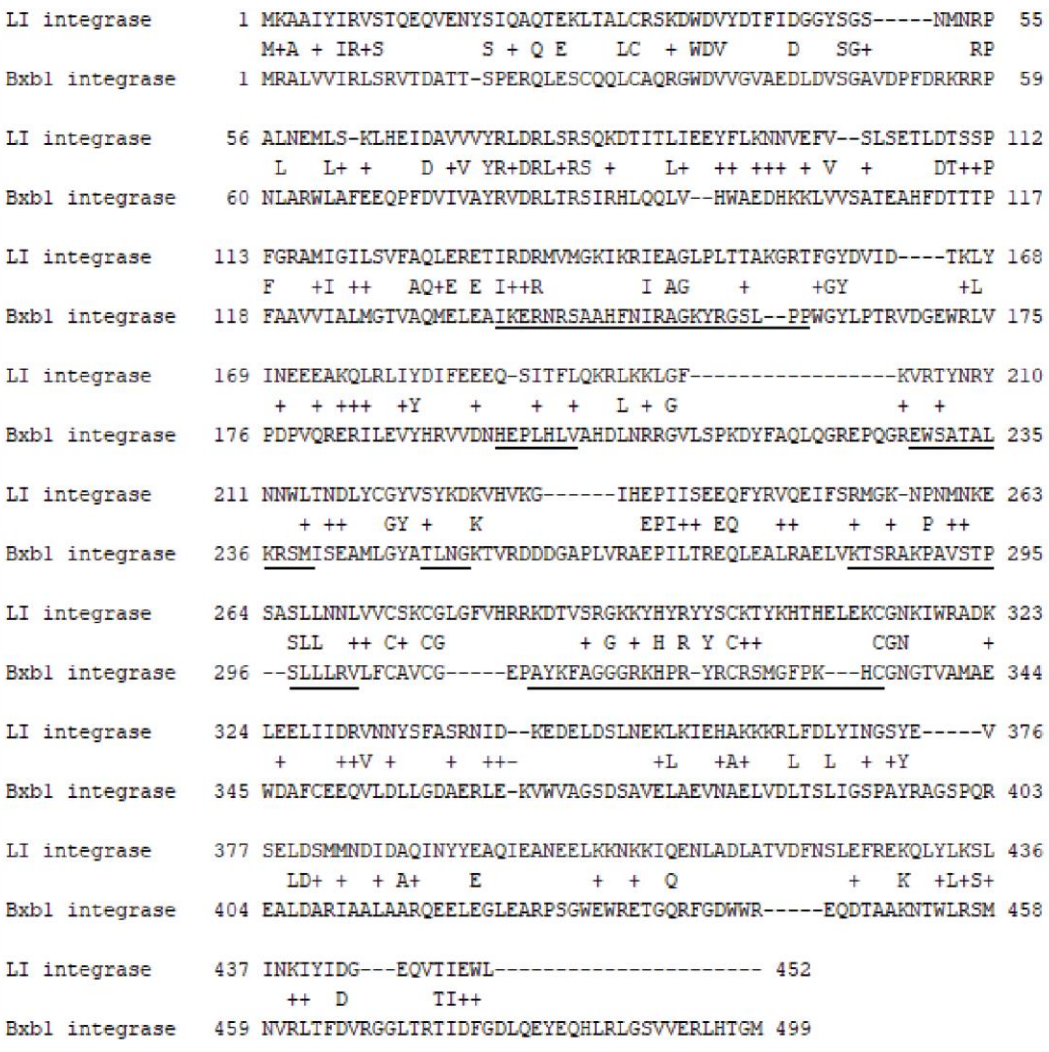
57 Tsai, S. Q. *et al.* GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat Biotechnol* **33**, 187-197 (2015). https://doi.org:10.1038/nbt.3117

58 Tareen, A. & Kinney, J. B. Logomaker: beautiful sequence logos in Python. *Bioinformatics* **36**, 2272-2274 (2020). https://doi.org:10.1093/bioinformatics/btz921

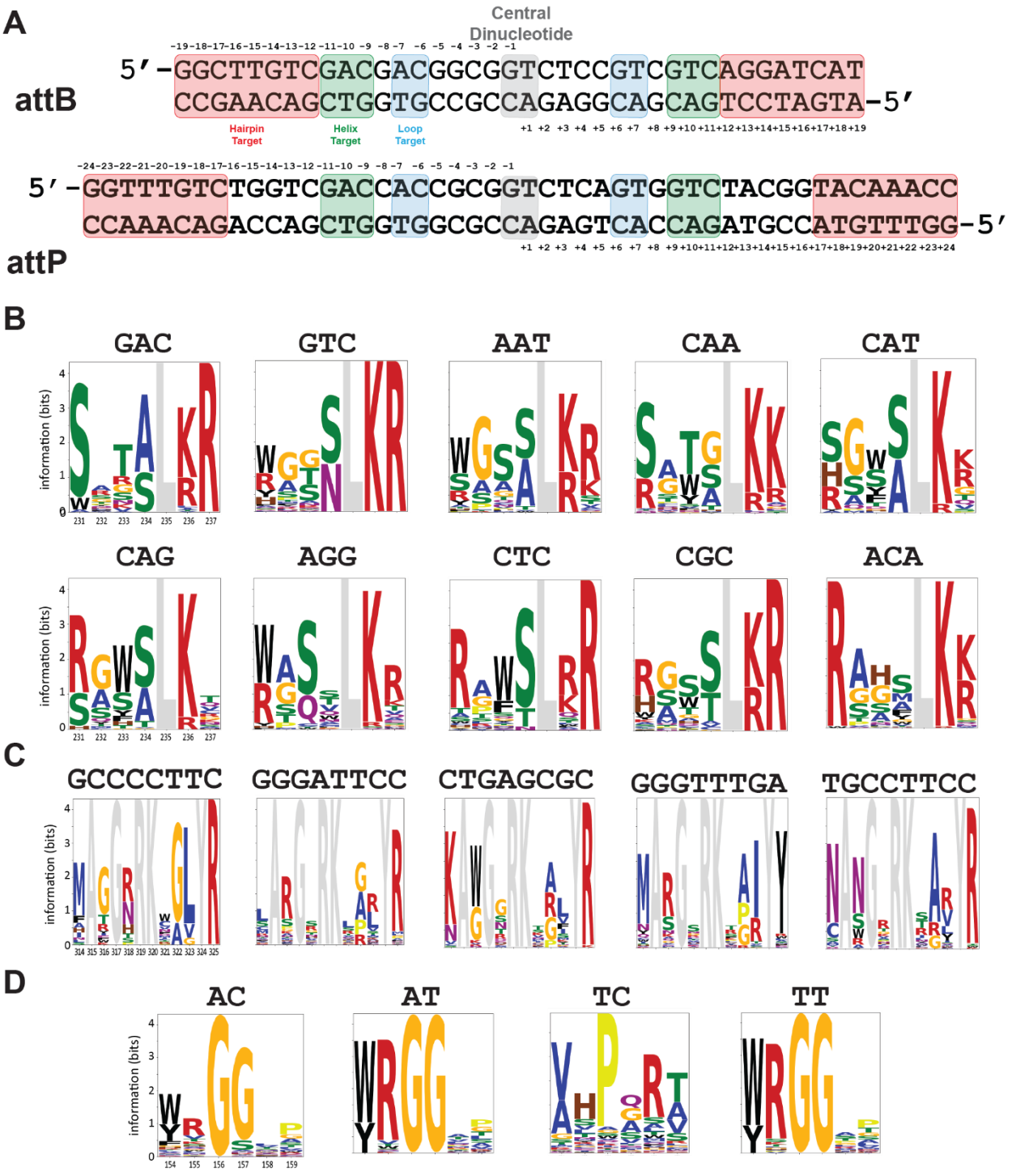


969

970 **Extended Data Figure 1. Performance of Zinc Finger Serine Recombinase fusion proteins**
971 **and wild-type Bxb1. a.** Performance of Zinc Finger (ZF)-targeted Serine Recombinases Gin and
972 Tn3. The data shown in this graph is derived from separate experiments and is representative for
973 similar work that has been performed over a period of more than two years. In contrast to Bxb1,
974 both Gin and Tn3 fusion proteins can result in high levels of indels causing low product purity and
975 hindering further improvement of target integration frequencies. We also observed indels within
976 the assayed TI junction site. Data is derived from a PCR-based NGS. **b.** Performance of wild-type
977 Bxb1 against natural attB and attP target sequences in human cells. We first established K562
978 landing pad cell lines by installing the natural attB or attP sequence in the human AAVS1 locus.
979 We noticed improved performance of Bxb1 with a C-terminal NLS compared the construct with a
980 N-terminal NLS. This guided future Bxb1 designs where all evolved variants presented in this
981 study have a C-terminal NLS. We also noticed higher targeted integration into the attB landing
982 pad. Notably, no or only minimal levels of indels were observed within the landing pad target
983 sequences. Data is derived from a PCR-based NGS.



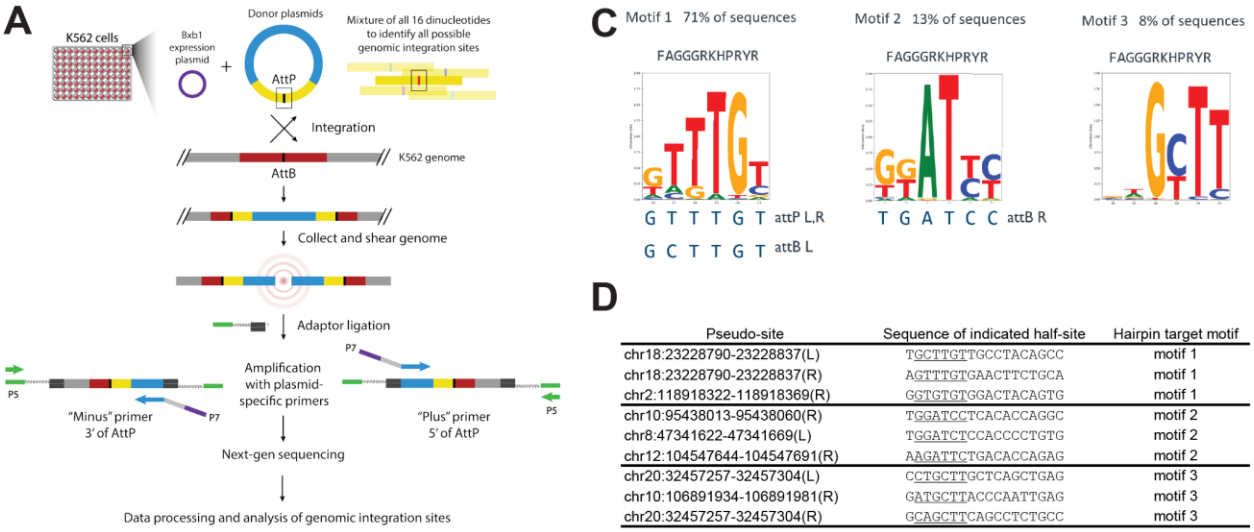
Extended Data Figure 2. Sequence alignment of the large serine recombinase from the Listeria innocua prophage and Bxb1. The alignment was modified to reflect predicted secondary structures. The region of Bxb1 that was probed with saturation mutagenesis is underlined.



988

989 **Extended Data Figure 3. a.** Schematic of attB and attP target sites with each base numbered and
990 regions varied for Bxb1 hairpin, helix, and loop selections annotated. **b.** Example sequence logo
991 plots summarizing enriched peptide motifs at amino acid residue positions 231-237 for helix
992 selections with the corresponding helix DNA targets for each selection shown above each sequence
993 logo. Position 235 was not randomized during selection and is shown in grey. Helices with an
994 arginine (R) at position 237 were frequently observed in selections for a C at position -9 of the

DNA target, helices with a lysine (K) at position 237 were often observed in selections for a A at position -9 of the DNA target, helices with an asparagine (N) at position 234 were often observed in selections for a T at position -10 of the DNA target, helices with an alanine (A) or glycine (G) at position 234 were often observed in selections for A at position -10 of the DNA target, and a tryptophan (W) was often observed in selections with a C at position -11 of the DNA target. These correlations mimic interactions observed with engineered zinc fingers if one considers position 233 to correspond to +2 of the zinc finger recognition helix, 234 to correspond to +3 of the zinc finger recognition helix, and 237 to correspond to +6 of the zinc finger recognition helix and if one considers the zinc finger DNA triplet to be the reverse complement of the Bxb1 helix target triplet. For example, from Ichikawa *et al.* (2023), the zinc finger helix QSGTLR**R** can target **G**CA, TKAYLL**K** can target **T**GA, QSS**N**LRT can target A**A**A, DPS**A**LIR can target A**T**C, and RK**W**TLQQ can target AA**G.** This implies some similarities between how the Bxb1 helix and how the zinc finger recognition helix interacts with target DNA. **c.** Example logo plots summarizing the results of selections at amino acid positions 314-325 against different hairpin DNA targets. Residues that were not randomized are shown in gray. **d.** Example logo plots summarizing the results of selections at amino acid positions 154-159 against different loop DNA targets.



D

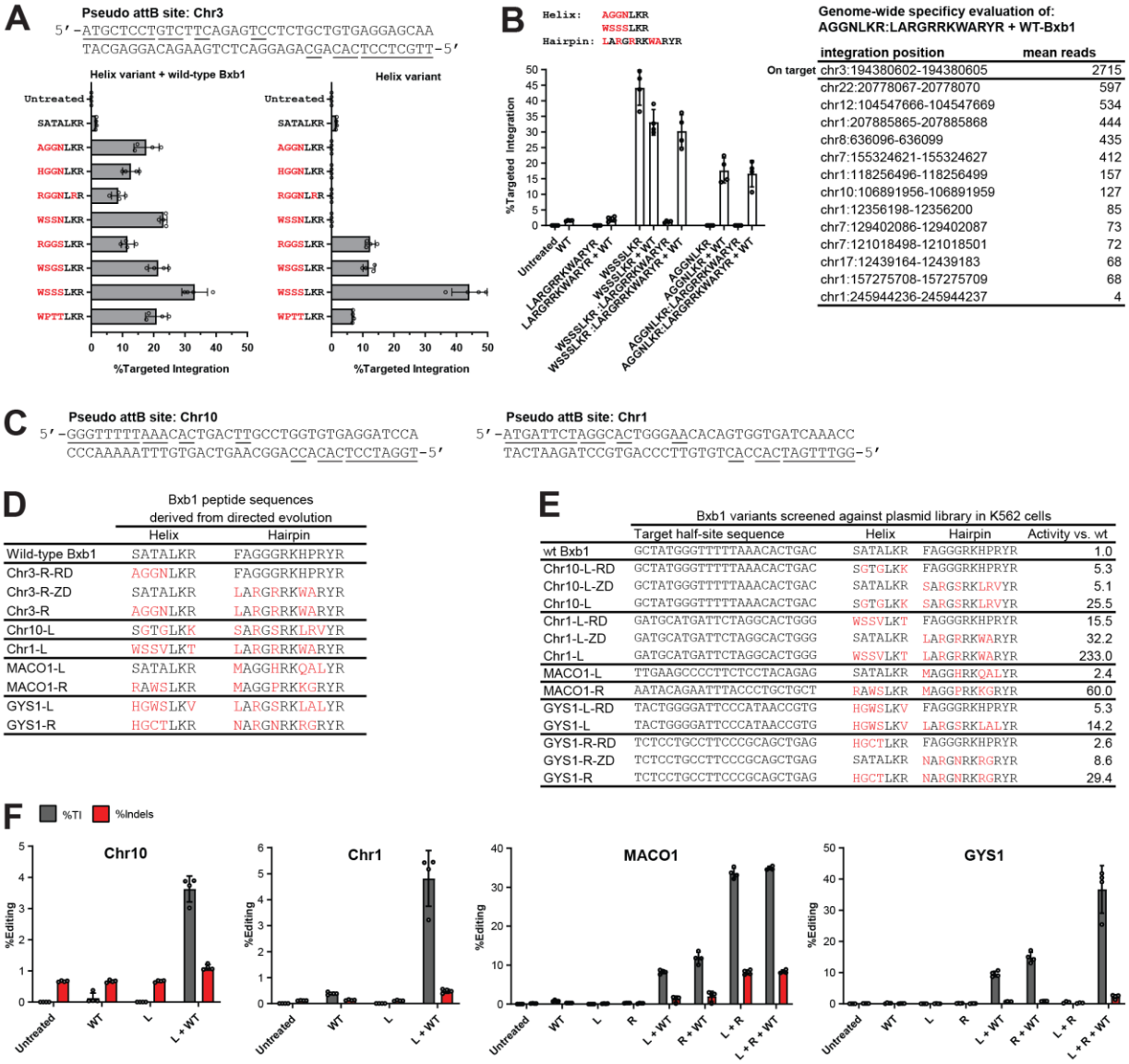
Pseudo-site	Sequence of indicated half-site	Hairpin target motif
chr18:23228790-23228837(L)	TGCTTGTTGCCTACAGCC	motif 1
chr18:23228790-23228837(R)	AGTTTGTGAACTTCTGCA	motif 1
chr2:118918322-118918369(R)	GGTGTGTGGACTACAGTG	motif 1
chr10:95438013-95438060(R)	TGGATCCTCACACCAGGC	motif 2
chr8:47341622-47341669(L)	TGGATCTCCACCCCTGTG	motif 2
chr12:104547644-104547691(R)	AAGATTCTGACACCAGAG	motif 2
chr20:32457257-32457304(L)	CCTGCTTGCTCAGCTGAG	motif 3
chr10:106891934-106891981(R)	GATGCTTACCCAATTGAG	motif 3
chr20:32457257-32457304(R)	GCAGCTTCAGCCTCTGCC	motif 3

B

Genomic coordinates	Name in ED Fig. 5	Target sequence	%TI	Source
chr12:104547644-104547691		GTGTATCTGACACCAGCACACTCTGGTGTCAGAATCTT	6.92	Experimental
chr10:106891934-106891981		CCTTCTTCTACCACAGCGTTCTCAATTGGGTAAGCATC	6.27	Experimental
chr3:194380580-194380627	Chr3	TTGCTCCTCACAGCAGAGGACTCTGAAGACAGGAGCAT	2.45	Computational
chr20:32457257-32457304		CCTGCTTGCTCAGCTGAGACGGCAGAGGCTGAAGCTGC	2.11	Experimental
chr1:25477416-25477463	MACO1	GCCCCTTCTCCTACAGAGCAAGCAGCAGGGTAAATTCT	1.59	Experimental
chr3:157294086-157294133		TGGTACTATACTACAGAGAATGCTGATGTTAAAGAGCT	1.49	Experimental
chr2:118918322-118918369		CTGATACTTGCAACTGAACACACTGTAGTCCACACACC	1.29	Computational
chr8:47341622-47341669		TGGATCTCCACCCCTGTGTTTGCTGTTGTCTAAGTTTT	1.13	Experimental
chr6:87076238-87076285		CCAGCCTGGGCAACAGAGCACTCTGTTGTCAAAAAAAA	1.09	Experimental
chr5:179235153-179235200		TTCTTTTAAACCACTGATCATGCAGAAGTATAAGCTCA	0.85	Experimental
chr1:12564789-12564836		TTGCTTCTCACCGTAGATGACCCTCTTGTCCACAAACC	0.64	Computational
chrX:50461699-50461746		GTTTTGTTCACTGTAGAATCCTCAGTGGTTAGCACAAT	0.49	Computational
chr22:20778045-20778092		TTGCTCCTGACAGCAGTGGGAGCTATTGTCTAAGAGAT	0.43	Experimental
chr7:129402064-129402111		GAAACTTACAACACAGAGTTTCCAGTTGTTCAATCTCT	0.39	Experimental
chr3:37249472-37249519		CAATCTTGGCTCACTGCAACCTCTGCAGTTGAAGCAAT	0.36	Experimental
chr18:33954336-33954383		CAGATCTATACGGCTGACTACACAGTGGTGAGAACCAT	0.34	Computational
chr1:114204100-114204147	Chr1	ATGATTCTAGGCACTGGGAACACAGTGGTGATCAAACC	0.32	Computational
chr5:179764006-179764053		TGAGTCCTGACTGCAGCACATGCAGGTGAGGAAGCAGT	0.26	Experimental
chr19:48970994-48971041	GYS1	GGGATTCCCATAACCGTGCACTCAGCTGCGGGAAGGCA	0.26	Experimental
chr1:207885846-207885893		TGTTTGGCCCCAACTGCCTCTGCCACTGTCGACACACT	0.25	Computational
chrX:103694808-10369485		ATGTTCCTCACCACAGCTAACCCAGTCTTCAAAACTCA	0.25	Experimental
chr10:95438013-95438060	Chr10	GGGTTTTTAAACACTGACTTGCCTGGTGTGAGGATCCA	0.20	Computational
chr11:58774490-58774537		AGAAATAGTACAACAGCTAACGCAGATGTTAAAAATCA	0.12	Experimental
chr18:23228790-23228837		TGCTTGTTGCCTACAGCCTCTGCAGAAGTTCACAAACT	0.04	Computational
chr1:203596841-203596888		CTCTTGATGACTGCAGAGTATTCCATTGTTGACAAATC	0.04	Computational
chr4:78331572-78331619		GGGTTGGTGGCAATGGAGATCTCAGTGGTGAAAAATCA	0.03	Computational

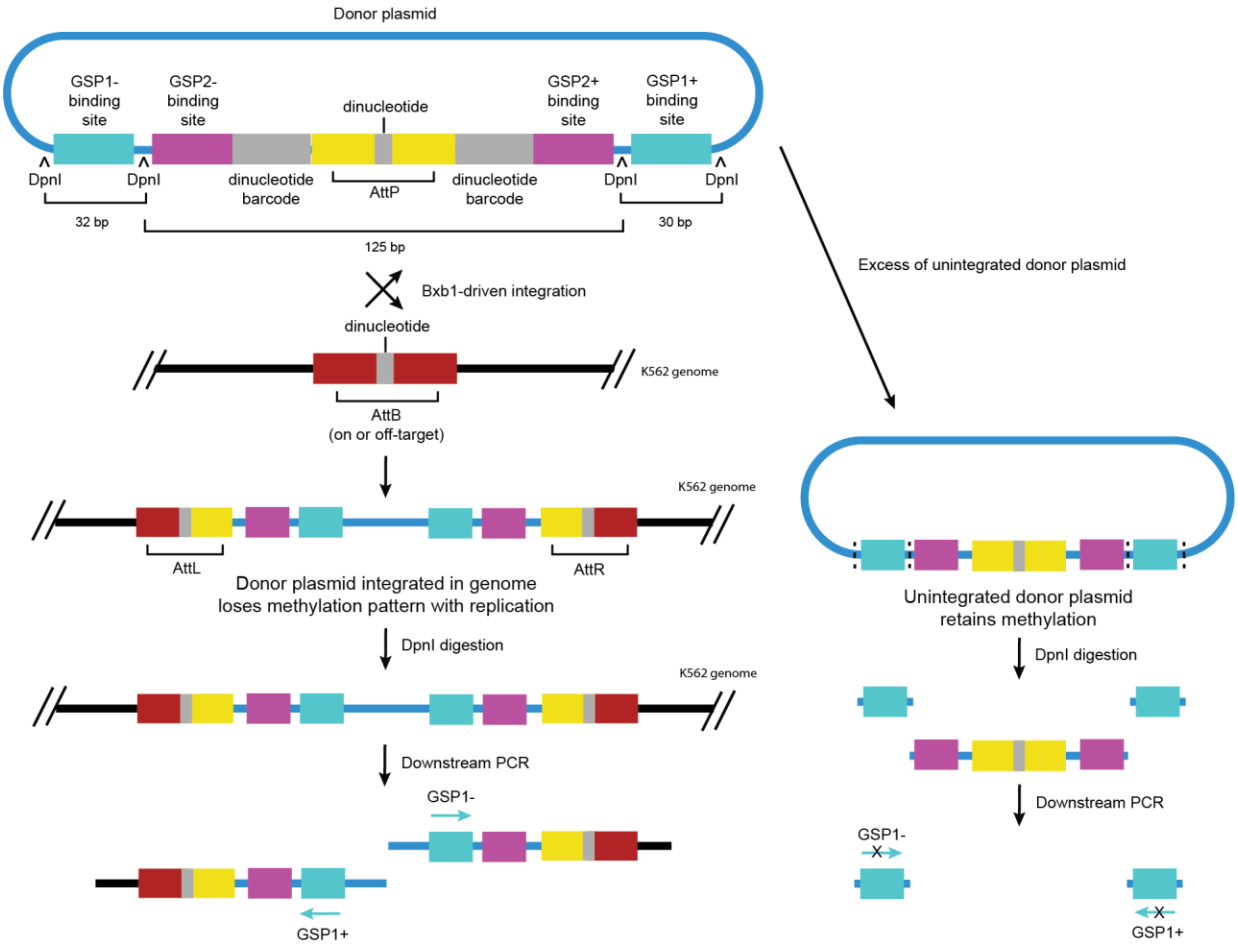
Extended Data Figure 4. Identification of Bxb1 pseudo-sites in the human genome using an unbiased genome-wide specificity assay. a. Schematic of the unbiased-genome specificity assay used in this study to experimentally identify pseudo-sites for wild-type Bxb1 in the human genome. **b.** List of 26 validated Bxb1 pseudo-sites in the human genome that were either identified through a computational search or experimentally using the assay shown in panel **a**. The central dinucleotides in each site are underlined. Only the portion corresponding to an attB site are shown in the table. Site-specific donors were used to validate targeted integration using a PCR-based NGS assay. The 23 sites above the line have more than 0.1% TI **c.** Additional analysis of the Bxb1 hairpin specificity data for the wild-type Bxb1 hairpin in **Figure 2e** indicates the selected sequences belong to at least three distinct DNA motifs. Plots of this data split into three separate motifs are shown. The first motif is consistent with positions -18 to -13 of the hairpin targets in both left and right half-sites of the natural attP site and the left half-site of the natural attB site. The second motif is consistent with the right half-site of the natural attB site. **d.** Examples of half-sites

from validated human pseudo-sites for wild-type Bxb1 that correspond to each of the three DNA sequence motifs are shown in panel **c**. Positions -18 to -13 of each half-site are underlined.

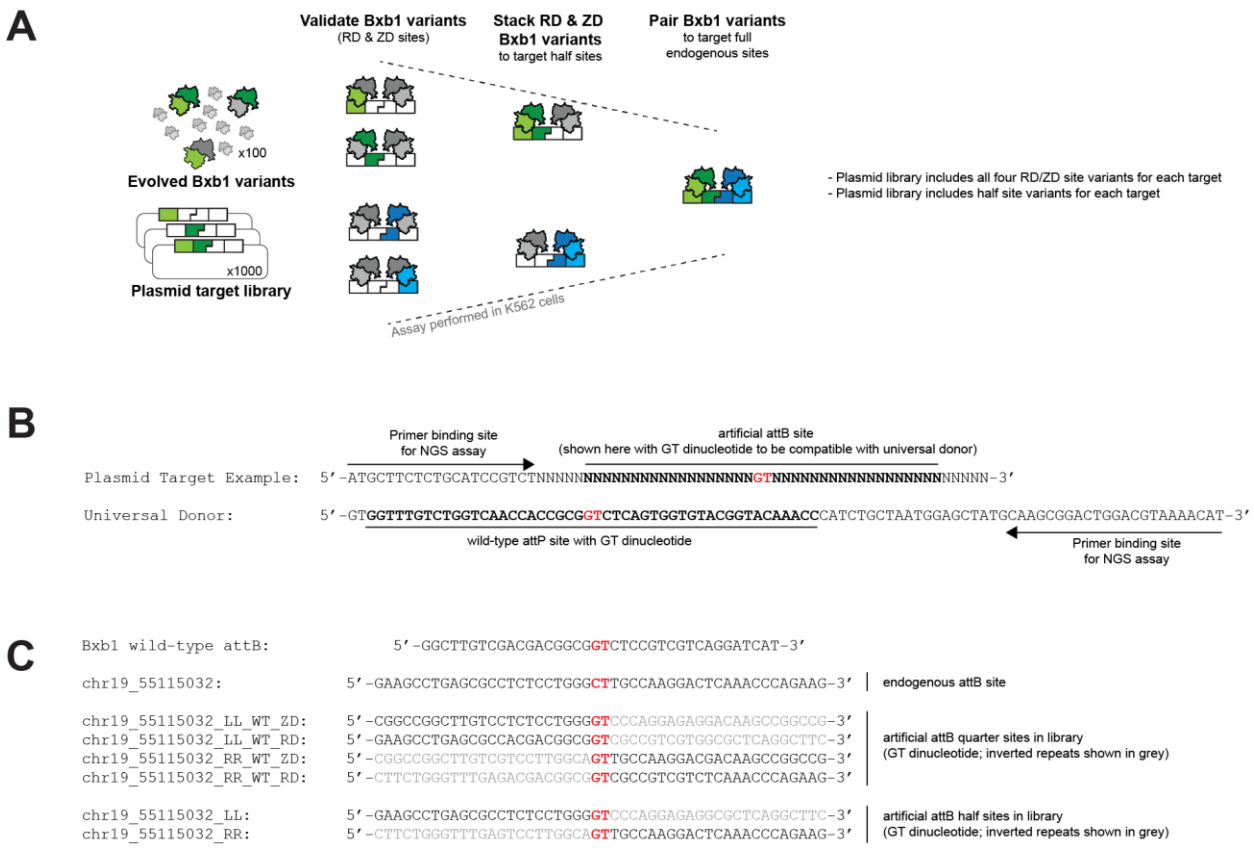


Extended Data Figure 5. Performance of Bxb1 variants against pseudo-sites in the human genome. a. Depicted is an attB pseudo-site on chromosome 3 and performance measurements of various Bxb1 helix domain variants against this target site using a PCR-based NGS assay. SATALKR is the wild-type helix peptide sequence and the target DNA sequence is shown as the reverse complement of the sequence in Extended Data Figure 4b to make it easier to visualize the region we are targeting in this experiment. **b.** On-target performance of a Bxb1 variant with combined helix and hairpin variations, and data summary of a genome-wide specificity evaluation using a modified version of the assay described in Extended Data Fig. 4 (see Extended Data Fig. 6). Note that the indicated Bxb1 variant has to be mixed with wild-type Bxb1 in order to be active at the intended chr3 target site so the genome-wide specificity assay was performed with a mixture of the indicated Bxb1 variant and wild-type Bxb1. The top 3 off-targets as well as two others are also targets for wild-type Bxb1 (Extended Data Figure 4b) and are likely caused by the presence of wild-type Bxb1 in the experiment. The experiment was also performed with a pool of donor constructs containing all 16 possible central dinucleotide sequences. The intended target on chromosome 3 is the only target on this list with a GA or TC central dinucleotide and thus the other

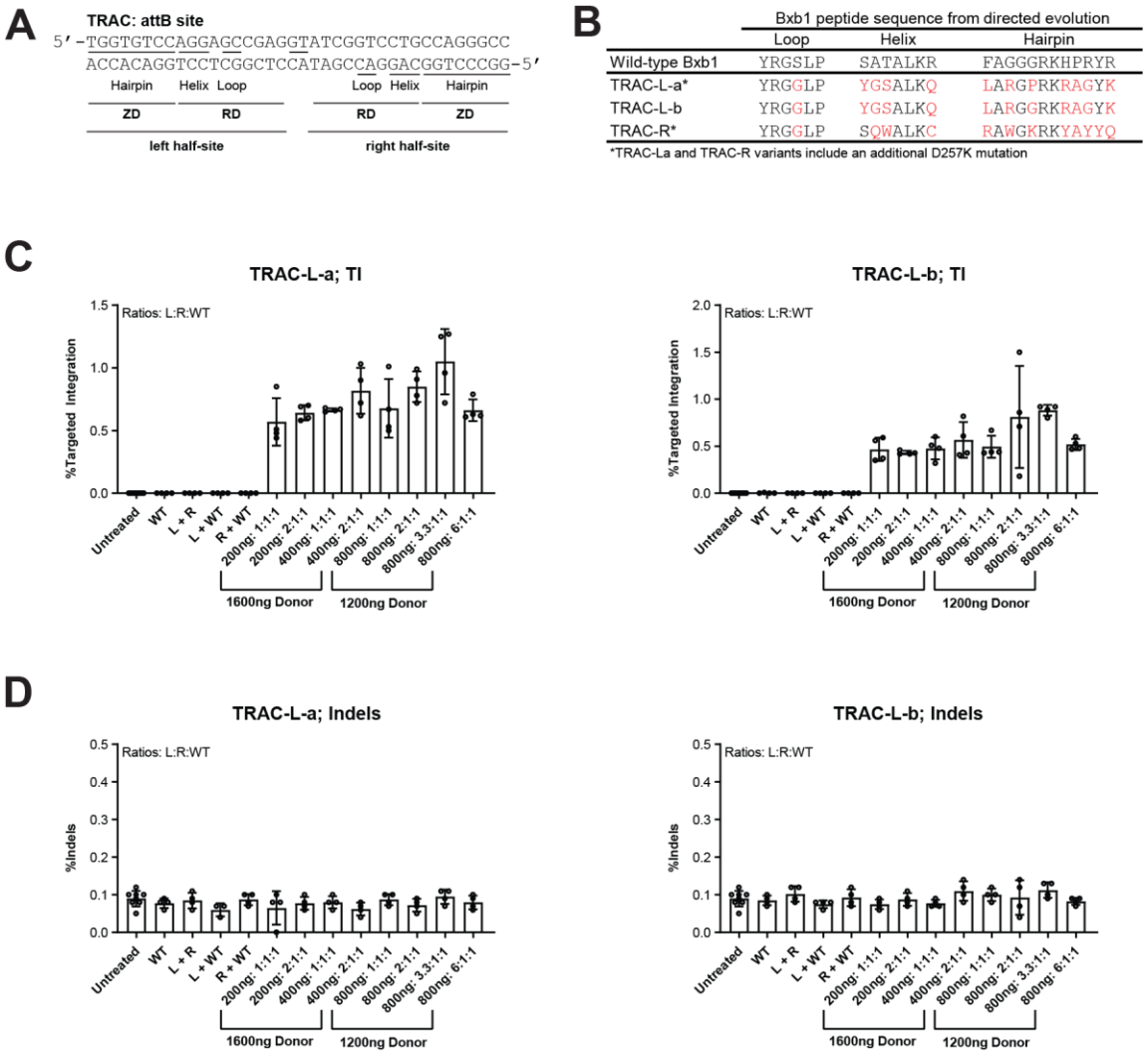
sites presumably wouldn't have been detected if only a single donor with a GA or TC dinucleotide had been used in the genome-wide specificity experiment. **c.** Sequence of four additional Bxb1 pseudo-sites in the human genome. **d.** Bxb1 peptide sequences of evolved Bxb1 variants that showed improved performance against the half-sites of pseudo-sites shown in **Figure 3b** and panel c. **e.** Screening data using synthetic DNA targets tested in human K562 cells that was used to identify the constructs shown in panel **d**. Activity is determined by the number of DNA sequence reads corresponding to recombined versions of each synthetic target; activity is normalized to the activity of wild-type Bxb1 alone against the same synthetic target site. **f.** Results from a PCR-based NGS assay demonstrating improved performance of evolved Bxb1 variants against their chromosomal endogenous targets in K562 cells. The presence of a wild-type Bxb1 expression construct is necessary to bind the wild-type attP sequence on the donor plasmid.



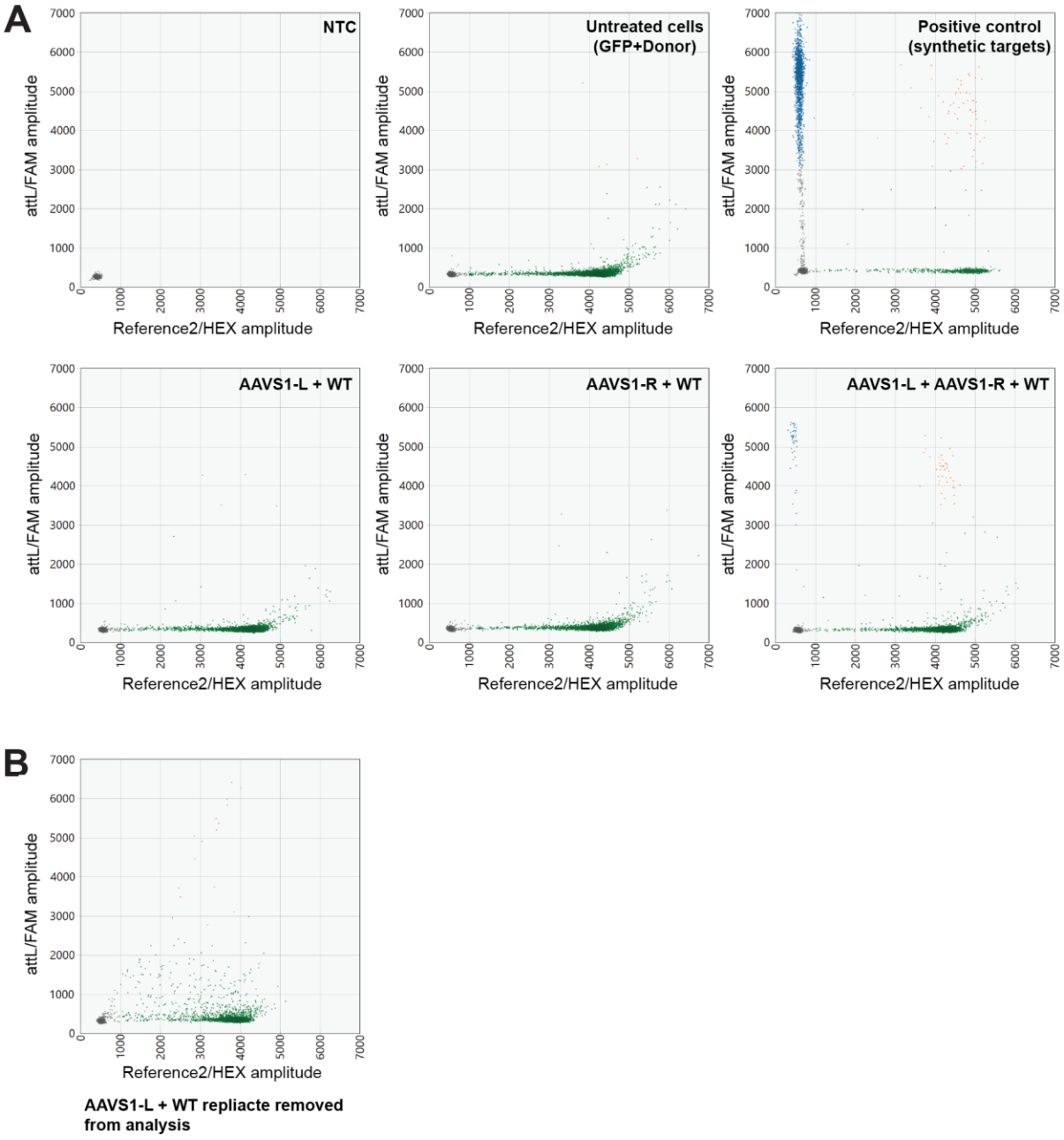
Extended Data Figure 6. Improved genome-wide specificity assay. Schematic of the modified unbiased-genome specificity assay used in this study to experimentally identify integration site shown in Extended Data Figure 5b. Strategically placing DpnI recognition sites in the donor molecule supports the enzymatic removal of excess unintegrated donor plasmid, resulting in a substantial reduction of donor plasmid-derived background signal.



Extended Data Figure 7. Screening system for testing Bxb1 variants against a plasmid library of artificial target sites in human K562 cells. a. Schematic of a system for testing evolved Bxb1 variants against thousands of artificial targets in a plasmid target library as described in **Figure 3a.** **b.** Sequence examples for the target library and the donor, including assay design information for PCR-based detection of recombination events. **c.** Sequence example for all plasmid target library members of the AAVS1 site highlighted in **Figure 4**. The plasmid target library includes up to six distinct members for each endogenous target site. Four members are designed to support the identification of Bxb1 variants targeting the RD and ZD motifs of both endogenous half-sites. Two additional members are designed to confirm activity of stacked Bxb1 variants against the corresponding half-sites. Bxb1 variants are screened individually against the plasmid target library.



Extended Data Figure 8. Bxb1 retargeting to the human TRAC locus. a. Depicted is an attB-like site in the human TRAC locus. **b.** Bxb1 peptide sequences of evolved Bxb1 variants that showed improved performance against the half-sites of the TRAC site shown in panel a. **c.** Results from a PCR-based NGS assay demonstrating targeted integration mediated by Bxb1 variants from panel b at the TRAC site shown in panel a. A dose titration was performed where the total dose was kept at 2000ng pDNA (Bxb1 expression constructs and donor combined). **d.** Indel analysis of the experiment shown in panel c. All experiments were performed in human K562 cells.



Extended Data Figure 9. ddPCR analysis of Bxb1 variants targeted to the human AAVS1 locus. a. Example 2-D ddPCR scatterplots of differently treated samples. Each 2-D plot contains 1-4 clusters of droplets: (1) double-negative droplets containing no targeted DNA templates (grey dots clustered at the left bottom in each plot); (2) reference-only droplets (green); (3) target integration/attL-only droplets (blue); and (4) double-positive droplets containing both target integration/attL and reference DNA templates (orange). **b.** The 2-D plot of an AAVS1-L + WT replicate discarded from the analysis with noisy FAM signal likely due to shredded droplets**.**

Supporting Sequence Information

Bxb1 C-NLS

MRALVVIRLSRVTDATTSPERQLESCQQLCAQRGWDVVGVAEDLDVSGAVDPFDRKRRPNLARWLAFEEQPFDVIVAYRVDRLTRSIRHLQQLVHWAEDHKKLVVS
ATEAHFDTTTPFAAVVIALMGTVAQMELEAIKERNRSAAHFNIRAGKYRGSLPPWGYLPTRVDGEWRLVPDPVQRERILEVYHRVVDNHEPLHLVAHDLNRRGVLS
PKDYFAQLQGREPQGREWSATALKRSMISEAMLGYATLNGKTVRDDDGAPLVRAEPILTREQLEALRAELVKTSRAKPAVSTPSLLLRVLFCAVCGEPAYKFAGGG
RKHPRYRCRSMGFPKHCGNGTVAMAEWDAFCEEQVLDLLGDAERLEKVWVAGSDSAVELAEVNAELVDLTSLIGSPAYRAGSPQREALDARIAALAARQEELEGLE
ARPSGWEWRETGQRFGDWWREQDTAAKNTWLRSMNVRLTFDVRGGLTRTIDFGDLQEYEQHLRLGSVVERLHTGMSGSGSGSHHHHHHGSGPKKKRKV**

MACO1

Donor fragment GGTTTGTCTGGTCAACCACCGCGCACTCAGTGGTGTACGGTACAAACCAAGCAGCTCGGAGCCTCTCTGTCACTTGCTCTTTAGA

NGS fw primer ACACGACGCTCTTCCGATCTNNNNCAATTAGTTGGCTGTATAATTTGG

NGS rev primer GACGTGTGCTCTTCCGATCTTCTAAAGAGCAAGTGACAGAGAGG

GYS1

Donor fragment GGTTTGTCTGGTCAACCACCGCGCACTCAGTGGTGTACGGTACAAACCAGCCCGAACTTAGCTTCCCTCATCGCCTA

NGS fw primer ACACGACGCTCTTCCGATCTNNNNCGATGTGTTCCTCCATGAAGCA

NGS rev primer GACGTGTGCTCTTCCGATCTTAGGCGATGAGGGAAGCTAAGT

Chr3

Donor fragment GGTTTGTCTGGTCAACCACCGCGGACTCAGTGGTGTACGGTACAAACCCTGTGACTACATTTAGTGAGCAGGTGGAATGAACAA

NGS fw primer ACACGACGCTCTTCCGATCTNNNNAGCCATTTCCTTCCTAGCAAATT

NGS rev primer GACGTGTGCTCTTCCGATCTTTGTTCATTCCACCTGCTCACT

Chr1

Donor fragment GGTTTGTCTGGTCAACCACCGCGAACTCAGTGGTGTACGGTACAAACCAGTCACCAATCTGGAGGGGATGAGGTGAGGGAGGGATAGACAATA

NGS fw primer ACACGACGCTCTTCCGATCTNNNNCAGCAGTCGATGTGGGAAC

NGS rev primer GACGTGTGCTCTTCCGATCTTATTGTCTATCCCTCCCTCACCT

Chr10

Donor fragment GGTTTGTCTGGTCAACCACCGCGTTCTCAGTGGTGTACGGTACAAACCAGCATCTTTCAAATAGCACCTCATTTTATCCTGAAGACCCAG

NGS fw primer ACACGACGCTCTTCCGATCTNNNNAGCCCAGAGTTAACCAAGCTAC

NGS rev primer GACGTGTGCTCTTCCGATCTCTGGGTCTTCAGGATAAAATGAGG

TRAC

Donor fragment GGTTTGTCTGGTCAACCACCGCGGTCTCAGTGGTGTACGGTACAAACCTCCCGTCATCCAGGTGCTCATATGCTGTAAGTTCCC

NGS fw primer ACACGACGCTCTTCCGATCTNNNNGCCAGGGTCATGCAACATGTAC

NGS rev primer GACGTGTGCTCTTCCGATCTGGGAACTTACAGCATATGAGC

AAVS1

Donor fragment GGTTTGTCTGGTCAACCACCGCGCTCTCAGTGGTGTACGGTACAAACCGAAGCAGCCCGGCCTTAGGGAAGCGGGACCCTGCTCTGGGCGGAGGAATATGTC

NGS fw primer ACACGACGCTCTTCCGATCTNNNNGCATGAGATGGTGGACGAGGA

NGS rev primer GACGTGTGCTCTTCCGATCTGACATATTCCTCCGCCCAGAG

1087