Reprogrammed Serine Integrases Enable Precise Integration of Synthetic DNA

Friedrich Fauser[1,2,3], Sebastian Arangundy-Franklin[1,2], Jessica E Davis[1,2], Lifeng Liu[1,2], Nicola J Schmidt[1], Luis Rodriguez[1], Danny F Xia[1], Nga Nguyen[1], Nicholas A Scarlott[1], Rakshaa Mureli[1], Irene Tan[1], Yuanyue Zhou[1], Lynn N Truong, Sarah J Hinkley[1], Bhakti N Kadam[1], Stephen Lam[1], Bryan Bourgeois[1], Emily Tait[1], Mohammad Qasim[1], Vishvesha Vaidya[1], Adeline Chen[1], Andrew Nguyen[1], Yuri R. Bendaña[1], David A. Shivak[1], Patrick Li[1], Andreas Reik[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 genomic site that is different from the natural recognition site of the integrase[5]. Here, we describe the development of a Modular Integrase (MINT) platform, a versatile protein-guided genome editing tool that can facilitate site-directed targeted integration (TI) 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 demonstrate the therapeutic potential of the MINT platform by retargeting Bxb1 to the human AAVS1 and TRAC loci where wild-type Bxb1 has no detectable activity. By combining the MINT platform with known activity-increasing Bxb1 mutants, we achieved 14% TI at the AAVS1 locus, and by additionally fusing zinc finger DNA binding domains to engineered Bxb1 variants, we achieved 35% TI at the TRAC locus in human K562 cells. To further demonstrate clinical potential, we achieved stable 15% TI of a functional donor construct in human T cells.

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, agriculture, synthetic biology, and basic research. Initially, there was excitement that homology-directed-repair (HDR) stimulated by engineered nucleases would be able to accomplish this goal, but numerous limitations were encountered when this was assessed, such as low efficiency caused by competing DNA repair pathways, undesired chromosomal rearrangements, and break-induced DNA damage response[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 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 has been 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[30-33] but delivering such complex systems to therapeutically relevant cells has yet to be demonstrated and other nucleic acid-guided systems have target site size limitations[34] that will make specifically targeting therapeutically relevant loci difficult.

Large serine recombinases (LSRs - also known as Lare Serine Integrases, LSIs) have long been proposed as the ideal tool for genome engineering[35,36] due to their unique properties including irreversible integration and no inherent limit to the size of the DNA constructs they can integrate (**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 in the center of each binding site which is referred to as the central dinucleotide or CDN. If the sequence of the CDN in the attP site matches the sequence of the CDN in the attB site then 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 is irreversible. 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 or natural attP target site into mammalian cells[37-39]. However, the requirement to first use prime editing to integrate the target site makes the process more complex for therapeutic development and manufacturing, requiring the simultaneous delivery of two independent genome editing systems 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[40], there are several technical challenges to overcome. 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 been used to engineer proteins such as ZFs that lack a simple DNA targeting code, but the modularity and compactness of the ZF repeat allows the engineering of one or two ZF repeats that can then be combined into a longer array[41,42]. Directed evolution can also be used on the entire coding sequence of non-modular proteins such as Cre, but this required over a hundred sequential rounds of directed evolution to achieve good performance at a desired target site[43,44]. A further challenge for reprogramming Bxb1 is that no experimentally determined structure of Bxb1 bound to its target site is currently available.

In this study, we 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 LSRs bound to target DNA 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[45]. 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 which we refer to as MINT constructs. We use this strategy to successfully target multiple

89 pseudo-sites in the human genome where wild-type Bxb1 has detectable activity as well as intron
90 1 of the therapeutically relevant human AAVS1 and TRAC loci where wild-type Bxb1 has no
91 detectable activity. We then add previously described mutations that increase the activity of wild-
92 type Bxb1 to increase integration activity of our MINT constructs that target both AAVS1 and
93 TRAC. Finally, we fuse engineered ZF arrays to our engineered LSRs that target the human TRAC
94 locus to achieve 35% TI in human K562 cells and stably integrate a functional GFP-expressing
95 donor with a frequency of 15% in primary human T cells.

96 **Bxb1-DNA interaction mapping**

97 To compensate for the lack of a known structure of Bxb1 bound to its DNA target site, we
98 performed an initial experiment designed to map interactions between residues in the Bxb1 RD
99 and ZD domains and key regions of its natural DNA target site. Based on a sequence alignment of
100 Bxb1 and the regions of the LI integrase known to interact with its DNA target[45,46] (Extended Data
101 Fig. 2), we identified 89 residues in Bxb1 that are likely to interact with DNA. We performed
102 scanning mutagenesis on these 89 residues and co-delivered the resulting Bxb1 variants along with
103 plasmids bearing variant attB and attP sites and measured recombination in human K562 cells.
104 This experiment identified two clusters of residues which, when mutated, could alter specificity at
105 portions of the DNA target site. A first cluster of residues at positions 231, 233, and 237 of Bxb1
106 was able to alter specificity at positions -10 and -9 of the DNA target site. A second cluster of
107 residues at positions 314, 315, 316, 318, 323, and 325 of Bxb1 was able to alter specificity at
108 positions -18 through -13 of the DNA target site. The residue at position 158 of Bxb1 was also able
109 to alter specificity at positions -7 and -6 of the DNA target site. Saturation mutagenesis of these
110 residues was able to achieve further specificity shifts. Mapping the Bxb1 mutations that shifted
111 targeting specificity onto a RoseTTAFold[47] structural model of Bxb1 bound to DNA indicated two
112 regions in the RD domain and one region in the ZD domain where focused protein engineering
113 should be able to alter the DNA targeting specificity. We refer to these regions as the helix
114 (positions 231-237), hairpin (positions 314-325), and loop (positions 154-159) respectively based
115 on the structure of each region in our Bxb1 structural model (**Figure 1b-d**).

116 **Evolving Bxb1-DNA interaction domains**

117 Next, we developed a directed evolution system to engineer Bxb1 target site specificity.
118 This system is based on a previously described method[48] whereby a "stuffer" sequence flanked by

119 artificial attB and attP target sites is placed within the coding sequence of an antibiotic resistance
120 gene such that recombination by a Bxb1 variant able to target the attB and attP sites restores the
121 open reading frame and allows the bacterial cell to survive an antibiotic challenge (**Figure 1e**;
122 Extended Data Fig. 3a). In our system we placed the Bxb1 variant library and the DNA target sites
123 on separate plasmids to enable more rapid testing of the same library of Bxb1 variants against
124 different DNA target sites and we utilized a different antibiotic resistance gene to improve selection
125 performance. A related system was used to assess the DNA targeting specificity of a selected Bxb1
126 variant by testing a library of different target sites against a single Bxb1 variant (Extended Data
127 Fig. 3e).

128 Our structural analysis of the LI integrase indicated that residues in the structure
129 corresponding to residues 231-237 in Bxb1 form an alpha helix that docks with its target DNA in
130 a manner reminiscent of the interaction between a zinc finger (ZF) and its target trinucleotide[49].
131 Thus, we also adopted the same residue randomization scheme used with ZFs[42] and kept residue
132 235 fixed as a leucine since this seems analogous to the leucine that is often at +4 of a ZF
133 recognition helix. This resulted in a library of Bxb1 helix variants with residues 231-234, and 236-
134 237 fully randomized (Extended Data Fig. 3a,b). Since ZFs can target 3 bp DNA sequences we
135 hypothesized that the Bxb1 helix might be able to specify the DNA bases at positions -11, -10, and
136 -9 of the target site. Thus, we performed 64 separate selections using our Bxb1 helix variant library
137 against all 64 possible DNA triplets at positions -11, -10, -9 (and mirrored at positions +9, +10,
138 +11) of the attB and attP target sites. Many selected helices demonstrated dramatic changes in
139 target specificity at these positions in the DNA target site. A comparison of the target preference
140 for the wild-type Bxb1 helix SATALKR and 19 selected helix sequences is shown in **Figure 1f**
141 and a ranked list of 366 selected helices organized by their preferred DNA triplet can be found in
142 Supplementary Tables 1,2.

143 Next, we performed selections using a library of Bxb1 variants where positions 154-159
144 (the loop region) were randomized (Extended Data Fig. 3a,c). We performed 16 separate selections
145 using this library against all possible DNA dimer sequences at positions -7 and -6 and found that,
146 surprisingly, a single residue change of S157G could shift the sequence preferences of wild-type
147 Bxb1 towards nearly every single dinucleotide at positions -7 and -6 except for CG (Extended Data
148 Fig. 3g) and we note that the same mutation was also identified by a recent effort to increase the

activity of Bxb1[39]. 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.

Finally, we used our directed evolution system to select Bxb1 variants with mutations at the hairpin region (residues 314-325). Since a fully randomized hairpin region would not be efficiently covered by a typical library generated in E. coli, we fully randomized residues 314, 316, 318, 321, 323, and 325 that partially face the DNA major groove in our model, and partially randomized residue 322 to correspond to residues conserved in many large serine recombinases (Extended Data Fig. 3a,d). Correspondingly, the DNA target sequence of the hairpin encompasses a sequence space too large to easily perform a selection against each possible target site variant so we performed selections against specific target sites where positions -19 to -12 matched potential Bxb1 target sites of interest in the human genome. Our selections returned hairpin motifs with wide sequence divergence (Extended Data Fig. 3f) and obvious target preference differences in comparison to the wild-type hairpin.

Identifying Bxb1 pseudo-sites in the human genome

As an initial test of engineered Bxb1 variants in human cells, we decided to target Bxb1 pseudo-sites in the human genome where wild-type Bxb1 already has detectable activity so that even modest improvements could be measured. At the time, no human Bxb1 pseudo-sites were known so we attempted to identify such sites in human K562 cells by performing both a computational search based on published Bxb1 specificity data[36] and a genome-wide experimental approach that mapped integrated donor constructs using a method similar to anchored multiplex PCR[50,51] (Extended Data Fig. 4a). In order to comprehensively identify pseudo-sites, we ignored the CDN sequence in the computational approach and we used a mixture of donors with all possible 16 CDNs for the experimental approach.

Combining the results of both approaches, we were able to identify 23 sites with at least 0.1% integration in human K562 cells. We chose 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, Supplementary Tables 3,4). Notably, both computational and experimental approaches identified only active attB

179 pseudo-sites, and no attP pseudo-sites, mirroring the direction of natural Bxb1-directed integration
180 into its host genome. This is consistent with our finding that "landing pad" cell lines with attB sites
181 pre-integrated into the genome support higher levels of integration than landing pad cells lines
182 with attP sites pre-integrated (Extended Data Fig. 1b).

183 **Increasing Bxb1 activity at human pseudo-sites**

184 Having identified suitable Bxb1 pseudo-sites in the human genome, we decided to test if
185 selected Bxb1 helix variants could improve activity at an intergenic pseudo-site on chromosome 3
186 (Extended Data Fig. 5a). This pseudo-site has one half-site with a T at position -10 that is expected
187 to be recognized poorly by the wild-type Bxb1 helix (**Figure 1f**). To improve activity, we
188 performed a helix selection against the relevant portion of this half-site and the selection yielded
189 multiple families of helix sequences. We tested a total of 58 selected helices from these different
190 sequence families for their ability to carry out site-specific integration in human K562 cells. Data
191 for eight helices that represent the most active variants from each sequence family with activities
192 ranging from 8% to over 30% are shown in Extended Data Fig. 5a. Note that the selected helix
193 AGGNLKR demonstrates a dramatic shift in DNA preference towards a T at position -10 (**Figure**
194 **1f**). Next, we wanted to investigate whether helix and hairpin variants, located in separate domains
195 of Bxb1, can be engineered independently and then combined to further engineer Bxb1 towards
196 this target site. We performed a selection using our Bxb1 variant hairpin library against the relevant
197 portion of the same half-site of the chromosome 3 target and identified the hairpin peptide
198 LARGRRKWARYR that can further improve performance when combined with selected helix
199 variants (Extended Data Fig. 5b). The combination of the AGGNLKR helix and
200 LARGRRKWARYR hairpin yielded a Bxb1 variant with a substantial shift in genome-wide DNA
201 target site specificity relative to wild-type Bxb1. For this variant, the intended pseudo-site in
202 chromosome 3 has the most integrations within the entire human genome (Extended Data Fig. 5b)
203 as assayed by an improved genome-wide specificity assay (Extended Data Fig. 6). Lastly, we
204 confirmed that our Bxb1 variants are also active in other cell types, such as human HepG2 cells
205 (Extended Data Fig. 5c).

206 We hypothesized that engineering separate variants for the left and right half-site of a
207 pseudo-site could yield additional improvements in performance. But we were concerned that
208 simply testing variants at the full endogenous pseudo-site would make it difficult to determine how

recognition of each half site was being effected by a given Bxb1 variant so we established a system that utilizes synthetic DNA target sites on plasmids where recombination at an inverted repeat of a given left half-site could be monitored separately from recombination at an inverted repeat of a given right half-site. This system also allows us to test the activity of Bxb1 helix, loop, or hairpin variants against hybrid half-sites where the portion of the half-site recognized by either the RD or ZD domain can be replaced with the relevant portion of the wild-type Bxb1 attB target site (we refer to these as "quarter-sites"). Helix and/or loop variants that perform well at an RD quarter-site can then be combined with hairpin variants that perform well at the corresponding ZD quarter-site in order to recognize the desired half-site (**Figure 2a**). We used this approach with two endogenous pseudo-sites identified by the computational genome scan and two endogenous pseudo-sites identified by the experimental genome-wide screen. This system allowed us to rapidly identify evolved helix and hairpin variants that, when combined, are active against their desired half-sites (**Figure 2b,c**; Extended Data Fig. 5d-f), demonstrating the modularity of our MINT platform and we refer to such combinations of Bxb1 variants in the RD and ZD domains as MINT constructs. 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. 5g).

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 MINT constructs that worked best at the corresponding left and right half-sites, achieving ~35% TI at both loci (**Figure 2d**). For *MACO1* we observed 0.85% TI with wild-type Bxb1 and achieved a 41-fold increase in activity with a combination of our *MACO1* MINT constructs, while we detected 0.23% TI with wild-type Bxb1 at *GYS1* and achieved a 163-fold increase with a combination of our *GYS1* MINT constructs. We then investigated the genome-wide specificity of these two sets of MINT constructs (**Figure 2e**; Supplementary Tables 5-8). The top candidate integration site for both variants was the intended locus, with the signal for all other sites at least 20-fold lower. We did observe some integration at the *MACO1* pseudo-site with the *GYS1* MINT constructs, but this appeared as #13 in the candidate integration site list with 100-fold lower signal than the intended *GYS1* target. Conversely, integration at the *GYS1* target was not observed for the *MACO1* MINT constructs. Thus, we were able to retarget Bxb1 to previously weakly-active pseudo-sites with active and specific novel Bxb1 variants, indicating we had accomplished proper retargeting and did not simply relax the specificity of the parental constructs.

240 **Targeting clinically-relevant sites in the human genome**

241 Encouraged by our success targeting chromosomal pseudo-sites with our MINT constructs,
242 we proceeded to the more difficult challenge of targeting clinically relevant regions of the human
243 genome, such as the well-established safe harbor AAVS1 locus[52] and the T-cell receptor α constant
244 (TRAC) locus (**Figure 3a**), in which wild-type Bxb1 has no detectable activity. We first screened
245 Bxb1 variants against a library of potential target sites within intron 1 of these regions using our
246 mammalian system (Extended Data Fig. 7, Supplementary Tables 9-13). We then performed
247 custom hairpin selections for promising sites derived from this initial screen. This resulted in two
248 pairs of MINT constructs (**Figure 3b**) that can recognize the AAVS1 and TRAC loci respectively
249 and that each achieved up to ~1% TI (**Figure 3c,d**; Extended Data Fig. 8,9). We utilized DNA
250 donors with wild-type attP target sites with CDNs matched to the intended genomic target site and
251 co-delivered wild-type Bxb1 to recognize the attP site in the donor. Notably, we did not observe
252 any integration activity above background levels when we omitted the Bxb1 variant targeted to
253 either the left or right half-site, indicating full retargeting of our MINT constructs to TRAC and
254 AAVS1.

255 **Optimization of TRAC-targeted MINT constructs**

256 While we were able to reprogram Bxb1 to target two clinically relevant sites in human
257 cells, the efficiencies that we achieved are not high enough for many therapeutic applications. In
258 order to improve the efficiency of our fully reprogrammed Bxb1 variants that target the human
259 TRAC locus, we investigated two separate approaches: a first approach that involved fusing
260 engineered ZF arrays to our Bxb1 variants and a second approach that involved adding additional
261 mutations that have been shown to increase the activity of wild-type Bxb1 for its natural target
262 site[39].

263 Molecular modeling efforts indicated that fusing ZF arrays to the N-terminus of our MINT
264 constructs would likely cause numerous severe steric clashes, so we focused on fusing ZF arrays
265 to the C-terminus of our MINT constructs. Our modeling efforts also indicated which strand of
266 DNA we should target with our ZF arrays such that the N-terminus of each array sits as close as
267 possible to the C-terminus of the cognate Bxb1 variant when both are bound to their intended target
268 sites. In order to evaluate this approach, we designed and tested a panel of linkers and ZF arrays
269 to identify the optimal distance between the 3' edge of the ZF target site and the 5' edge of the

270 cognate MINT construct target site. We observed activity improvements when ZF arrays were
271 fused to either the left or right MINT construct and observed the largest improvements when the
272 separation between binding sites was between 4 and 10 bp (**Figure 4a**; Extended Data Fig. 9a).
273 Combining left and right MINT constructs both bearing a fused ZF array further improved
274 performance, yielding more than a 10-fold improvement relative to the same MINT construct
275 without fused ZFs (**Figure 4b**). Encouraged by this result, we next investigated the genome-wide
276 specificity of our TRAC constructs with and without the best performing combination of ZF arrays.
277 We observed a selective increase in signal for the on-target with addition of the ZF arrays with the
278 on-target having 9-fold higher signal than candidate OT1 with the ZF arrays compared to 1.5-fold
279 lower signal than candidate OT1 without these ZF arrays (**Figure 4c)**.

280 As an orthogonal method to improve integration activity, we added the three mutations
281 utilized in eeBxb1 (V74A, E229K, and V375I)[39] that have been shown to increase the activity of
282 wild-type Bxb1 for its target site. The addition of these three mutations increased the activity of
283 our AAVS1 MINT constructs by 14-fold (**Figure 3c**) and increased the activity of our TRAC MINT
284 constructs by 17-fold (**Figure 3d**). Finally, we combined both approaches with our TRAC MINT
285 constructs and achieved 35% TI into the TRAC locus in human K562 cells (**Figure 4d**) although
286 the activity increasing mutants did also increase off-target activity (Extended Data Fig. 9b and 9c,
287 Supplementary Tables 18-20). To characterize the nature of the integration, we performed single
288 cell cloning in K562 cells and confirmed full-length TI (Extended Data Fig. 10).

289 **Performance of MINT constructs in primary human T cells**

290 Encouraged by the high TI levels at the TRAC locus observed in human K562 cells, we
291 decided to test our best performing MINT constructs in the more therapeutically relevant setting
292 of primary human T cells. Since plasmid DNA has been shown to be poorly tolerated by primary
293 human cells[53,54], we delivered the Bxb1 variants as mRNAs and delivered the donor construct as
294 a nanoplasmid (circular double-stranded DNA lacking a bacterial original of replication and a
295 bacterial antibiotic resistance marker). In order to mimic a therapeutic targeted insertion event, we
296 added a GFP expression cassette to the donor and monitored integration stability and GFP
297 expression. Primary human CD4+/CD8+ T cells were activated and the Bxb1 variants and the
298 donor were delivered to the cells by nucleofection 3 days post-activation. After nucleofection, T
299 cells were incubated for 14 days to allow the GFP signal from unintegrated donor to declined and

300 then TI was characterized by DNA sequencing and GFP expression was analyzed by flow
301 cytometry (**Figure 5a**). MINT construct treatment did not markedly affect T cell growth after
302 recovery from the transfection and T cell markers were stably maintained over the course of the
303 experiments at comparable levels to untreated control cells. Integration activity with the
304 unoptimized TRAC MINT constructs was noticeably higher in T cells than in K562 cells perhaps
305 due to the higher levels of TRAC expression and greater chromosomal accessibility of the TRAC
306 locus in T cells. As in K562 cells, ZF fusions also markedly increased activity in T cells and we
307 achieved 10% TI and 11% stable GFP expression. MINT constructs with both ZF fusions and the
308 activity-increasing mutations yielded 15% TI (**Figure 5b**) and 26% stable GFP expression 14 days
309 post nucleofection (**Figure 5c**).

310 We observed subpopulations of cells exhibiting distinct levels of GFP expression, with the
311 high GFP expressing subpopulations more closely mirroring the TI results obtained by NGS
312 (**Figure 5b,c**). To further characterize the subpopulations, we sorted T cells based on GFP
313 expression levels and analyzed TI levels in the subpopulations. The result shows that cells with TI
314 into the TRAC locus are located in the high GFP expressing subpopulations, and that TI
315 percentages in the high GFP expressing subpopulations in MINT-ZF treated cells reach up to 71%,
316 indicating a high frequency of biallelic integration in the highest GFP expressing subpopulation
317 (Extended Data Fig. 11).

318 **Discussion and future directions**

319 In this work we have demonstrated the direct reprogramming of a site-specific LSR which
320 has been a long-standing challenge for genome engineering[16]. We achieved this through a
321 combination of AI-assisted structural modeling, directed evolution, and screening in human cells
322 that should be applicable to a wide range of site-specific DNA binding proteins. We were able to
323 use this approach to generate numerous reprogrammed Bxb1 variants capable of integrating
324 synthetic DNA into therapeutically relevant endogenous loci in the human genome with
325 efficiencies of up to 35%. To demonstrate that the MINT platform is compatible with AAV-
326 mediated donor delivery, we successfully tested the use of both single-stranded AAV and self-
327 complementary AAV (Extended Data Fig. 1c,d). By combining the MINT platform with known
328 activity-increasing Bxb1 mutants, and fusing ZF proteins to engineered Bxb1 variants, we were
329 able to achieve 15% TI of a functional donor construct into human T cells. We also observed that

330 on-target activity gains from ZF fusions translated better to human T cells than mutations intended
331 to make Bxb1 more active in general.

332 In contrast to methods that use DNA or RNA replication, such as homology-dependent
333 repair (HDR) or target-primed reverse transcription (TPRT), integrase techniques that use a cut-
334 and-paste mechanism have no inherent limit to the size of the DNA fragment that can be integrated
335 and remove transgene mutation risks caused by use of polymerases. This is not only beneficial for
336 the integration of synthetic DNA into AAVS1 and TRAC, as demonstrated in this study, but also
337 for targeting the first intron of a mutated gene to integrate the correct copy of the corresponding
338 coding sequence linked to a splice acceptor. Furthermore, targeting endogenous attP-like sites in
339 addition to attB-like sites would support other genomic rearrangements such as deletions and
340 inversions. Additionally, the ability to target two different endogenous loci in close proximity
341 would enable recombinase mediated cassette exchange (RMCE) at endogenous loci which could
342 replace an entire genomic locus with a linear synthetic DNA donor construct. This would enable
343 unprecedented flexibility in types of genomic alterations that can be generated[55]. We were able to
344 develop MINT constructs capable of targeting a site within the ~4 kb human AAVS1 locus and a
345 ~2 kb intron of the human TRAC locus which demonstrates a high targeting density that would
346 support RMCE at an endogenous locus.

347 Our results also provide insights into how serine integrases interact with their DNA target
348 sites. Our structural model suggests that residues 231-237 of Bxb1 interact with DNA in a manner
349 reminiscent of how the recognition helix of a ZF interacts with DNA. To our surprise, we observed
350 similar patterns of amino acids at specific positions of the Bxb1 recognition helix that have been
351 observed in ZF engineering efforts (Supplementary Fig. 1). In contrast, the hairpin region appears
352 to use a distinctly different mechanism to recognize DNA. Through an analysis of Bxb1 targets in
353 the human genome, we observed that the hairpin region of wild-type Bxb1 can recognize at least
354 four distinct DNA motifs implying that the hairpin region can either adopt multiple conformations
355 or bind using multiple DNA docking geometries (Extended Data Fig. 4c,d). This is consistent with
356 the observation that the ZD domain of wild-type Bxb1 can specifically recognize two completely
357 different sequences in the left and right half-sites of the natural attB target. The loop region appears
358 to recognize DNA using yet another mechanism. The fact that a single mutation, S157G, can
359 dramatically relax specificity implies that the loop region may act more as a switch that reduces

integration activity when the preferred sequence is not detected rather than the more standard model of interactions that increase binding affinity when the preferred sequence is present.

We envision that our MINT platform will also accelerate efforts in other research areas where Bxb1 has been successfully used, such as for metabolite pathway assembly[56] and various agrobiotechnology applications in model plants[57] and crops[58,59]. We believe 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.

In this study we established a method for retargeting Bxb1 that will likely enable the reprogramming of other serine integrases including those that have recently been described[38,60] (Supplementary Fig. 2). 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 integrase helix and hairpin variants. Such a pre-characterized archive would allow the *in silico* design of Bxb1 variants to target any sequence of interest by simply combining the appropriate helix variant with the appropriate hairpin variant without the need for custom directed evolution experiments. This work provides an initial archive of 26 hairpin variants and 24 helix variants that are active in human cells (Supplementary Fig. 3,4; Supplementary Tables 22-24) that can be used in conjunction with activity-increasing mutations and ZF fusions to target desired genomic loci for therapeutic applications and beyond.

Data availability

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

Code availability

All custom python scripts will be made available on GitHub after acceptance of this manuscript at a peer-reviewed scientific journal.

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 thank Trevor Collingwood for scientific discussions. We also thank Ed Rebar, Jon Melnick, Deepak Patil, and Charles Paine for contributions to the early development of ZF-targeted Serine Recombinases. And we thank Sandy Macrae, Adrian Woolfson, and Jason Fontenot for their relentless support of this project.

Author contributions

F.F., S.A.-F., J.E.D., L.L, N.J.S., L.R., N.A.S., R.M., I.T., B.N.K., M.Q, P.L., A.R. and J.C.M. designed experiments. F.F., S.A-F., J.E.D., L.L., N.J.S., L.R., D.F.X., N.N., N.A.S., R.M., I.T., Y.Z., L.N.T., S.J.H., B.N.K., S.L., B.B., E.T., M.Q., V.V., A.C., A.N., and P.L. performed experiments. F.F., S.A.-F., J.E.D., L.L., N.J.S., L.R., D.F.X., N.N., N.A.S., R.M., I.T., Y.Z., L.N.T., S.J.H., B.N.K., S.L., B.B., E.T., M.Q., V.V., A.N., P.L., A.R. and J.C.M. performed data analysis. F.F., S.A.-F., J.E.D., L.L., S.J.H., P.L., A.R., D.E.P., G.D.D., and J.C.M. supervised experiments. J.E.D., Y.R.B., D.A.S., and J.C.M. developed custom computer code. F.F., S.A.-F., and J.C.M wrote the manuscript with input from all authors.

Material requests and correspondence to F. Fauser and J.C. Miller.

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.

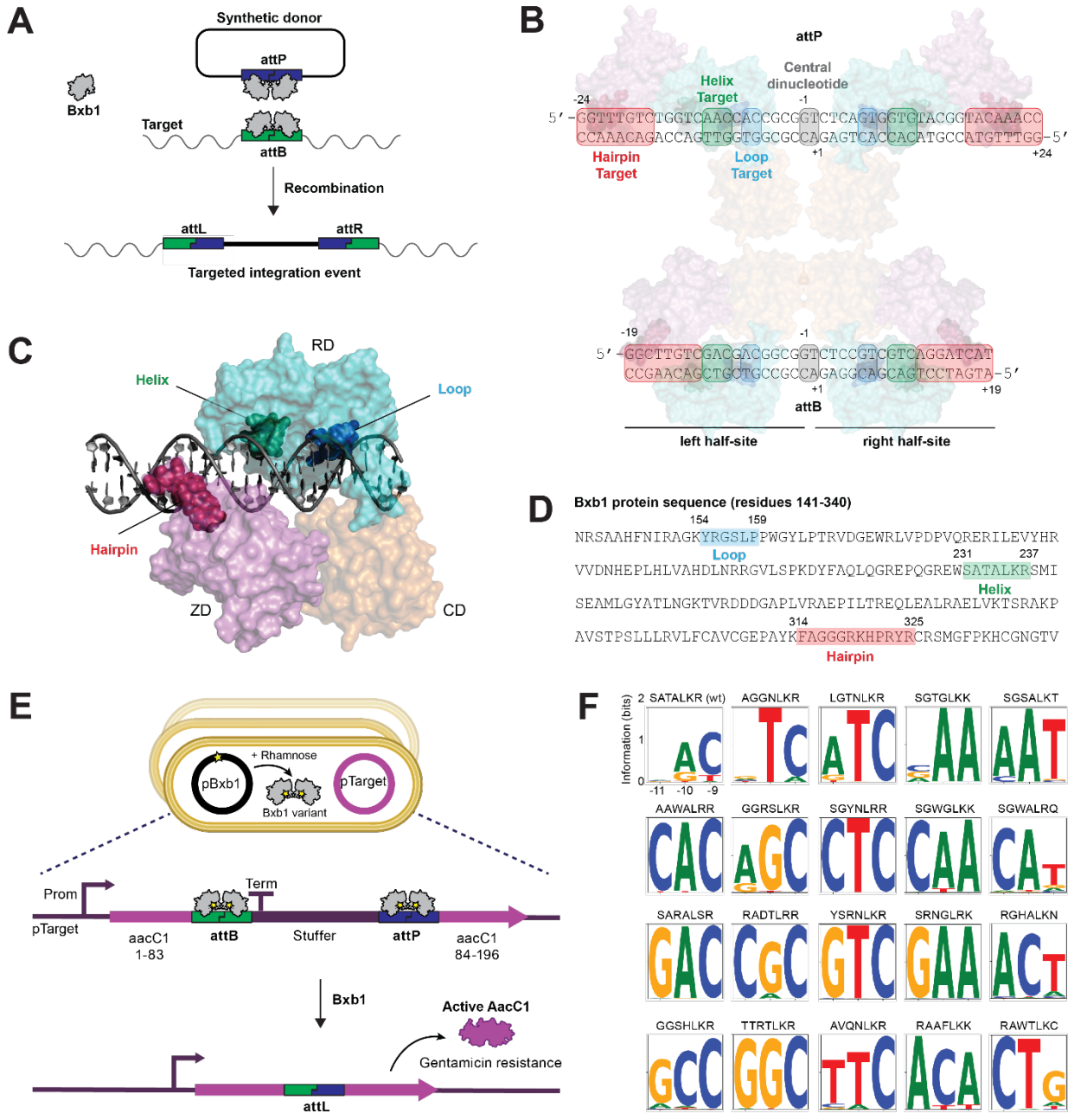


Figure 1. Bxb1-DNA interaction mapping and DNA binding domain engineering using directed evolution. a. Schematic of Bxb1-induced TI of a synthetic circular donor into a genomic target. Attachment sites are bound by Bxb1 dimers and non-reversible recombination is facilitated by a Bxb1 tetramer. **b.** The sequence of the natural Bxb1 attP and attB attachment sites with the portions recognized by the hairpin, helix, and loop regions of Bxb1 highlighted. 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. **c.** Structural model of a Bxb1 monomer bound to an attB attachment half-site. **d.** Amino acid sequence of wild-type Bxb1 with residues investigated in this study highlighted. Bxb1-DNA interaction mapping identified three specificity-determining regions of Bxb1 that can be reprogrammed: the loop region (residues 154-159; shown in blue), the helix region (residues 231-237; shown in green) which are both in the RD domain of Bxb1, and

419 the hairpin region (residues 314-325; shown in red) within the ZD domain of Bxb1. **e.** Schematic
420 of our two-plasmid directed evolution system. The expression plasmid pBxb1 encodes an integrase
421 gene, while the reporter plasmid pTarget contains an antibiotic marker disrupted by a stuffer
422 sequence that is flanked by modified attB and attP sequences. Recombination of these sites by an
423 active integrase excises the stuffer sequence and restores the open reading frame of the antibiotic
424 resistance marker. The selection system using gentamicin resistance is shown. The T in the stuffer
425 sequence indicates the position of a transcriptional terminator sequence. This system enables either
426 screening of libraries of integrase variants against a single DNA target or screening of a library of
427 DNA targets against a single integrase. **f.** Example DNA specificity plots of selected helix variants.
428 Data in panel d is the mean of four biological replicates. Plotted data are provided in the Source
429 Data file.

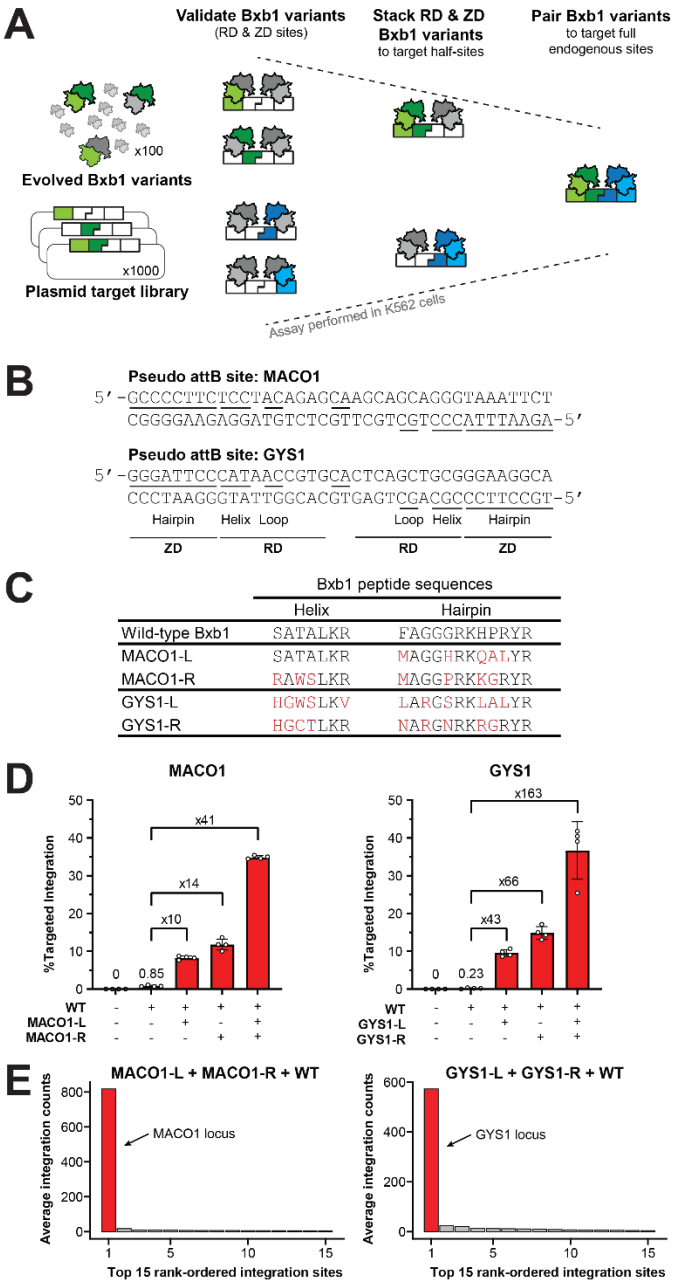


Figure 2. Performance of engineered Bxb1 variants at pseudo-sites in the human genome. a. Schematic of plasmid-based system for testing evolved Bxb1 variants against artificial DNA targets. Our LSR engineering strategy divides each endogenous target site into four "quarter-sites" where the left and right 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 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. Data are presented as the mean ± s.d. from four biological replicates. **e.** Results from an unbiased genome-wide specificity assay demonstrating a strong on-target preference of our engineered Bxb1 variants. The top 15 candidate integration sites are shown (out of total 112 sites for MACO1 constructs and 103 for GYS1 constructs) with the entire datasets provided in Supplementary Tables 5 and 6 and corresponding control datasets in Supplementary Tables 7 and 8. Average integration counts were determined using the average deduplicated read numbers across 2 replicates and 2 samples each using primers that amplify the attL and attR plasmid-genome junctions. Plotted data are provided in the Source Data file.

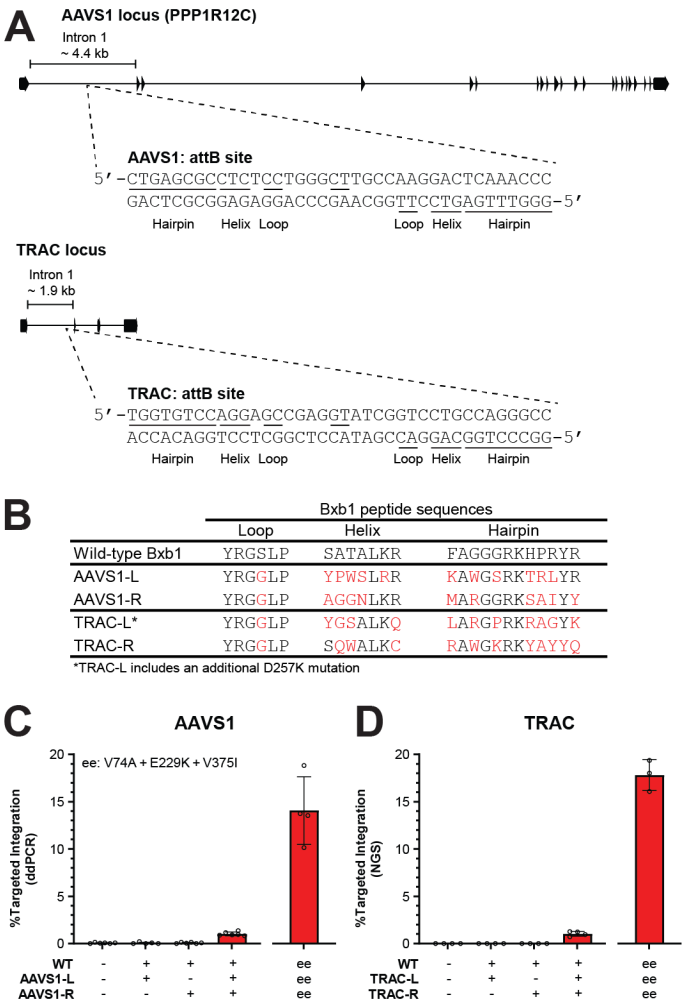


Figure 3. Retargeting Bxb1 to the human TRAC and AAVS1 loci. a. Sequence of the attB sites recognized by the TRAC and AAVS1 Bxb1 variants shown in panel b. Both attB target sites are within the first intron of each gene. **b.** Bxb1 peptide sequences of evolved Bxb1 variants for each of the two TRAC and AAVS1 half-sites generated by the strategy shown in **Figure 3a**. **c.** Results from a digital droplet PCR-based assay demonstrating TI at the depicted AAVS1 site. See Extended Data Fig. 8 for additional details. **d.** Results from our PCR-based NGS assay demonstrating TI at the depicted TRAC site. Data are presented as the mean ± s.d. from four biological replicates for TRAC, and six biological replicates for AAVS1 respectively. Plotted data are provided in the Source Data file.

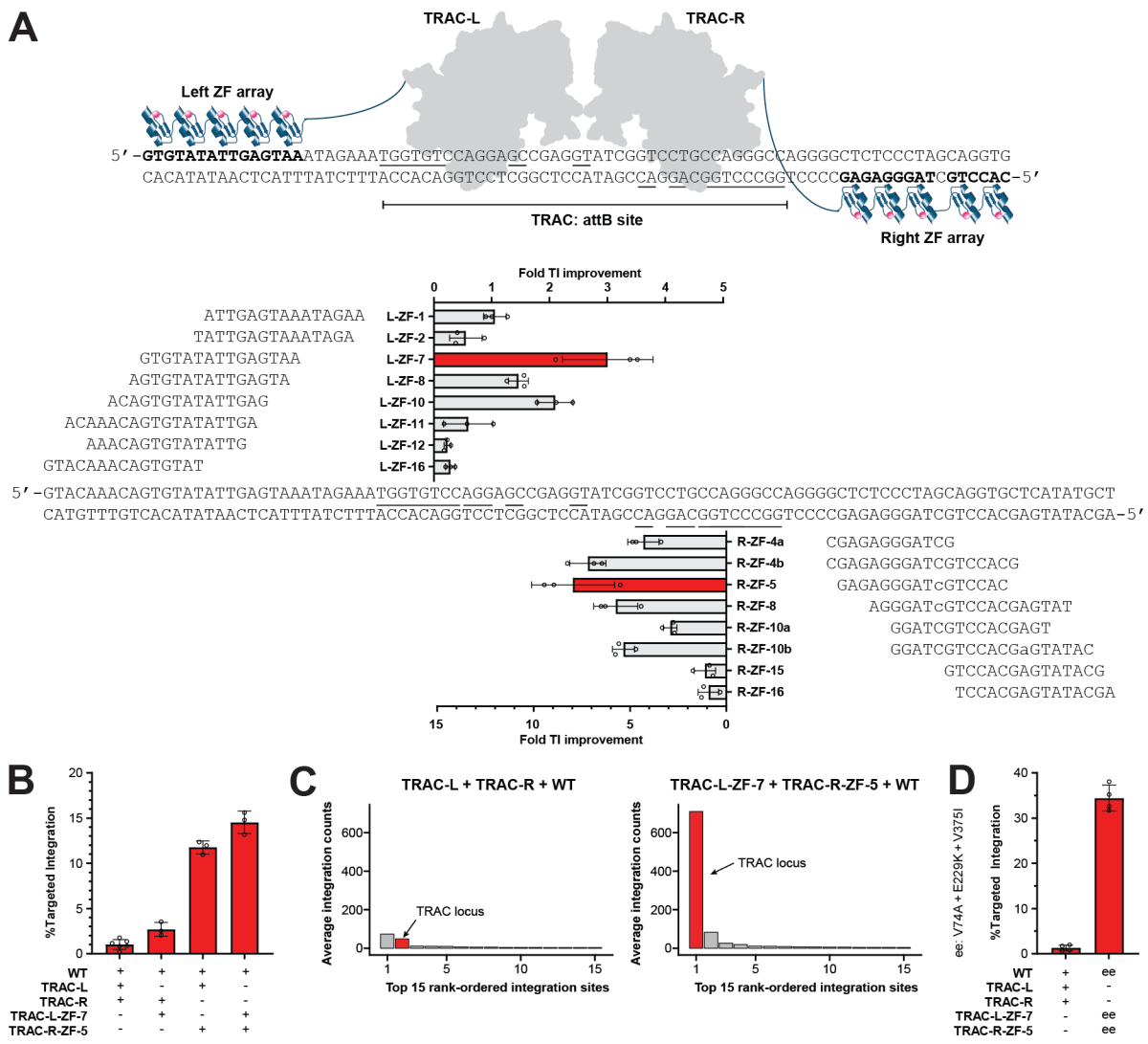


Figure 4. Optimization of TRAC constructs. a. Illustration of ZF-fused MINT constructs targeting the human TRAC locus. Eight different ZF arrays were tested for both TRAC-L and TRAC-R. Shown are the respective DNA binding sites and the fold TI improvement when tested by themselves (e.g. L-ZF-1-TRAC-L + TRAC-R + WT). **b.** Demonstration that the best performing constructs from panel a can be paired to further improve efficiency. **c.** Results from an unbiased genome-wide specificity assay demonstrating the selective increase ZF-fusion adds to the on-target integration signal for TRAC MINT constructs. The top 15 candidate integration sites are shown (out of total 85 sites for TRAC constructs and 94 for ZF-fused TRAC constructs) with the entire datasets provided in Supplementary Tables 14 and 15 and corresponding control datasets in Supplementary Tables 16 and 17. Average integration counts were determined using the average deduplicated read numbers across 2 replicates and 2 samples each using primers that amplify the attL and attR plasmid-genome junctions. **d.** Demonstration that activity-increasing Bxb1 variants such as eeBxb1 can be combined with ZF-fused MINT constructs. Data in panels a-c derived from a PCR-based NGS assay and presented as the mean ± s.d. from three to four biological replicates. Plotted data are provided in the Source Data file.

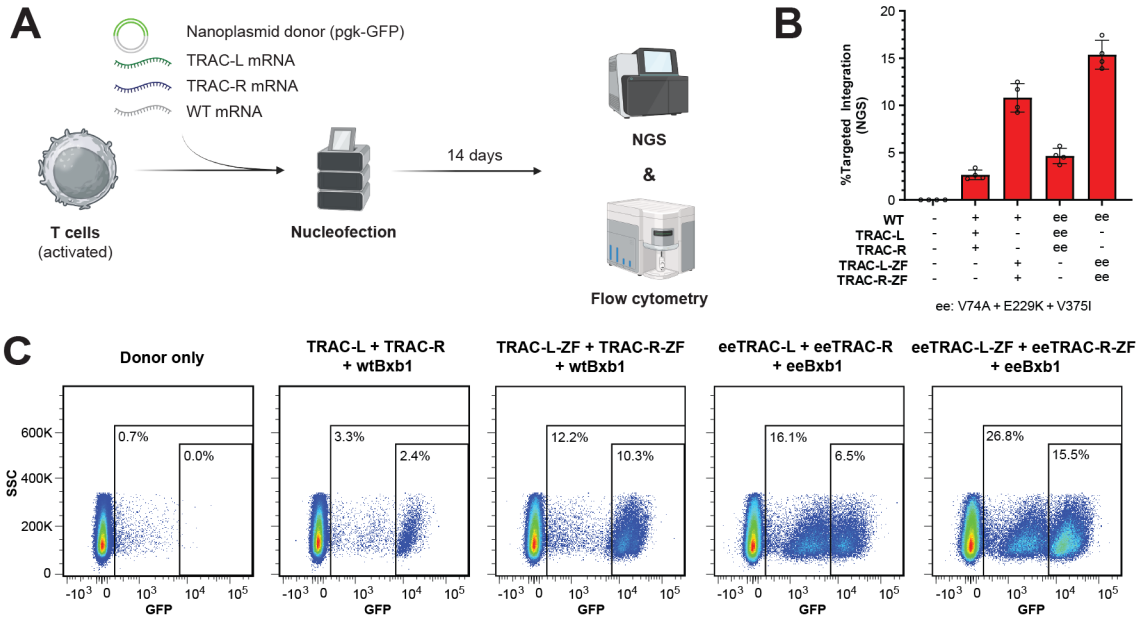


Figure 5. Testing MINT TRAC constructs in primary human T cells. a. Experimental procedure of T cell nucleofection and TI quantification. **b.** TI percentage determined by next generation sequencing analysis on DNA isolated 14 days after nucleofection with the indicated constructs. **a** was created with BioRender.com. **b.** Flow cytometric analysis of GFP expression 14 days after nucleofection. The percentage of the GFP-positive cells is indicated. TRAC-L-ZF-7 and TRAC-R-ZF-5 from **Figure 4** were used in all T cell nucleofections. Next generation sequencing data are presented as the mean ± s.d. from 4 biological replicates. The flow cytometry plots show individual representative samples. Plotted data are provided in the Source Data file.

490 **Methods**

491 **Cloning of expression constructs and donors used in mammalian cells**

492 Most constructs were cloned using NEBuilder® HiFi DNA Assembly (NEB, Catalog
493 #E2621X) or Q5® Site-Directed Mutagenesis (NEB, Catalog #E0554S), or synthesized and cloned
494 by Twist Biosciences utilizing their clonal gene services. The gBlocks or eBlock DNA fragments
495 used for constructs assembly were synthesized at IDT (Integrated DNA Technologies). DNA
496 sequences for all constructs can be found in supporting sequence information. All constructs were
497 sequence confirmed using Sanger sequencing services provided by Elim Biopharm Inc., and whole
498 plasmid sequences were verified using either the Nextera XT DNA library prep kit (Illumina,
499 Catalog #FC-131-1096) or whole plasmid sequencing services (Plasmidsaurus Inc.; Elim
500 Biopharma Inc.).

501 The ZF constructs have the following amino acid sequences where recognition helix
502 sequences are underlined and structured basepair skipping linkers are in italics:

503 **High throughput Bxb1 variants assembly**

504 The Bxb1 variant gene fragments (bases 460 – 1087, corresponding amino acids 144 – 362
505 with loop, helix and hairpin regions included) were synthesized as eBlocks (Integrated DNA
506 Technologies) and assembled to full length fusion expression cassette through 2-step PCR. In the
507 first step PCR, the eBlocks were amplified with overlapping gene fragments DF148 and DF164
508 with AccuPrime *Pfx* SuperMix (Invitrogen, Catalog #12344040) and the following thermocycler
509 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
510 2 min 30 s; followed by a final extension at 68 °C for 5 min; hold at 4 °C. The full-length expression
511 cassette PCR products were then amplified with forward primer 5'-
512 GCAGAGCTCTCTGGCTAACTAGAG-3' and reverse primer 5'-
513 TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTAG
514 ACAGGCGGGGAGGCGGC-3' and the following thermocycler conditions: initial melt of 95 °C
515 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
516 extension at 68 °C for 5 min; hold at 4 °C. The final PCR products were examined on a gel and
517 purified with AMPure XP beads (Beckman# A63881). The purified PCR products were used for
518 mRNA production.

519 **mRNA production**

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

524 **AAV production**

525 AAV genome plasmids containing the relevant attP sequences for Bxb1-mediated
526 recombination, flanked by inverted repeat terminal sequences (ITR) for AAV packaging, were
527 cloned using NEBuilder® HiFi DNA Assembly (NEB, Catalog #E2621X). ITRs were mutated to
528 produce scAAV[61]. Bovine growth hormone (bGH) polyA termination sequence was inserted
529 upstream of the 3' ITR to aid transcription termination as well as AAV titration. All constructs were
530 sequence confirmed using Sanger sequencing services provided by Sequetech Corporation. The
531 whole plasmid sequences were verified using Nextera XT DNA library prep kit (Illumina, Catalog
532 #FC-131-1096) or whole plasmid sequencing services provided by Plasmidsaurus Inc. to ensure
533 ITRs are intact prior to AAV production.

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

539 **Directed evolution system**

540 **Preparation of Integrase Mutant libraries**. As a starting point for engineering of Bxb1
541 variants in E. coli, we cloned a codon optimized ORF into plasmid pBxb1, which contains a
542 pBR322 origin of replication as well as the L-rhamnose inducible pRhaBAD promoter. Libraries
543 were prepared using inverse PCR and primers encoding degenerate bases using an NNK
544 degeneracy scheme. In order to diversify the "loop" submotif, primers were designed to target
545 residues 154-159. In order to diversify the "helix" submotif, primers were designed to target
546 residues 231-234, and residues 236-237. In order to diversify the "Beta-hairpin" submotif, primers
547 were designed to target residues 314, 316, 318, 321, 323, and 325 using an NNK randomization

548 scheme, while residue 322 was randomized using an SSK randomization scheme. Briefly, iPCR
549 reactions were set up in 50 uL volumes, using 1x KOD ONE Master mix, 0.5uM forward primer,
550 0.5 uM reverse primer, and 10 ng of template plasmid DNA. Following purification of the PCR
551 products via silica column (Qiagen PCR cleanup kit), and elution in 50 uL of buffer EB, compatible
552 overhangs for ligation were generated by digestion of PCR products using 5 units each of BsaI
553 (NEB, R3733S) and DpnI (NEB, R0176S) in 60 uL of 1x CutSmart buffer. The resulting digest
554 products were then ligated at 10 ng/uL in 1x T4 Ligase buffer, at 8 U/uL T4 Ligase, for 1 hour at
555 room temperature. Following ligation, DNA products were purified using the Qiagen PCR cleanup
556 kit and eluted in 50 uL of buffer EB. The resulting ligated products were used to transform 50 uL
557 of electrocompetent cells (NEB 10Beta, C3020K; or ThermoFisher OneShot Top10, C404052)
558 using a BMX electroporator and a 96-well cuvette (BTX, 45-0450-M) using the manufacturer's
559 protocol and allowed to recover for 1 hour at 37 °C in 1ml total volume of SOC media. The
560 resulting transformations yielded libraries in the order of 4e8 to 5e9 CFUs. After recovery, cells
561 were transferred to 1 L of LB media containing 34 ug/mL chloramphenicol and grown overnight
562 with shaking at 37 °C. The resulting culture was harvested, and plasmids were purified using a
563 Qiagen Plasmid Plus Giga Kit.

564 **Preparation of selection cassette plasmids.** Selection plasmid pTarget was based on a
565 p15A origin of replication and contained a spectinomycin resistance cassette. The selection
566 cassette contained a Tet promoter upstream of an open reading frame encoding a gentamicin
567 resistance gene (aacC1) disrupted by a "stuffer sequence". The stuffer sequence was flanked with
568 extra sequence upstream of attB and downstream of attP to enable changing of reading frames if
569 attL recombination produced an unviable peptide stuffer sequence. The attB and attP sequences
570 were modified such that recombination by an active integrase would leave behind an attL sequence
571 encoding an in-frame peptide insertion, giving rise to an active aacC1. Each selection cassette was
572 generated by the cloning of two ~1050 bp DNA fragments into a linearized pTarget backbone
573 plasmid using the NEB HiFi assembly master mix, according to the manufacturer's instructions.
574 Sequence-confirmed plasmids were either prepared by Mini- or Midi- Kits (Qiagen) according to
575 the manufacturer's instructions.

576 **Selection of Bxb1 variants by aacC1 reassembly.** Typical selection reactions used 1ug of
577 Bxb1 library, and 1ug selection cassette plasmid. OneShot Top10 electrocompetent cells were

578 mixed with library and selection plasmids before being transferred to a 96-well cuvette, and
579 electroporated as above. After recovery, cells were transferred to 4 mL of LB containing 34 ug/mL
580 chloramphenicol, 50 ug/mL spectinomycin and 0.2% L-rhamnose, then the cultures were shaken
581 at 37 °C for 16 h to allow recombination to take place. For selections using liquid broth, cells were
582 spun down and resuspended in 50 mL Super Broth containing 34 ug/mL chloramphenicol and 20
583 ug/mL gentamicin. After overnight culture, 1 mL of each culture was used to extract plasmid DNA
584 using the Qiagen Miniprep kit. For selections using agar plates, cells were instead spun down and
585 resuspended in 1 mL of Superbroth and plated on 25x25 cm LB agar plates containing 34 ug/mL
586 chloramphenicol and 20 ug/mL gentamicin. After overnight incubation at 37 °C, colonies were
587 scraped from plates, resuspended in 10 mL LB broth, and a 500 uL volume was removed and used
588 in plasmid purification using Qiagen Miniprep kits.

589 **NGS sequencing of Bxb1 variant submotifs post-selection.** Samples were prepared for
590 sequencing using Illumina MiSeq or NextSeq using an integrase-adaptor hybrid primer pair
591 (PCR1) followed by an Illumina-adaptor specific primer pair (PCR2). Briefly, 50uL reactions were
592 prepared using 1x Q5 Hotstart Master mix (NEB M0494S), 0.5uM Forward primer, 0.5uM Reverse
593 primer, and 50ng of plasmid DNA. PCR was performed using the following cycling conditions,
594 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
595 at 72 °C for 30s. The resulting PCR product was used as template for PCR2. Briefly, 50 uL
596 reactions were prepared using 1x KOD ONE polymerase master mix (Millipore-Sigma), 0.5 uM
597 Adapter primer 1, 0.5 uM Adapter primer 2, and 1 uL of PCR1 as template. PCR was performed
598 using the following cycling conditions 98 °C 30 s, then 13 cycles of 98 °C 5s, 60 °C 7 s, 72 °C 5
599 s followed by a final extension step at 72 °C for 30 s. The resulting PCR products were column
600 purified using a Qiagen PCR cleanup kit, and samples were sequenced using standard Illumina
601 Kits for MiSeq or NextSeq.

602 **Molecular specificity assays.** We built a plasmid reporter to assay the molecular
603 specificity of novel integrase submotifs. Three versions of the pBxb1 plasmids were used as
604 starting point to avoid background activity of the wild-type integrase during cloning. By placing
605 an extra adenine residue within the loop, helix, or hairpin submotif, we created pBxb1 variants
606 where Bxb1 was inactivated by a frameshift-inducing mutation but could be rescued by subsequent
607 mutagenesis. Recombination cassettes from the pTarget selection plasmids were amplified by PCR

608 and cloned into the pBxb1 plasmids by Gibson assembly upstream of the Bxb1 ORF. To assay the
609 specificity of selected loops, a single reverse primer and different forward primers which encoded
610 the new loop residues were used in iPCR using the small library of loop target plasmids (16
611 targets). A similar procedure was used to generate helix mutants using the small library of helix
612 target plasmids (64 targets).

613 To generate a library of ZD hairpin targets, we amplified and cloned an oligo pool (Twist
614 Biosciences) which comprised all possible attB and attP sites where positions -18 to -13 have been
615 randomized into our pBxb1 plasmid, downstream of the integrase ORF. Such a library ensures that
616 the ZD hairpin targets on attBL and attPL are identical, and the targets on attBR and attPR are
617 inverted repeats of the left side targets. Mutant hairpins are generated by two different iPCR
618 primers, which together encode the novel hairpin sequence.

619 iPCR reactions were carried out in 50 uL volumes, with 1 x KOD ONE master mix, 0.5uM
620 forward primer, 0.5uM reverse primer, and 20 ng of a plasmid pool containing the relevant
621 recombination cassettes. Thermocycling was carried out using an initial denaturation step of 98 °C
622 for 30s, followed by 35 cycles of 98 °C 10s, 60 °C 5s, 68 °C 30s. PCR amplicons were purified
623 using AMPure XP beads (Beckman) and eluted in 40 uL EB buffer (Qiagen). 30 uL of the purified
624 PCR amplicons were then digested and ligated in a 50 uL one-pot reaction containing 1x T4 ligase
625 buffer (NEB), 20U DpnI (NEB), 20 U BsaI-HFv2 (NEB), and 400 U of Salt-T4 DNA Ligase
626 (NEB). All reactions were then incubated at 37 °C 30 minutes, 20 °C 30 minutes, 37 °C for 30
627 minutes. 20 uL of thawed chemically competent NEB 5alpha cells were added 2 uL of the
628 digested/ligated amplicons, and cells were transformed according to the manufacturer's
629 instructions. After recovery, cells were added to 800 uL of media containing 34ug/mL
630 chloramphenicol and 0.2% L-rhamnose (w/v). The resulting culture was incubated for 16 hours at
631 37 °C with shaking in 96-well deep-well plates, then harvested by centrifugation. Plasmid DNA
632 was extracted using a Qiaprep 96 Turbo Miniprep kit.

633 To assess the specificity of each clone, primers which flank the resulting attL sequence and
634 contain Illumina adaptor sequences were used to amplify the products of recombination. Briefly,
635 PCR reactions were carried out in 20uL volumes, using 1 x Hotstart Taq master mix (NEB,
636 M0496S), attB_MiSeq forward primer, and attP_Miseq reverse primer, using 1uL of purified
637 plasmid as template. Thermocycling was carried out as follows 98 °C 30s; 25 cycles of [98 °C 5s,

638 53 °C 10s, 72 °C 5s] Final extension 72 °C 30seconds. A second PCR to install sequencing
639 barcodes was carried out using specific barcoding primers, with the recombination products from
640 each clone being represented by a unique combination of forward and reverse barcoding primers.
641 PCR reactions were carried out in 20 uL volumes, using 1 x KOD ONE master mix, 0.5 uM
642 forward primer, 0.5 uM reverse primer, and 1 uL of the previous PCR product. Thermocycling was
643 carried out as follows 98C 30s; 12 cycles of [98 °C 5s, 60 °C 5s, 68 °C 3s]. The resulting PCR
644 products were column purified using Qiagen PCR cleanup kits, then sequenced using standard
645 Illumina kits for NextSeq or MiSeq.

646 **General mammalian cell culture conditions**

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

650 CD4+/CD8+ primary human T cells were purified from healthy donor leukopaks. Primary
651 human T cells were cultured in X-Vivo 15 (Lonza, BP04-744Q) supplemented with HEPES
652 (Gibco, 15630-080), GlutaMAX (Gibco, 35050-061), Sodium Pyruvate (Gibco, 11360-070),
653 MEM vitamins (Corning 25-020-c1), human AB serum (Valley BioMedical, HS1017HI) & IL-2
654 (Invitrogen CTP0023).

655 **K562 tissue culture nucleofection protocol and genomic DNA preparation**

656 Expression constructs were routinely dosed as plasmid DNA (pDNA) in K562 cells. K562
657 cells were electroporated with pDNA using the SF cell line 96-well Nucleofector kit (Lonza,
658 Catalog#V4SC-2960) or SF Cell Line 384-well Nucleofector Kit (Lonza, Catalog#V5SC-2010),
659 using manufacturer's protocol. Prior to electroporation, K562 cells were centrifuged at ~300 x g
660 for 5 min, and washed with 1x DPBS (Corning, Catalog#21-031-CV). For 96-well nucleofection,
661 cells were resuspended at 2e5 cells per 12 µl of supplemented SF cell line 96-well Nucleofector
662 solution. 12 µl of cells were mixed with 8 µl of pDNA and transferred to the Lonza Nucleocuvette
663 plate. Nucleofector program 96-FF-120 was used to electroporate K562 cells with the pDNA mix
664 on the Amaxa Nucleofector 96-well Shuttle System (Lonza). After electroporation, cells were
665 incubated for 10 min at room temperature and transferred to a 96-well tissue culture plate
666 containing 180 µl of complete medium (prewarmed to 37 °C).

667 For 384-well nucleofection, cells were resuspended at 1e5 cells per 14 µl of supplemented
668 SF cell line 384-well Nucleofector solution. 14 µl of cells were mixed with 6 µl of pDNA including
669 1400ng donor, 66ng of wt, 66ng of left and 66ng right bxb1 variants or bxb1 variant-ZFP fusion
670 DNA and transferred to the Lonza Nucleocuvette plate. Nucleofector program FF/120/DA was
671 used to electroporate K562 cells with the pDNA mix on the Amaxa HT Nucleofector System
672 (Lonza, AAU-1001). After electroporation, cells were incubated for 10 min at room temperature
673 and transferred to a 384-well tissue culture plate containing 60 µl of complete medium (prewarmed
674 to 37 °C). K562 cells were incubated for ~72 h and then harvested for quantification of editing
675 events.

676 **HepG2 tissue culture nucleofection protocol and genomic DNA preparation**

677 HepG2 cells were electroporated with pDNA using the SF cell line 96-well Nucleofector
678 kit (Lonza, Catalog #V4SC-2960) using manufacturer's protocol. 1000 ng of Bxb1 or GFP control
679 along with 5000 ng of plasmid donor was used. Prior to electroporation, HepG2 cells were
680 centrifuged at ~200 x g for 10 min, and washed with 1x DPBS (Corning, Catalog #21-031-CV).

681 For 96-well nucleofection, cells were resuspended at 2e5 cells per 20 µl of SF Nucleofector
682 Solution with supplement. 16.5 µl of cells were mixed with 3.5 µl of pDNA, then transferred to
683 the Lonza 96-well Shuttle plate. Nucleofector program 96-EH-100 was used to electroporate
684 HepG2 cells with the pDNA mix on the Amaxa Nucleofector 96-well Shuttle System (Lonza).
685 After electroporation, cells were incubated for 10 min at room temperature, 180 ul of growth
686 medium (MEM plus 10% FBS) were added per well and cells transferred to a 48-well tissue culture
687 plate containing 300 µl medium (prewarmed to 37 °C). Electroporated HepG2 cells were placed
688 in a 30 °C incubator for 24 hours and then transferred to a 37 °C incubator. HepG2 cells were
689 incubated for ~72 h and then harvested for quantification of editing events.

690 **Digital Droplet PCR quantification of targeted integration at AAVS1 target site**

691 For AAVS1 target integration ddPCR quantification, 2 TI probes (targeting attL and attR)
692 and 2 reference probes were designed. The target integration probes were designed targeting either
693 attL or attR region which would not be detected in non-transfected cells or plasmid donors. The
694 reference probes were designed within a 10 kb region of the AAVS1 attB target site to mitigate the
695 risk of copy number difference due to abnormal karyotype in K562 cells. The probes and primers

696 were designed with PrimerPlus3 (https://www.primer3plus.com/) based on the recommendation
697 by BioRad manual and were synthesized at IDT (Integrated DNA Technologies, Inc.). The primers
698 and probes were tested in duplex format at different annealing temperatures and with synthetic
699 eblock mixes, non-transfected cell lysate and NTC. The attL and reference2 probes and primers
700 and 57.1°C were selected for AAVS1 target site ddPCR quantification.

701 DNA was extracted with QuickExtract DNA Extraction Solution (Lucigen#QE09050). For
702 each reaction, 50 µl of QuickExtract DNA solution was added to approximately 0.5-1 million
703 pelleted cells, followed by mixing and incubation at 65 °C for 15 min and heat inactivation at 98 °C
704 for 5 min. The cell lysates were mixed by vortexing for 15 seconds before ddPCR. Each ddPCR
705 reaction was prepared and analyzed with a QX200 ddPCR system (Bio-Rad) and ddPCR Supermix
706 for Probes (No dUTP) (Bio-Rad, Catalog #1863024) per Bio-Rad's standard recommendations.
707 All reactions were mixed to 22 µl including 10 U of HindIII-HF (NEB, Catalog #R3104L) and up
708 to 2 µl of QuickExtract lysates. Forward primer, reverse primer and probe were at a 3.6:3.6:1 ratio.
709 Droplets were generated in the droplet generator per Bio-Rad's protocol. Thermocycler conditions:
710 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
711 8 °C. QX Manager Software 2.1 Standard Edition (Bio-Rad) was used for QC and the analysis.
712 The thresholds were set manually at 3000 for channel1/FAM and 1000 for channel2/HEX. All final
713 data was exported into Microsoft Excel for further analysis. The target integration ratio was
714 calculated by the equation: TI (%) = 100*CattL/Cref2 (C: volumetric concentration (copies/µl)).

715 **AAV transduction**

716 For evaluating rAAVs. K562 cells were electroporated with 800ng of Bxb1 pDNA using
717 the SF cell line 96-well Nucleofector kit (Lonza, Catalog#V4SC-2960), using manufacturer's
718 protocol. K562 cells were centrifuged at ~300 x g for 5 min, and washed with 1X PBS (Corning,
719 Catalog#21-031-CV). For 96-well nucleofection, cells were resuspended at 2E5 cells per 12 µl of
720 supplemented SF cell line 96-well Nucleofector solution. 12 µl of cells were mixed with 8 µl of
721 Bxb1 pDNA and transferred to the Lonza Nucleocuvette plate. Nucleofector program 96-FF-120
722 was used to electroporate K562 cells with the pDNA mix on the Amaxa Nucleofector 96-well
723 Shuttle System (Lonza). After electroporation, cells were incubated for 10 min at room temperature
724 and transferred to a 96-well tissue culture plate containing 180 µl of complete medium (prewarmed

725 to 37 °C). 30 mins post-electroporation, rAAV constructs were dosed at MOI (multiplicity of
726 infections) at 500,000 vg/cell. rAAV donor only control wells were included in parallel. K562 cells
727 were incubated for ~72 h and then harvested for quantification of editing and circularization
728 events. A PCR-based NGS assay was used to measure TI events.

729 **K562 landing pad cell line generation**

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

746 **PCR-based NGS assay for targeted integration and indel quantification**

747 3 days post transfection, cells were spun down at ~500 x g for 5 min. Supernatant was
748 discarded, cells were washed in PBS (Corning, Catalog #21-031-CV), and cells were resuspended
749 in 50 µl of QuickExtract DNA Extraction Solution (Lucigen, Catalog #QE09050). Genomic DNA
750 was extracted by treating the cells to the following protocol: 65 °C for 15 min, 98 °C for 8 min.
751 Target sites were amplified from the genomic DNA using Accuprime HiFi reagents (Invitrogen,
752 Catalog #12346094) and the following PCR conditions: initial melt of 95 °C for 5 min; 30 cycles
753 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.
754 Primers containing adapters (forward primer adapter: ACACGACGCTCTTCCGATCT; reverse

755 primer adapter: GACGTGTGCTCTTCCGATCT), targeting specific target sites were used at a
756 final concentration of 0.1 μM. Sequences for the primers used can be found in supporting sequence
757 information. The PCR products obtained were then subjected to a second PCR to add Illumina
758 barcodes to the PCR fragments generated in the first PCR. We used Phusion High-Fidelity PCR
759 MasterMix with HF Buffer (NEB, Catalog #M0531L) for the second PCR and used the following
760 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
761 °C for 40 s; and a final extension at 72 °C for 10 min. PCR libraries generated from the second
762 PCR were pooled and purified using QIAquick PCR purification kit (Qiagen, Catalog #28106).
763 Samples were diluted to a final concentration of ~2 nM after they were quantified using the Qubit
764 dsDNA HS Assay kit (Invitrogen, Catalog #Q33231). The libraries were then run on either an
765 Illumina MiSeq using a standard 300-cycle kit or an Illumina NextSeq 500 or an Illumina NextSeq
766 2000 using a mid-output 300-cycle kit using standard protocol.

767 **Pooled screening of attB sites in K562 cells**

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

773 **Genome-wide specificity assay**

774 Our genome-wide specificity assay was adapted from the GUIDE-seq protocol[51], with
775 modifications designed to measure Bxb1-induced off-target editing events. K562 cells were
776 transfected using conditions similar as described above. Transfections involving integrations with
777 multiple donors with different core dinucleotides were performed separately per donor plasmid
778 and then cells were pooled and expanded for 1 week before being spun down for genomic DNA
779 extraction. Without the DpnI site addition to donor plasmids, the cells were grown out for 3-4
780 weeks prior to DNA extraction and the DpnI digestion step below was not followed.

781 Genomic DNA was extracted from K562 cells using Qiagen DNeasy Blood & Tissue kits
782 (Catalog #69504) following the Purification of Total DNA from Animal Blood or Cells Spin-
783 Column Protocol for cultured cells. The optional RNase A incubation was followed for all samples.

784 DNA was eluted in 60 µL Elution Buffer and quantified using the Qubit fluorometer and the Qubit
785 dsDNA HS Assay Kit (Invitrogen, Catalog #Q33231) following the recommended protocol.

786 Next, 10 uM adapters were prepared by annealing the MiSeq common oligo to each GS_i5
787 oligo in a 96-well plate format to make a barcoded Y adapter plate. Annealing was performed with
788 1X oligo annealing buffer (10 mM Tris HCL pH 7.5, 50 mM NaCl, and 0.1 mM EDTA) by
789 following the below thermocycling method: initial melt of 95 °C for 2 min, step-down from 80 °C
790 to 4 °C with -1 °C per cycle and 1 min incubation at each temperature, hold at 4 °C until further
791 use. Adapters were stored at -20 °C, and before use were thawed on ice400 ng (133,000 haploid
792 human genomes) genomic DNA was brought up to 50 uL using 1X IDTE pH 7.5 (IDT, Catalog
793 #11-05-01-05) in each tube of a Covaris 8 microTUBE-130 AFA Fiber H Slit Strip V2 (Covaris,
794 Catalog #520239). Samples were sonicated on a Covaris ME220 using the following settings on a
795 ME220 Rack 8 AFA-TUBE TPX Strip (Covaris, Catalog #PN500609) using the waveguide
796 (Covaris PN 500526): Power 0.0 W, Temperature 19.7 °C, Duration(s) 65.0, Peak Power 40.0,
797 Duty %Factor 10.0, Cycles/Burst 10000, Avg. Power 4.0.

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

806 The reaction was brought up to 50 uL with the addition of CutSmart (final concentration
807 1X) and 1 uL DpnI (NEB, Catalog #R0176S) and incubated for 1 hour at 37 °C. DNA was purified
808 using 0.8x Ampure XP beads using the same bead clean-up protocol as before. Next, the following
809 End repair and A-tailing mixture was added to each 14.5 uL DNA mixture, while the reaction was
810 kept on ice: 0.5 uL 10 mM dNTP mix (Invitrogen, Catalog #18427013), 2.5 uL 10X T4 DNA
811 Ligase Buffer (Enzymatics, Catalog #B6030), 2 uL End-repair mix (Enzymatics, Catalog #Y9140-
812 LC-L), 2 uL 10X Platinum Taq Polymerase PCR Rxn Buffer (-Mg2 free) (Invitrogen, Catalog
813 #10966034), 0.5 uL Taq DNA Polymerase Recombinant (5u/ uL) (Invitrogen, Catalog

814 #10342020), and 0.5 uL dsH2O to a total of 22.5 uL. The solution was mixed and incubated on a
815 thermocycler with the following program: 12 °C for 15 min, 37 °C for 15 min, 72 °C for 15 min,
816 hold at 4 °C until further use.

817 Next, 2 uL T4 DNA Ligase (Enzymatics, Catalog #L6030-LC-L) and 1 uL of one of the 10
818 uM annealed Y adapters (chosen from a 96-well plate of GS_i5 adapters) was added per end-
819 repaired and A-tailed reaction. Each reaction was mixed and incubated on a thermocycler with the
820 following program: 16 °C for 30 min, 22 °C for 30 min, hold at 4 °C until further use.DNA was
821 then purified using 0.9 volumes Ampure XP beads using the same bead clean-up protocol but using
822 23 uL IDTE pH 7.5 to resuspend the DNA-bead mixture for elution and collecting 22 uL
823 supernatant after incubation on the magnet.

824 Next, ligated and sheared DNA fragments were amplified using a primer specific to all
825 adapters (P5_1) and a primer specific to the sequence of interest (GSP1 +/-). To each tube on ice,
826 the following reagents were added: 22 uL DNA from previous step, 3 uL 10X Platinum Taq
827 Polymerase PCR Rxn Buffer (-Mg2 free) (Invitrogen, Catalog #10966034), 1.2 uL 50 mM MgCl2
828 (Invitrogen, Catalog# 10966034), 0.6 uL 10 mM dNTP mix (Invitrogen, Catalog #18427013), 0.5
829 uL 10 µM P5_1 primer, 1 uL 10 µM PSP1+/-, 1.5 uL 0.5 M TMAC (Sigma Aldrich, Catalog
830 #T3411), and 0.3 uL Platinum Taq DNA polymerase (5 U/µL) (Invitrogen, Catalog #10966034) to
831 a total of 30.1 uL. Each reaction was mixed and incubated with the following thermocycler
832 conditions: initial melt of 95 °C for 2 min; 15 cycles of 95 °C for 30 s, 70 °C (-1 °C/cycle) for 2
833 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
834 30 s; followed by a final extension at 72 °C for 5 min; hold at 4 °C until further use.

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

838 Next, the second round of PCR amplification was performed to add plate barcodes (GS_i7
839 sequences). To each tube on ice, the following reagents were added: 1.5 ul 10 µM Plate adapter
840 (GS_i7), 20.4 ul DNA from previous step, 3 ul 10X Platinum Taq Polymerase PCR Rxn Buffer (-
841 Mg2 free) (Invitrogen, Catalog #10966034), 1.2 ul 50 mM MgCl2 (Invitrogen, Catalog
842 #10966034), 0.6 ul 10 mM dNTP mix (Invitrogen, Catalog #18427013), 0.5 ul 10 µM P5_2 primer,
843 1 ul 10 µM PSP2+/-, 1.5 ul 0.5 M TMAC (Sigma Aldrich, Catalog #T3411), and 0.3 ul Platinum

844 Taq DNA polymerase (5 U/µL) (Invitrogen, Catalog #10966034) to a total of 30 ul. Each reaction
845 was mixed and incubated with the following thermocycler conditions: initial melt of 95 °C for 5
846 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
847 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
848 °C for 5 min; hold at 4 °C until further use.

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

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

866 For NextSeq reactions, 1.98 ul of 100 uM custom sequencing primer Read2 was added to
867 600 ul Illumina HP21 primer mix for a 0.3 uM final concentration and 3.98 ul of 100 uM custom
868 sequencing primer Index1 was added to 600 ul BP14 primer mix for a 0.6 uM final concentration.
869 Then 550 µl of custom primer mix was added to custom 1 well or custom 2 well on the reagent
870 cartridge separately.

871 **T cell Transfection**

872 Primary human T cells were thawed on day -3 and activated with anti CD3/CD28 beads
873 (Thermo Fisher Scientific, Catalog #CTS40203D). Three days later, 'day 0', T cells were

874 centrifuged at ~300 x g for 10 min, and washed with 1x DPBS (Corning, Catalog #21-031-CV).
875 For 96-well nucleofection, cells were resuspended at 1E6 cells per 20 µl of P3 Primary Cell
876 Nucleofector Solution (Lonza, Catalog #V4SP-3096) with supplement. 20 µl of cells were mixed
877 with 9.6 µl of Bxb1 mRNA and nanoplasmid donor (Aldevron), then transferred to the Lonza 96-
878 well Shuttle plate. Nucleofector program EO-115-AA was used to electroporate T cells with the
879 mRNA/DNA mix on the Amaxa Nucleofector 96-well Shuttle System (Lonza). After
880 electroporation, cells were incubated for 10 min at room temperature, 80ul of T-cell medium were
881 added per well and cells transferred to a 48-well tissue culture plate containing 400 µl medium
882 (prewarmed to 37 °C). Electroporated T cells were placed in a 30°C incubator for 24 hours and
883 then transferred to a 37°C incubator. Cells were cultured until day 14 by replenishing culture
884 medium every 3 to 4 days. On day 14, GFP expression was determined by Flow Cytometry and T
885 cell aliquots were harvested, DNA isolated and PCR amplified with the specified primers to assess
886 target integration into the TRAC by Illumina next generation sequencing.

887 **Flow Cytometry of T cells**

888 On day 14 post transfection, ~5x10^5 cells were harvested, washed once with PBS, then
889 stained with Fixable Viability Dye eFluor™ 780 (Thermo Fisher, Catalog #65-0865) according to
890 Manufacturer's instructions. The GFP flow analysis was performed using an Attune NxT Acoustic
891 Cytometer (Thermo Fisher Scientific). Subsequently, the flow data was analyzed with FlowJo
892 software (BD).

893 For cell sorting 3ml (~ 6E6) cells were spun down, washed once with PBS, de-beaded using
894 a Dynal MPC-L magnetic particle concentrator (Thermo Fisher) and then resuspend in 1.5ml of
895 EasySep buffer (Stemcell Technologies, Catalog #20144). The cell suspension was loaded on a
896 Sony cell sorter SH800S for sorting & GFP fraction collection. For the sorting, a 70um sorting
897 chip (Sony, Catalog #LE-C3207) and sorting pressure 4 was employed. A small volume of the
898 mock transfected control sample was used to set the GFP negative gate and a MINT transfected
899 sample to set gates for the different expression levels of GFP as shown in Extended Data Fig. 11.+.
900 For cell collection, 0.4 ml PBS were pre-aliquoted to each 15 ml collection tube, desired numbers
901 (when available) of GFP negative & GFP positive fractions were collected.

902 After GFP fractions were collected, an aliquot of the sorted samples was centrifuged and
903 appropriate GFP fractionation was confirmed by flow cytometry analysis. The remaining cells
904 were lysed with QuickExtract buffer for DNA extraction and analyzed for TI as described above.

905 **Computational analysis of directed evolution experiments**

906 Data processing was performed with custom python scripts and logo plots were generated
907 using Logomaker[62]. Sequence reads that don't match sequences encoded by the directed evolution
908 library were filtered out as were sequence reads that were likely artifacts due to being rare sequence
909 reads with only one or two differences from much more common sequence reads. The remaining
910 sequence reads were translated into peptide sequence, and the rest of the analysis only considered
911 the peptide sequence of the randomized region. For hairpin selections, the partially randomized
912 position 322 could be either a G, A, R, or P and the motif analysis was conducted separately based
913 on the identity of position 322. The clearest signal was 4 residue motifs within the 6 fully
914 randomized positions. For example, the helix library with a XXXXLXX randomization scheme
915 had a strongly enriched motif of XGGNLXR where X is any amino acid. To account for potential
916 biases in the starting library, motifs were scored based on the FDR corrected p-value of a given
917 motif occurring by chance given the amino acid frequency at each position when considered
918 independently. For helix selections, enriched 4-residue motifs that were observed in selections with
919 a wide variety of different DNA target sequences were considered to be non-specific and were
920 filtered out. Compatible four residue motifs (e.g. XGXNLKR and XXGNLKR) were then
921 combined to create motifs. Several peptides that best matched the most enriched 4 residue motifs
922 in each combined motif were then chosen for additional characterization.

923 **Computational analysis of pooled screening of attB sites in K562 cells**

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

929 **Computational analysis of chromosomal targeted integration events**

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

937 **Computational identification of Bxb1 pseudo-sites**

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

953 **Computational analysis of the unbiased genome-wide specificity assay**

954 NGS reads were demultiplexed, adapter trimmed, and filtered for a minimum quality
955 threshold of 14 over all bases. Samples then underwent analysis for plasmid integration site
956 detection. NGS samples were processed to remove remaining contaminant unintegrated plasmid
957 reads due to incomplete DpnI digestion or no DpnI digestion in the case of week 3 and 4 samples.
958 Reads were then aligned to the hg38 genome, and candidate integration sites were summarized.
959 First, read 2 reads that contained both attP sequence 5' and 3' of the dinucleotide were removed

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

979 **Single cell cloning and targeted integration characterization**

980 K562 cells were nucleofected with eebxb1-ZFP constructs as detailed above. Three days
981 post-nucleofection, cells were examined via microscopy and quantified using automated cell
982 counters (Thermo Fisher Scientific). Subsequently, cells were seeded into 96-well plates at an
983 average density of 0.3 cells per well. On day 10, wells exhibiting GFP fluorescence and containing
984 a single cell cluster were maintained for an additional 10 days. Thereafter, cells were subjected to
985 flow cytometry analysis and genomic DNA was extracted using QuickExtract (Lucigen
986 #QE09050) for TI analysis.
987 To verify full-length and single-copy donor integration, DNA from single-cell clones
988 demonstrating >99% target integration efficiency and 100% GFP positive was amplified using
989 primers LL-P125 (5'-GATCAATCTGAGTGCTGTACGG-3') and LL-P126 (5'-

990 ACCCCTGTCTTACCTGTTTCAA-3'). Briefly, PCR reactions were carried out in 20uL volumes,
991 using 2 x Phusion High-Fidelity PCR master mix (NEB M0531L), and 2 uL of quick extracted
992 DNA as template. Thermocycling was carried out as follows 98 °C 1min; 35 cycles of [98 °C
993 15sec, 65 °C 30sec, 72 °C 10min] Final extension 72 °C 5 min. The resulting PCR products were
994 resolved by gel electrophoresis and confirmed by Nextera sequencing following the instruction of
995 Illumina Nextera XT DNA Library Preparation Kit (FC-131-1096).

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 Pabo, C. O., Peisach, E. & Grant, R. A. Design and selection of novel Cys2His2 zinc finger proteins. *Annu Rev Biochem* **70**, 313-340 (2001). https://doi.org:10.1146/annurev.biochem.70.1.313

11 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

12 Carroll, D. Genome engineering with zinc-finger nucleases. *Genetics* **188**, 773-782 (2011). https://doi.org:10.1534/genetics.111.131433

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

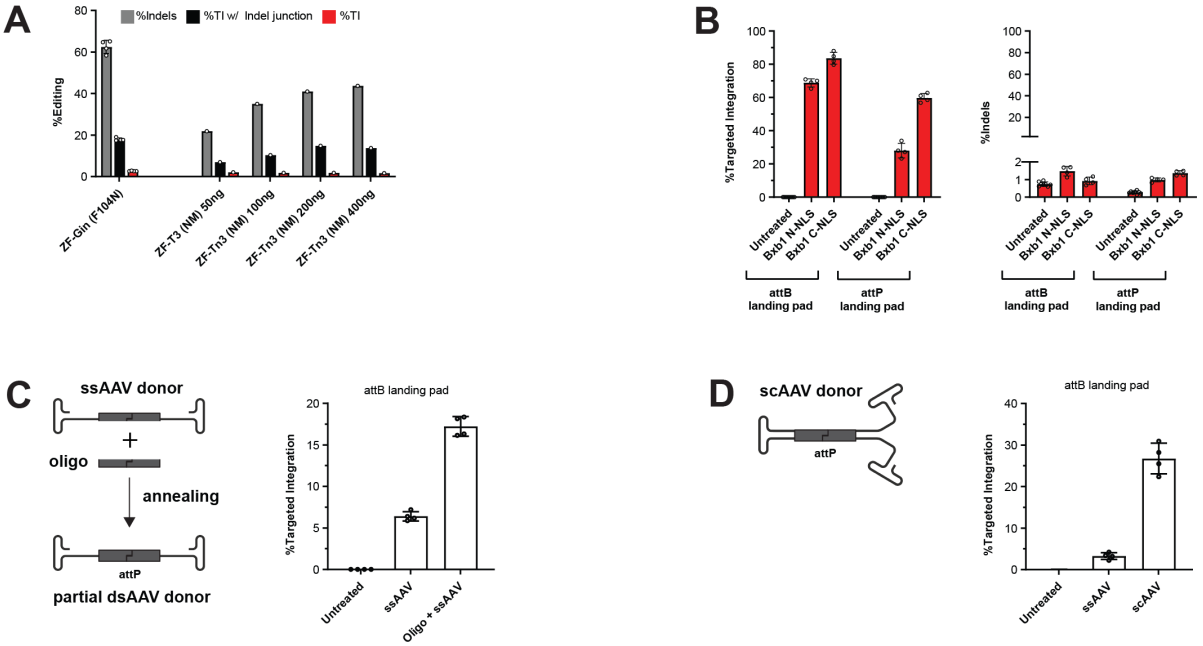
17 Chaikind, B., Bessen, J. L., Thompson, D. B., Hu, J. H. & Liu, D. R. A programmable Cas9-serine recombinase fusion protein that operates on DNA sequences in mammalian cells. *Nucleic Acids Res* **44**, 9758-9770 (2016). https://doi.org:10.1093/nar/gkw707

1039 18 Gaj, T., Mercer, A. C., Sirk, S. J., Smith, H. L. & Barbas, C. F., 3rd. A comprehensive
1040 approach to zinc-finger recombinase customization enables genomic targeting in human
1041 cells. *Nucleic Acids Res* **41**, 3937-3946 (2013). https://doi.org:10.1093/nar/gkt071
1042 19 Gordley, R. M., Gersbach, C. A. & Barbas, C. F., 3rd. Synthesis of programmable
1043 integrases. *Proc Natl Acad Sci U S A* **106**, 5053-5058 (2009).
1044 https://doi.org:10.1073/pnas.0812502106
1045 20 Mercer, A. C., Gaj, T., Fuller, R. P. & Barbas, C. F., 3rd. Chimeric TALE recombinases
1046 with programmable DNA sequence specificity. *Nucleic Acids Res* **40**, 11163-11172 (2012).
1047 https://doi.org:10.1093/nar/gks875
1048 21 Prorocic, M. M. *et al.* Zinc-finger recombinase activities in vitro. *Nucleic Acids Res* **39**,
1049 9316-9328 (2011). https://doi.org:10.1093/nar/gkr652
1050 22 Standage-Beier, K. *et al.* RNA-Guided Recombinase-Cas9 Fusion Targets Genomic DNA
1051 Deletion and Integration. *CRISPR J* **2**, 209-222 (2019).
1052 https://doi.org:10.1089/crispr.2019.0013
1053 23 Voziyanova, E., Li, F., Shah, R. & Voziyanov, Y. Genome targeting by hybrid Flp-TAL
1054 recombinases. *Sci Rep* **10**, 17479 (2020). https://doi.org:10.1038/s41598-020-74474-2
1055 24 Mukhametzyanova, L. *et al.* Activation of recombinases at specific DNA loci by zinc-
1056 finger domain insertions. *Nat Biotechnol* (2024). https://doi.org:10.1038/s41587-023-
1057 02121-y
1058 25 Feng, X., Bednarz, A. L. & Colloms, S. D. Precise targeted integration by a chimaeric
1059 transposase zinc-finger fusion protein. *Nucleic Acids Res* **38**, 1204-1216 (2010).
1060 https://doi.org:10.1093/nar/gkp1068
1061 26 Owens, J. B. *et al.* Transcription activator like effector (TALE)-directed piggyBac
1062 transposition in human cells. *Nucleic Acids Res* **41**, 9197-9207 (2013).
1063 https://doi.org:10.1093/nar/gkt677
1064 27 Owens, J. B. *et al.* Chimeric piggyBac transposases for genomic targeting in human cells.
1065 *Nucleic Acids Res* **40**, 6978-6991 (2012). https://doi.org:10.1093/nar/gks309
1066 28 Ye, L. *et al.* TAL effectors mediate high-efficiency transposition of the piggyBac
1067 transposon in silkworm Bombyx mori L. *Sci Rep* **5**, 17172 (2015).
1068 https://doi.org:10.1038/srep17172
1069 29 Yant, S. R., Huang, Y., Akache, B. & Kay, M. A. Site-directed transposon integration in
1070 human cells. *Nucleic Acids Res* **35**, e50 (2007). https://doi.org:10.1093/nar/gkm089
1071 30 Klompe, S. E., Vo, P. L. H., Halpin-Healy, T. S. & Sternberg, S. H. Transposon-encoded
1072 CRISPR-Cas systems direct RNA-guided DNA integration. *Nature* **571**, 219-225 (2019).
1073 https://doi.org:10.1038/s41586-019-1323-z
1074 31 Strecker, J. *et al.* RNA-guided DNA insertion with CRISPR-associated transposases.
1075 *Science* **365**, 48-53 (2019). https://doi.org:10.1126/science.aax9181
1076 32 Lampe, G. D. *et al.* Targeted DNA integration in human cells without double-strand breaks
1077 using CRISPR-associated transposases. *Nat Biotechnol* **42**, 87-98 (2024).
1078 https://doi.org:10.1038/s41587-023-01748-1
1079 33 Witte, I. P. *et al.* Programmable gene insertion in human cells with a laboratory-evolved
1080 CRISPR-associated transposase. *Science* **388**, eadt5199 (2025).
1081 https://doi.org:10.1126/science.adt5199
1082 34 Durrant, M. G. *et al.* Bridge RNAs direct programmable recombination of target and donor
1083 DNA. *Nature* **630**, 984-993 (2024). https://doi.org:10.1038/s41586-024-07552-4

1084 35 Merrick, C. A., Zhao, J. & Rosser, S. J. Serine Integrases: Advancing Synthetic Biology.
1085 *ACS Synth Biol* **7**, 299-310 (2018). https://doi.org:10.1021/acssynbio.7b00308
1086 36 Bessen, J. L. *et al.* High-resolution specificity profiling and off-target prediction for site-
1087 specific DNA recombinases. *Nat Commun* **10**, 1937 (2019).
1088 https://doi.org:10.1038/s41467-019-09987-0
1089 37 Anzalone, A. V. *et al.* Programmable deletion, replacement, integration and inversion of
1090 large DNA sequences with twin prime editing. *Nat Biotechnol* **40**, 731-740 (2022).
1091 https://doi.org:10.1038/s41587-021-01133-w
1092 38 Yarnall, M. T. N. *et al.* Drag-and-drop genome insertion of large sequences without double-
1093 strand DNA cleavage using CRISPR-directed integrases. *Nat Biotechnol* **41**, 500-512
1094 (2023). https://doi.org:10.1038/s41587-022-01527-4
1095 39 Pandey, S. *et al.* Efficient site-specific integration of large genes in mammalian cells via
1096 continuously evolved recombinases and prime editing. *Nat Biomed Eng* (2024).
1097 https://doi.org:10.1038/s41551-024-01227-1
1098 40 Xu, Z. *et al.* Accuracy and efficiency define Bxb1 integrase as the best of fifteen candidate
1099 serine recombinases for the integration of DNA into the human genome. *BMC Biotechnol*
1100 **13**, 87 (2013). https://doi.org:10.1186/1472-6750-13-87
1101 41 Rebar, E. J. & Pabo, C. O. Zinc finger phage: affinity selection of fingers with new DNA-
1102 binding specificities. *Science* **263**, 671-673 (1994).
1103 https://doi.org:10.1126/science.8303274
1104 42 Greisman, H. A. & Pabo, C. O. A general strategy for selecting high-affinity zinc finger
1105 proteins for diverse DNA target sites. *Science* **275**, 657-661 (1997).
1106 https://doi.org:10.1126/science.275.5300.657
1107 43 Karpinski, J. *et al.* Directed evolution of a recombinase that excises the provirus of most
1108 HIV-1 primary isolates with high specificity. *Nat Biotechnol* **34**, 401-409 (2016).
1109 https://doi.org:10.1038/nbt.3467
1110 44 Lansing, F. *et al.* Correction of a Factor VIII genomic inversion with designer-
1111 recombinases. *Nat Commun* **13**, 422 (2022). https://doi.org:10.1038/s41467-022-28080-7
1112 45 Rutherford, K., Yuan, P., Perry, K., Sharp, R. & Van Duyne, G. D. Attachment site
1113 recognition and regulation of directionality by the serine integrases. *Nucleic Acids Res* **41**,
1114 8341-8356 (2013). https://doi.org:10.1093/nar/gkt580
1115 46 Li, H., Sharp, R., Rutherford, K., Gupta, K. & Van Duyne, G. D. Serine Integrase attP
1116 Binding and Specificity. *J Mol Biol* **430**, 4401-4418 (2018).
1117 https://doi.org:10.1016/j.jmb.2018.09.007
1118 47 Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track
1119 neural network. *Science* **373**, 871-876 (2021). https://doi.org:10.1126/science.abj8754
1120 48 Gersbach, C. A., Gaj, T., Gordley, R. M. & Barbas, C. F., 3rd. Directed evolution of
1121 recombinase specificity by split gene reassembly. *Nucleic Acids Res* **38**, 4198-4206 (2010).
1122 https://doi.org:10.1093/nar/gkq125
1123 49 Pavletich, N. P. & Pabo, C. O. Zinc finger-DNA recognition: crystal structure of a Zif268-
1124 DNA complex at 2.1 A. *Science* **252**, 809-817 (1991).
1125 https://doi.org:10.1126/science.2028256
1126 50 Zheng, Z. *et al.* Anchored multiplex PCR for targeted next-generation sequencing. *Nat Med*
1127 **20**, 1479-1484 (2014). https://doi.org:10.1038/nm.3729

1128 51 Tsai, S. Q. *et al.* GUIDE-seq enables genome-wide profiling of off-target cleavage by
1129 CRISPR-Cas nucleases. *Nat Biotechnol* **33**, 187-197 (2015).
1130 https://doi.org:10.1038/nbt.3117
1131 52 Kotin, R. M., Linden, R. M. & Berns, K. I. Characterization of a preferred site on human
1132 chromosome 19q for integration of adeno-associated virus DNA by non-homologous
1133 recombination. *EMBO J* **11**, 5071-5078 (1992). https://doi.org:10.1002/j.1460-
1134 2075.1992.tb05614.x
1135 53 Roth, T. L. *et al.* Reprogramming human T cell function and specificity with non-viral
1136 genome targeting. *Nature* **559**, 405-409 (2018). https://doi.org:10.1038/s41586-018-0326-
1137 5
1138 54 Hornung, V. & Latz, E. Intracellular DNA recognition. *Nat Rev Immunol* **10**, 123-130
1139 (2010). https://doi.org:10.1038/nri2690
1140 55 Inniss, M. C. *et al.* A novel Bxb1 integrase RMCE system for high fidelity site-specific
1141 integration of mAb expression cassette in CHO Cells. *Biotechnol Bioeng* **114**, 1837-1846
1142 (2017). https://doi.org:10.1002/bit.26268
1143 56 Wang, X. *et al.* Bxb1 integrase serves as a highly efficient DNA recombinase in rapid
1144 metabolite pathway assembly. *Acta Biochim Biophys Sin (Shanghai)* **49**, 44-50 (2017).
1145 https://doi.org:10.1093/abbs/gmw115
1146 57 Thomson, J. G. *et al.* The Bxb1 recombination system demonstrates heritable transmission
1147 of site-specific excision in Arabidopsis. *BMC Biotechnol* **12**, 9 (2012).
1148 https://doi.org:10.1186/1472-6750-12-9
1149 58 Blechl, A., Lin, J., Shao, M., Thilmony, R. & Thomson, J. The Bxb1 Recombinase
1150 Mediates Site-Specific Deletion in Transgenic Wheat. *Plant Molecular Biology Reporter*
1151 **30**, 1357-1366 (2012). https://doi.org:10.1007/s11105-012-0454-2
1152 59 Jiang, L. *et al.* Target lines for recombinase-mediated gene stacking in soybean. *Theor Appl*
1153 *Genet* **135**, 1163-1175 (2022). https://doi.org:10.1007/s00122-021-04015-6
1154 60 Durrant, M. G. *et al.* Systematic discovery of recombinases for efficient integration of large
1155 DNA sequences into the human genome. *Nat Biotechnol* **41**, 488-499 (2023).
1156 https://doi.org:10.1038/s41587-022-01494-w
1157 61 McCarty, D. M. *et al.* Adeno-associated virus terminal repeat (TR) mutant generates self-
1158 complementary vectors to overcome the rate-limiting step to transduction in vivo. *Gene*
1159 *Ther* **10**, 2112-2118 (2003). https://doi.org:10.1038/sj.gt.3302134
1160 62 Tareen, A. & Kinney, J. B. Logomaker: beautiful sequence logos in Python. *Bioinformatics*
1161 **36**, 2272-2274 (2020). https://doi.org:10.1093/bioinformatics/btz921

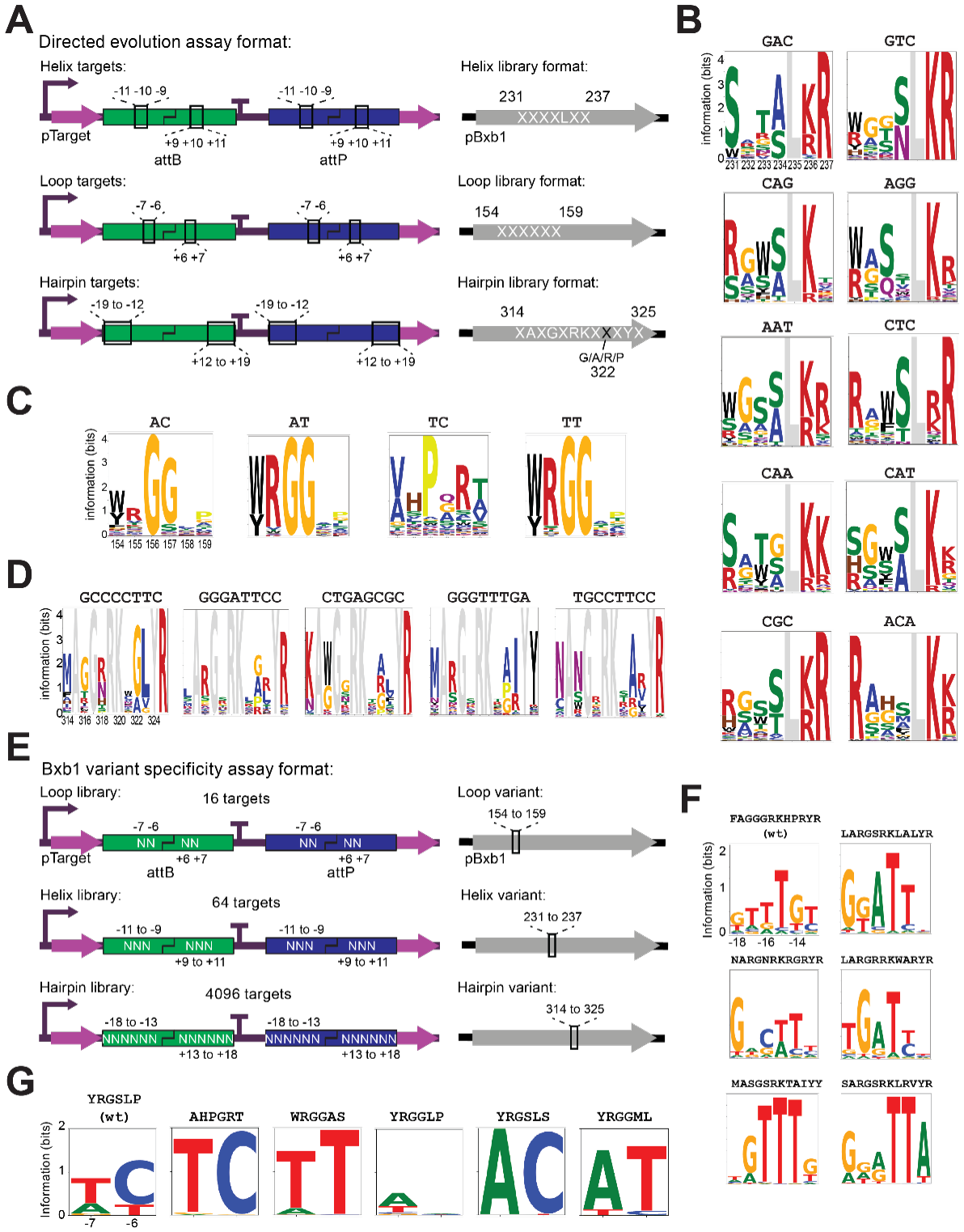
1162



Extended Data Figure 1. Performance of zinc fingers fused to small serine recombinase domains and performance of different wild-type Bxb1 constructs and donors. a. Performance of Zinc Finger (ZF)-targeted small serine recombinase Gin and Tn3 hyperactive domains. The data shown in this graph is derived from separate experiments. In contrast to Bxb1, both Gin and Tn3 hyperactive domain fusion proteins can result in high levels of indels causing low product purity and hindering further improvement of target integration frequencies. We also observed indels within the assayed TI junction site. Data is derived from a PCR-based NGS. Data are presented as the mean from four biological replicates for ZF-Gin, and individual data points are shown for ZF-Tn3 fusions. **b.** Performance of wild-type Bxb1 against natural attB and attP target sequences in human cells. We first established K562 landing pad cell lines by installing the natural attB or attP sequence in the human AAVS1 locus. We noticed improved performance of Bxb1 with a C-terminal NLS compared to the construct with a N-terminal NLS. This guided future Bxb1 designs where all evolved variants presented in this study have a C-terminal NLS. We also noticed higher TI into the attB landing pad. Notably, no or only minimal levels of indels were observed within the landing pad target sequences. Data is derived from a PCR-based NGS. Data are presented as the mean ± s.d. from four biological replicates (eight for untreated sample). **c.** Utilization of a single-stranded AAV (ssAAV) donor for Bxb1-mediated TI. 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. **d.** Utilization of a self-complementary AAV (scAAV) donor for Bxb1-mediated TI. Data are presented as the mean ± s.d. from at least four biological replicates.

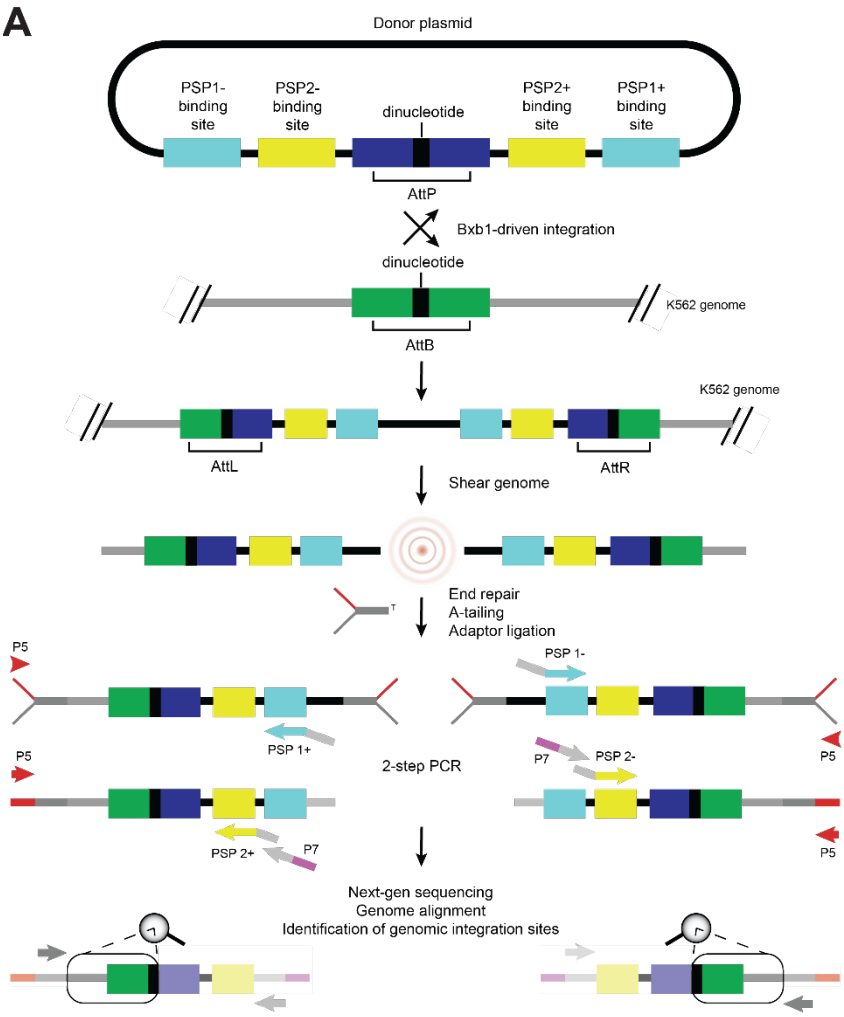
```
LI integrase       1 MKAAIYIRVSTQEQVENYSIQAQTEKLTALCRSKDWDVYDTFIDGGYSGS-----NMNRP  55
                     M+A + IR+S        S + Q E    LC  + WDV     D   SG+        RP
Bxbl integrase     1 MRALVVIRLSRVTDATT-SPERQLESCQQLCAQRGWDVVGVAEDLDVSGAVDPFDRKRRP  59

LI integrase      56 ALNEMLS-KLHEIDAVVVYRLDRLSRSQKDTITLIEEYFLKNNVEFV--SLSETLDTSSP 112
                      L   L+ +    D +V YR+DRL+RS +    L+  ++ +++ + V  +     DT++P
Bxbl integrase    60 NLARWLAFEEQPFDVIVAYRVDRLTRSIRHLQQLV--HWAEDHKKLVVSATEAHFDTTTP 117

LI integrase     113 FGRAMIGILSVFAQLERETIRDRMVMGKIKRIEAGLPLTTAKGRTFGYDVID----TKLY 168
                     F   +I ++   AQ+E E I++R        I AG    +     +GY         +L
Bxbl integrase   118 FAAVVIALMGTVAQMELEAIKERNRSAAHFNIRAGKYRGSL--PPWGYLPTRVDGEWRLV 175

LI integrase     169 INEEEAKQLRLIYDIFEEEQ-SITFLQKRLKKLGF-----------------KVRTYNRY 210
                      +  + +++  +Y    +    +  +   L + G                  +  +
Bxbl integrase   176 PDPVQRERILEVYHRVVDNHEPLHLVAHDLNRRGVLSPKDYFAQLQGREPQGREWSATAL 235

LI integrase     211 NNWLTNDLYCGYVSYKDKVHVKG------IHEPIISEEQFYRVQEIFSRMGK-NPNMNKE 263
                        + ++   GY +   K             EPI++ EQ   ++    +  +  P ++
Bxbl integrase   236 KRSMISEAMLGYATLNGKTVRDDDGAPLVRAEPILTREQLEALRAELVKTSRAKPAVSTP 295

LI integrase     264 SASLLNNLVVCSKCGLGFVHRRKDTVSRGKKYHYRYYSCKTYKHTHELEKCGNKIWRADK 323
                       SLL  ++ C+ CG           + G + H R Y C++         CGN      +
Bxbl integrase   296 --SLLLRVLFCAVCG-----EPAYKFAGGGRKHPR-YRCRSMGFPK---HCGNGTVAMAE 344

LI integrase     324 LEELIIDRVNNYSFASRNID--KEDELDSLNEKLKIEHAKKKRLFDLYINGSYE-----V 376
                      +    ++V +    +  ++-           +L   +A+   L  L  + +Y
Bxbl integrase   345 WDAFCEEQVLDLLGDAERLE-KVWVAGSDSAVELAEVNAELVDLTSLIGSPAYRAGSPQR 403

LI integrase     377 SELDSMMNDIDAQINYYEAQIEANEELKKNKKIQENLADLATVDFNSLEFREKQLYLKSL 436
                       LD+ +  + A+    E         +  +  Q              +   K  +L+S+
Bxbl integrase   404 EALDARIAALAARQEELEGLEARPSGWEWRETGQRFGDWWR-----EQDTAAKNTWLRSM 458

LI integrase     437 INKIYIDG---EQVTIEWL---------------------- 452
                       ++  D       TI++
Bxbl integrase   459 NVRLTFDVRGGLTRTIDFGDLQEYEQHLRLGSVVERLHTGM 499
```

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.



Extended Data Figure 3. a. Integrase variant library screening configuration with pBxb1 libraries where the helix, loop, or hairpin submotif has been randomized and are transformed along with pTarget plasmids where corresponding positions of the attB and attP target sites are altered. Primers with NNK codon usage were used for randomization of each residue indicated with an X to

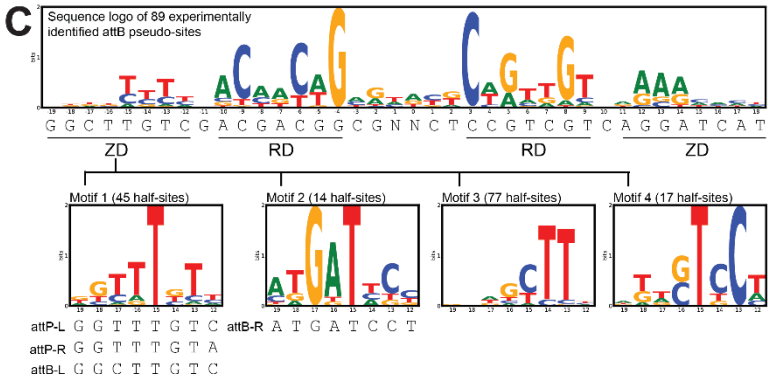
generate the Bxb1 variant libraries. For the hairpin library, SSK was used instead at residue 322 to bias the randomization to an A or P at this position. **b.** Example sequence logo plots summarizing enriched peptide motifs at amino acid residue positions 231-237 for helix selections with the corresponding helix DNA targets for each selection shown above each sequence logo. Position 235 was not randomized during selection and is shown in grey. Helices with an arginine (R) at position 237 were frequently observed in selections for a C at position -9 of the DNA target, helices and 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) at position 233 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 154-159 for each loop DNA target selection shown above each sequence logo. **d.** Example logo plots summarizing the results of selections at amino acid positions 314-325 for each hairpin DNA target selection shown above each sequence logo. Residues that were not randomized are shown in gray. **e.** Bxb1 variant specificity assay configuration where a single integrase variant previously isolated in the directed evolution assay is assayed for activity across DNA target libraries varied at the indicated positions with N. Each individual pTarget in the library contains the same modification at both half-sites of the attB and attP sites. **f.** Example DNA specificity plots of selected hairpin variants. Shown is the mean of a total of eight replicates; four biological replicates and two technical replicates thereof. **g.** Example DNA specificity plots of selected loop variants. Shown is the mean of four biological replicates.



Extended Data Figure 4. See legend on next page.

B

Genomic coordinates	Target sequence	%TI	Source	Locus	Name in ED Fig. 5
chr12:104547649-104547686	GTGTATCTGACACCAGCACACTCTGGTGTCAGAATCTT	6.92	Experimental	CHST11	
chr10:106891939-106891976	CCTTCTTCTACCACAGCGTTCTCAATTGGGTAAGCATC	6.27	Experimental	SORCS1	
chr3:194380585-194380622	TTGCTCCTCACAGCAGAGGACTCTGAAGACAGGAGCAT	2.45	Computational	chr3 intergenic	Chr3
chr20:32457262-32457299	CCTGCTTGCTCAGCTGAGACGGCAGAGGCTGAAGCTGC	2.11	Experimental	NOL4L	
chr1:25477421-25477458	GCCCCTTCTCCTACAGAGCAAGCAGCAGGGTAAATTCT	1.59	Experimental	MACO1	MACO1
chr3:157294091-157294128	TGGTACTATACTACAGAGAATGCTGATGTTAAAGAGCT	1.49	Experimental	VEPH1	
chr2:118918327-118918364	CTGATACTTGCAACTGAACACACTGTAGTCCACACACC	1.29	Computational	chr2 intergenic	
chr8:47341627-47341664	TGGATCTCCACCCCCTGTGTTTGCTGTTGTCTAAGTTTT	1.13	Experimental	SPIDR	
chr6:87076243-87076280	CCAGCCTGGGCAACAGAGCACTCTGTTGTCAAAAAAAA	1.09	Experimental	chr6 integenic	
chr5:179235158-179235195	TTCTTTTAAACCACTGATCATGCAGAAGTATAAGCTCA	0.85	Experimental	ADAMTS2	
chr1:84080777-84080814	AAGATCCTGACCTCCGAAGAAGCTGGAGTTGAGATCAT	0.64	Computational	chr1 integenic	
chrX:50461704-50461741	GTTTTGTTCACTGTAGAATCCTCAGTGGTTAGCACAAT	0.49	Computational	DGKK	
chr22:20778050-20778087	TTGCTCCTGACAGCAGTGGGAGCTATTGTCTAAGAGAT	0.43	Experimental	SERPIND1	
chr7:129402069-129402106	GAAACTTACAACACAGAGTTTCCAGTTGTTCAATCTCT	0.39	Experimental	AHCYL2	
chr3:37249477-37249514	CAATCTTGGCTCACTGCAACCTCTGCAGTTGAAGCAAT	0.36	Experimental	GOLGA4	
chr18:33954341-33954378	CAGATCTATACGGCTGACTACACAGTGGTGAGAACCAT	0.34	Computational	NOL4	
chr1:114204105-114204142	ATGATTCTAGGCACTGGGAACACAGTGGTGATCAAACC	0.32	Computational	chr1 integenic	Chr1
chr5:179764011-179764048	TGAGTCCTGACTGCAGCACATGCAGGTGAGGAAGCAGT	0.26	Experimental	MAML1	
chr19:48970999-48971036	GGCATTCCCATAACCGTGCACTCAGCTGCCGCAAGGCA	0.26	Experimental	GYS1	GYS1
chr1:207885851-207885888	TGTTTGGCCCCAACTGCCTCTGCCACTGTCGACACACT	0.25	Computational	CD34	
chrX:103694813-103694850	ATGTTCCTCACCACAGCTAACCCAGTCTTCAAAACTCA	0.25	Experimental	chrX intergenic	
chr10:95438018-95438055	GGGTTTTTAAACACTGACTTGCCTGGTGTGAGGATCCA	0.20	Computational	SORBS1	Chr10
chr11:58774495-58774532	AGAAATAGTACAACAGCTAACGCAGATGTTAAAAATCA	0.12	Experimental	chr11 intergenic	

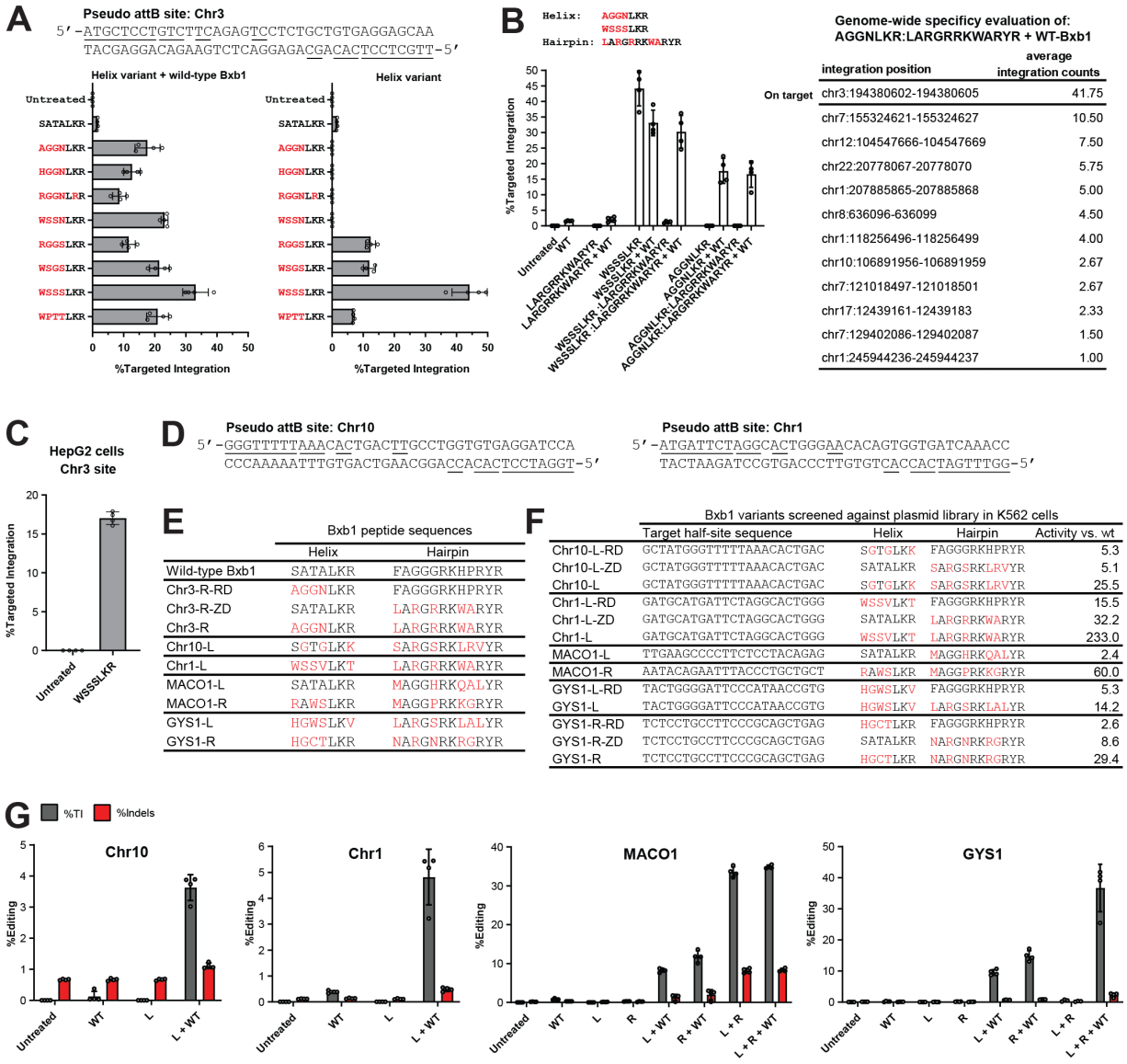


D

Pseudo-site	Motif	Half-site sequence	Pseudo-site	Motif	Half-site sequence
chr2 intergenic (R)	1	GGTGTGTGGACTACAGTG	SORCS1 (L)	3	CCTTCTTCTACCACAGCG
DGKK (L)	1	GTTTTGTTCACTGTAGAA	SORCS1 (R)	3	GATGCTTACCCAATTGAG
chr1 intergenic (R)	1	GGTTTGATCACCACTGTG	NOL4L (L)	3	CCTGCTTGCTCAGCTGAG
CD34 (L)	1	TGTTTGGCCCCAACTGCC	NOL4L (R)	3	GCAGCTTCAGCCTCTGCC
CHST11 (R)	2	AAGATTCTGACACCAGAG	chr3 intergenic (L)	4	TTGCTCCTCACAGCAGAG
chr2 intergenic (L)	2	CTGATACTTGCAACTGAA	chr3 intergenic (R)	4	ATGCTCCTGTCTTCAGAG
SPIDR (L)	2	TGGATCTCCACCCCTGTG	DGKK (R)	4	ATTGTGCTAACCACTGAG
chr1 intergenic (R)	2	ATGATCTCAACTCCAGCT	SERPIND1 (L)	4	TTGCTCCTGACAGCAGTG

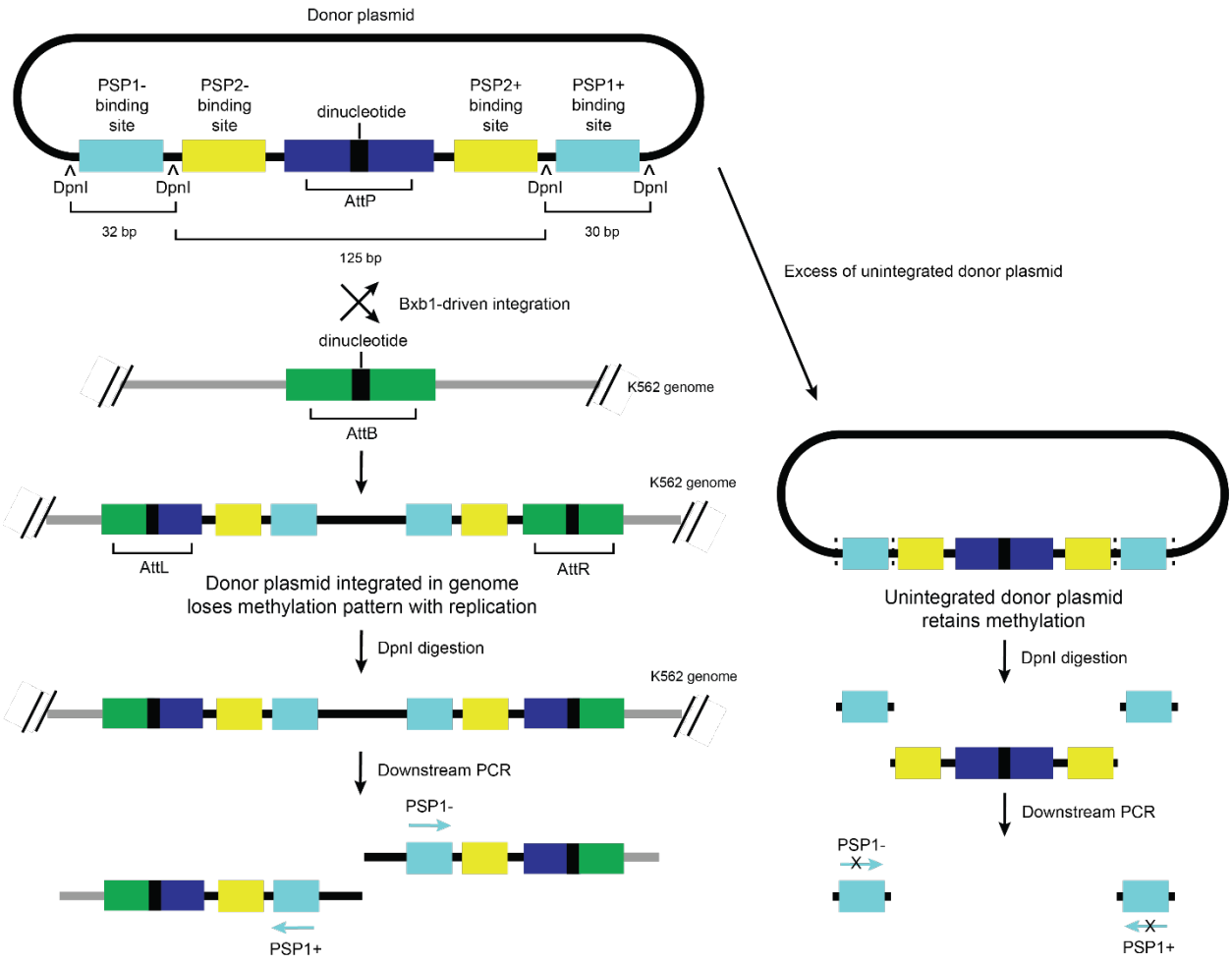
Extended Data Figure 4. Identification of Bxb1 pseudo-sites in the human genome using an unbiased genome-wide specificity assay. a. Schematic of the unbiased genome-wide specificity assay used in this study to experimentally identify potential Bxb1 integration sites in the human genome. Bxb1 drives integration of a donor plasmid at the intended on-target and pseudo-sites throughout the K562 genome. DNA is then extracted, sheared and adapters are ligated. Both genome-plasmid junctions are amplified separately using a 2-step PCR with plasmid-specific primers (PSP1 and 2) from either upstream or downstream of the original AttP site (+ or -). Then genomic integration sites are identified using NGS and a custom bioinformatics pipeline. **b.** List of 23 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 TI using a PCR-based NGS assay. CD34 and Chr1 pseudo-sites were identified in an independent computational search designed to identify targets for Bxb1 variants; control experiments indicated that these sites were better targeted with wild-type Bxb1. **c.** Sequence logo for all 89 potential pseudo-sites with at least two independent integrations identified using the experimental assay shown in panel **a**. Positions -19 to -12 of the 178 half-sites of these 89 potential pseudo-sites are a mixture of four different distinct DNA motifs

1244 and sequence logos for these four different motifs are shown below the sequence logo of the full
1245 sites. **d.** Examples of half-sites from validated human pseudo-sites for wild-type Bxb1 that
1246 correspond to each of the four DNA sequence motifs shown in panel **c**. Positions in each half site
1247 that correspond to key bases in the relevant motif are underlined.

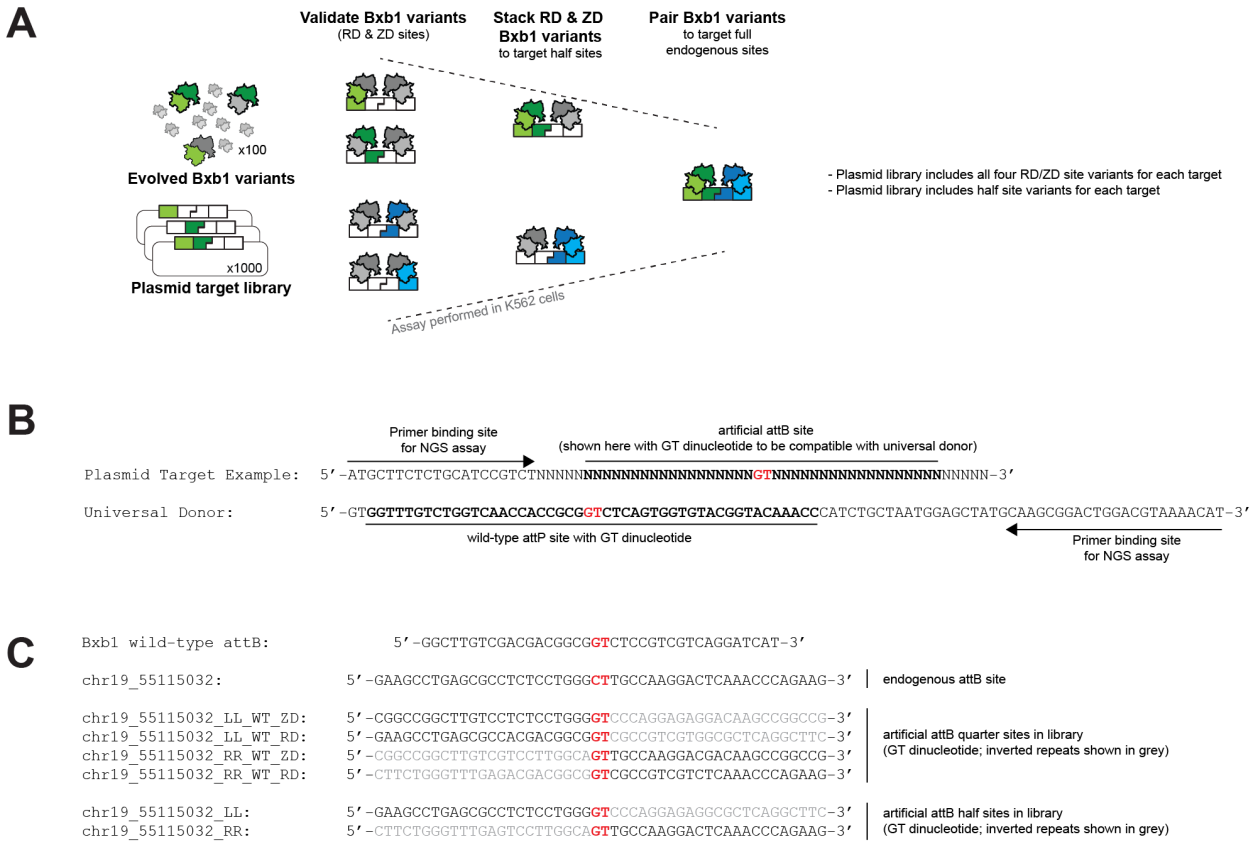


Extended Data Figure 5. Performance of Bxb1 variants against pseudo-sites in the human genome. a. The attB pseudo-site on chromosome 3 used for initial targeting and performance measurements of various Bxb1 helix domain variants against this target site using a PCR-based NGS assay. SATALKR is the wild-type helix peptide sequence and the target DNA sequence is shown as the reverse complement of the sequence in Extended Data Fig. 4b to make it easier to visualize the region we are targeting in this experiment. Selected helix sequences that match enriched motifs and contain an S or T at position 234 were active at the chromosome 3 target even in the absence of wild-type Bxb1 indicating that such helices can interact with both half-sites in the genomic target as well as the attP target in the donor. Similar sequences were enriched in selections with other target sites, so we concluded that these types of helix sequences represent helices with poor specificity, and we computationally removed similar sequences from subsequent 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. DNA targeting specificity characterization of this helix indicated a dramatic shift in specificity towards a T at position -10 (Figure 1f) consistent with the GTC sequence at positions -11 to -9 of the half-site we

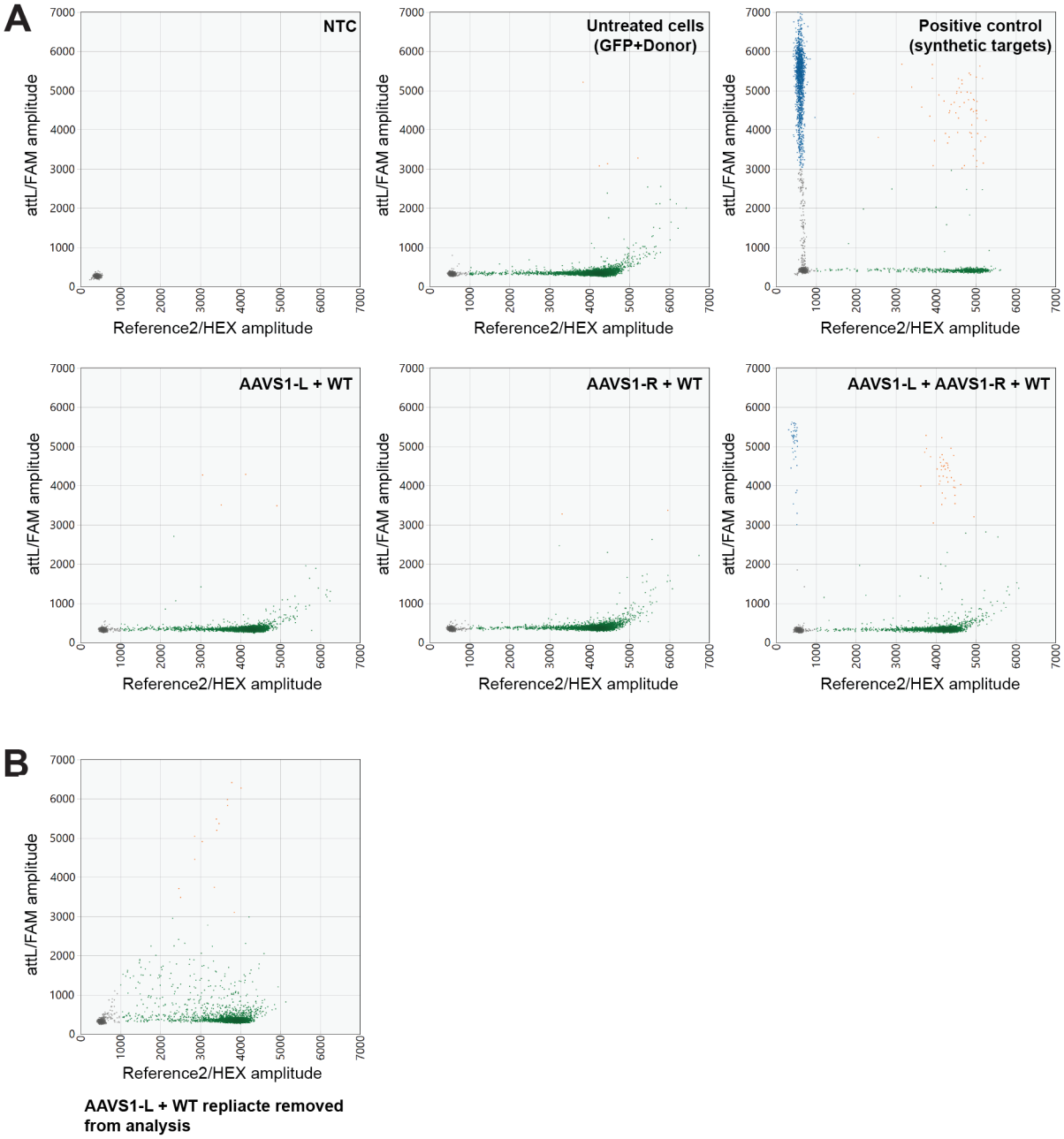
1264 targeted with this helix selection **b.** On-target performance of a Bxb1 variant with combined helix
1265 and hairpin variations, and data summary of a genome-wide specificity evaluation using a
1266 modified version of the assay described in Extended Data Fig. 4 (see Extended Data Fig. 6). Here,
1267 we combined the most promising selected hairpin LARGRRKWARYR with both the AGGNLKR
1268 helix and the more active, but less specific WSSSLKR helix. Both combinations retained full
1269 activity at the chromosome 3 site when combined with wild-type Bxb1 but show a substantial
1270 decrease without wild-type Bxb1. Thus, WSSSLKR in combination with LARGRRKWARYR
1271 now requires wild-type Bxb1 to achieve full integration activity. This indicates improved
1272 specificity vs. the WSSSLKR helix variant alone which is able to interact with both half sites of
1273 the chromosome 3 target as well as the attP site in the donor in the absence of wild-type Bxb1.
1274 Since both AGGNLKR and LARGRRKWARYR were able to target the intended chromosome 3
1275 half-site more specifically than the wild-type Bxb1 helix and hairpin, we reasoned the combination
1276 of this helix and hairpin should be even more specific for the chromosome 3 site and characterized
1277 this combination of helix and hairpin using an unbiased genome-wide specificity assay. This fully
1278 engineered Bxb1 variant had fewer candidate integration sites within the human genome than wild-
1279 type Bxb1 and we observed that the intended site on chromosome 3 had higher levels of
1280 integrations than any other site in the human genome. Note that the indicated Bxb1 variant has to
1281 be mixed with wild-type Bxb1 (included to bind the attP site on the donor) in order to be active at
1282 the intended chr3 target site so the genome-wide specificity assay was performed with a mixture
1283 of the indicated Bxb1 variant and wild-type Bxb1. The experiment was performed with a pool of
1284 16 donor constructs containing all possible central dinucleotides. Average integration counts were
1285 determined using the average deduplicated read numbers across 2 samples using primers that
1286 amplify the attL and attR plasmid-genome junctions. The corresponding control sample where
1287 GFP was substituted for Bxb1 resulted in 0 sites (Supplementary Table 21) **c.** Activity-testing of a
1288 Bxb1 helix variant from panel a. in human HepG2 cells. **d.** Sequence of two additional Bxb1
1289 pseudo-sites in the human genome. **e.** Bxb1 peptide sequences of evolved Bxb1 variants that
1290 showed improved performance against the half-sites of pseudo-sites shown in **Figure 2b** and panel
1291 c. **f.** Screening data using synthetic DNA targets tested in human K562 cells that was used to
1292 identify the constructs shown in panel d. Activity is determined by the number of DNA sequence
1293 reads corresponding to recombined versions of each synthetic target; activity is normalized to the
1294 activity of wild-type Bxb1 alone against the same synthetic target site. **g.** Results from a PCR-
1295 based NGS assay demonstrating improved performance of evolved Bxb1 variants against their
1296 chromosomal endogenous targets in K562 cells. The presence of a wild-type Bxb1 expression
1297 construct is necessary to bind the wild-type attP sequence on the donor plasmid. Data are presented
1298 as the mean ± s.d. from four biological replicates.



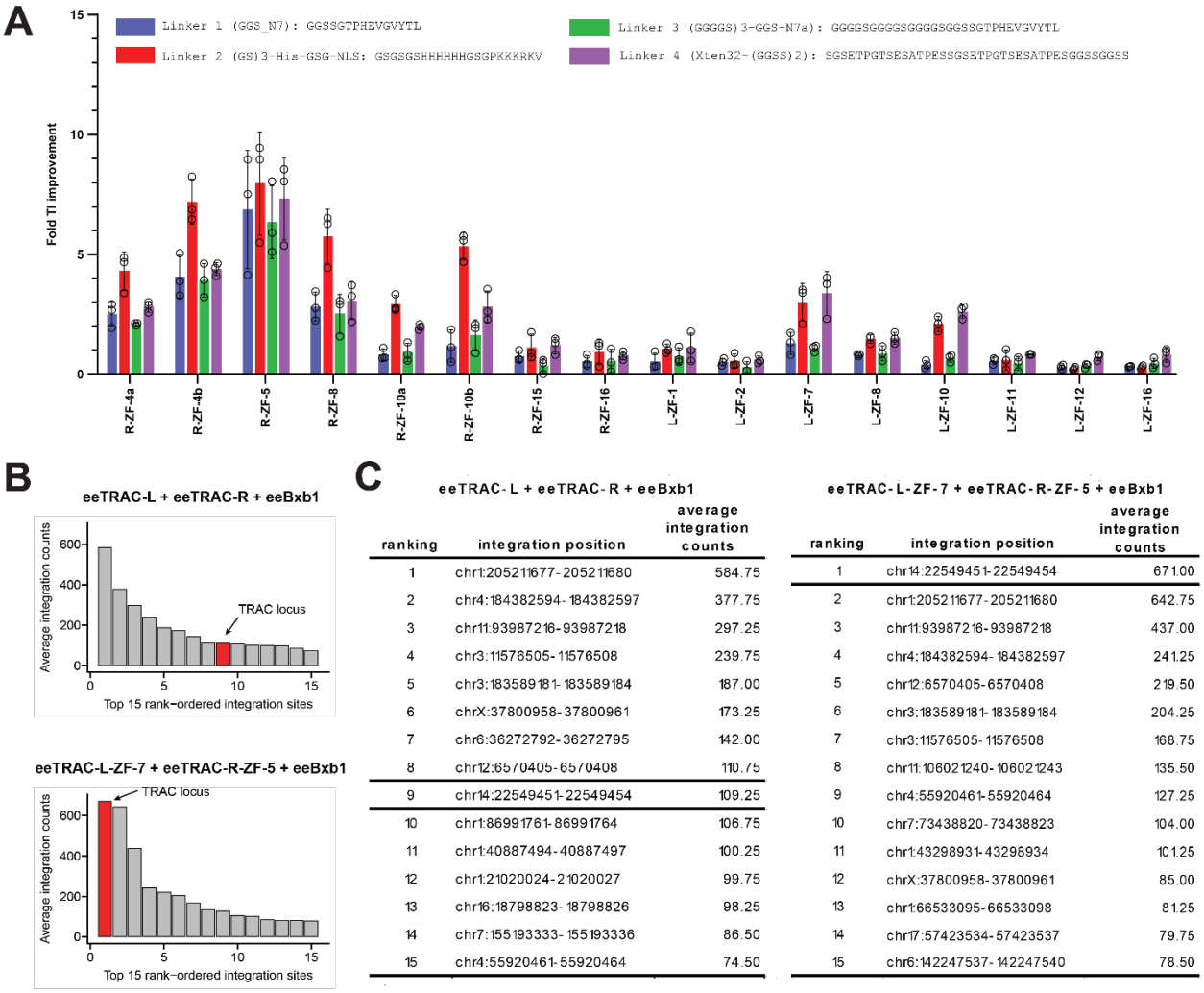
Extended Data Figure 6. Improved genome-wide specificity assay. Schematic of the modified unbiased genome-wide specificity assay used in this study to experimentally nominate candidate integration sites shown in Extended Data Fig. 5b. Strategically placed DpnI recognition sites in the donor molecule supports the enzymatic removal of excess unintegrated donor plasmid since these sites should lose their bacterial methylation marks after integration and cell replication in the human genome. The DpnI removal method resulted in a substantial reduction of donor plasmid-derived background signal and enabled culturing the cells for less time compared to previous method where many cell passages were required to dilute out excess plasmid.



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 3**. 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. 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**.**



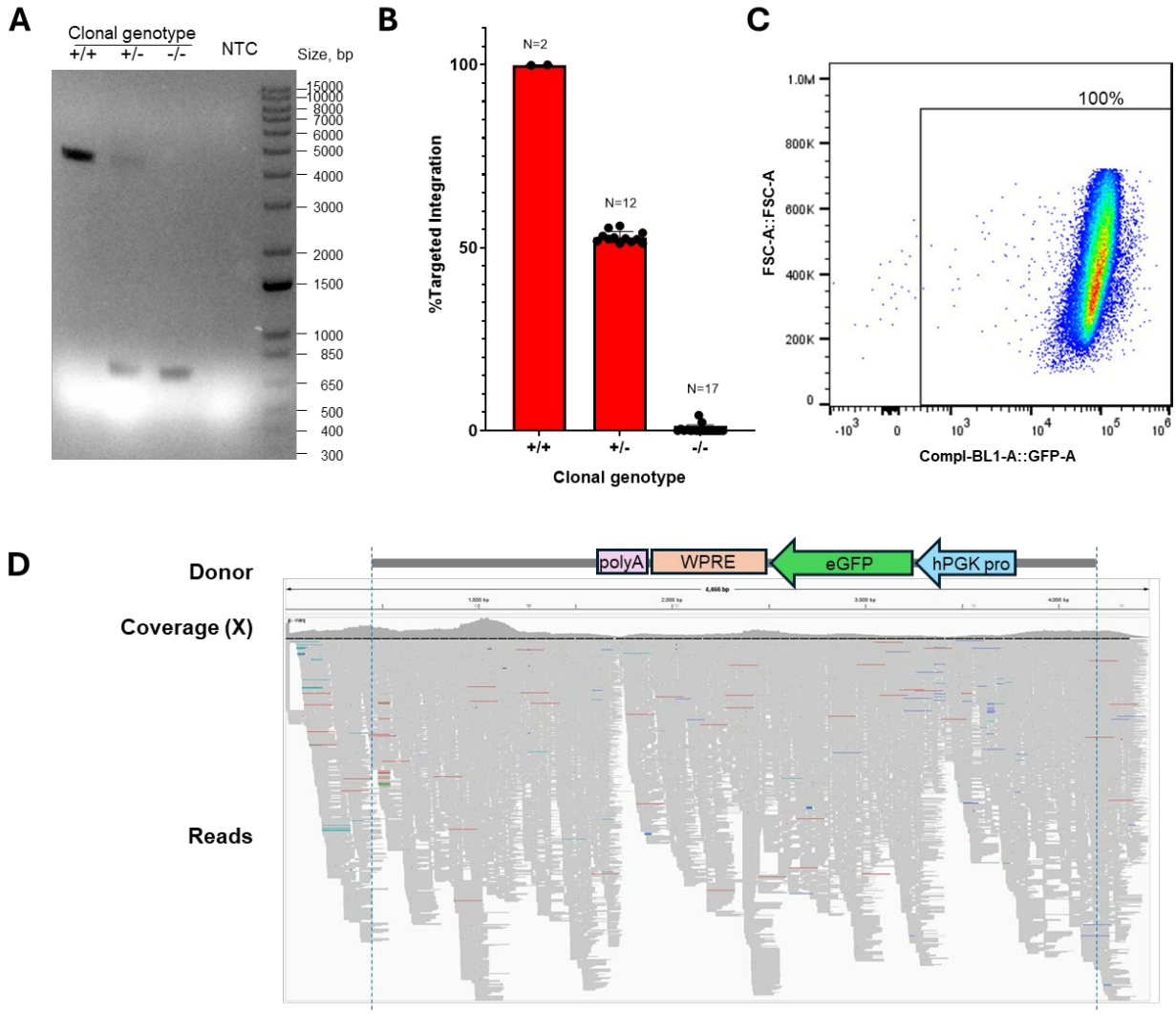
eeTRAC-L + eeTRAC-R + eeBxb1

ranking	integration position	average integration counts
1	chr1:205211677-205211680	584.75
2	chr4:184382594-184382597	377.75
3	chr11:93987216-93987218	297.25
4	chr3:11576505-11576508	239.75
5	chr3:183589181-183589184	187.00
6	chrX:37800958-37800961	173.25
7	chr6:36272792-36272795	142.00
8	chr12:6570405-6570408	110.75
9	chr14:22549451-22549454	109.25
10	chr1:86991761-86991764	106.75
11	chr1:40887494-40887497	100.25
12	chr1:21020024-21020027	99.75
13	chr16:18798823-18798826	98.25
14	chr7:155193333-155193336	86.50
15	chr4:55920461-55920464	74.50

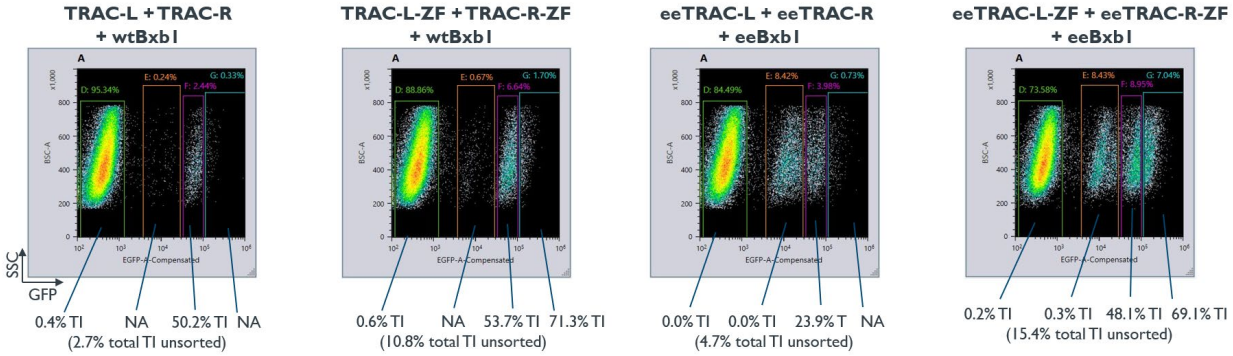
eeTRAC-L-ZF-7 + eeTRAC-R-ZF-5 + eeBxb1

ranking	integration position	average integration counts
1	chr14:22549451-22549454	671.00
2	chr1:205211677-205211680	642.75
3	chr11:93987216-93987218	437.00
4	chr4:184382594-184382597	241.25
5	chr12:6570405-6570408	219.50
6	chr3:183589181-183589184	204.25
7	chr3:11576505-11576508	168.75
8	chr11:106021240-106021243	135.50
9	chr4:55920461-55920464	127.25
10	chr7:73438820-73438823	104.00
11	chr1:43298931-43298934	101.25
12	chrX:37800958-37800961	85.00
13	chr1:66533095-66533098	81.25
14	chr17:57423534-57423537	79.75
15	chr6:142247537-142247540	78.50

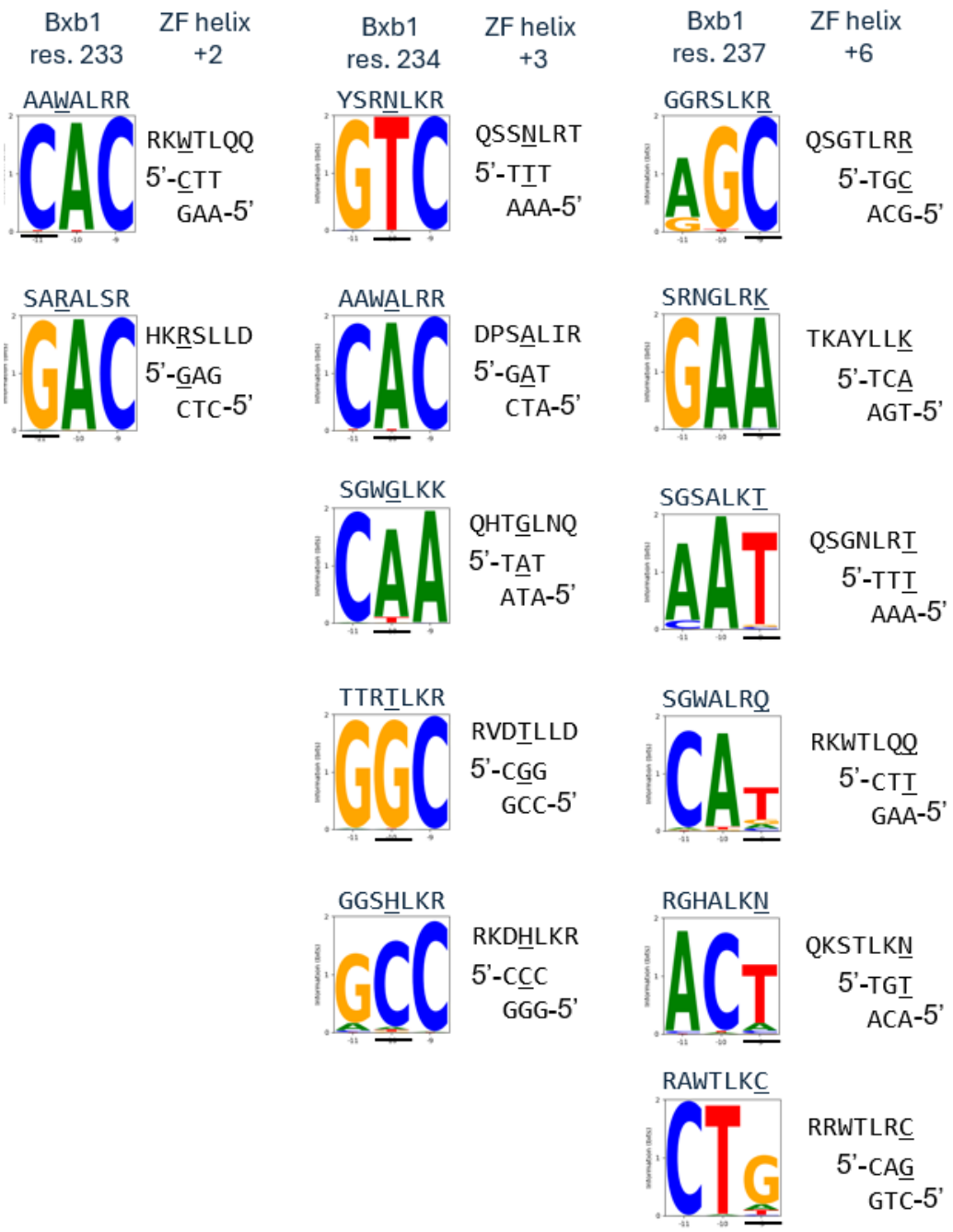
Extended Data Figure 9. ZF and activity-increasing TRAC Bxb1 construct optimizations. a. Four linkers with different lengths were tested in the context of the constructs highlighted in **Figure 4**. Linker 2 was chosen for all subsequent experiments. **b.** Results from an unbiased genome-wide specificity analysis of the indicated constructs that contain the eeBxb1 mutations. The top 15 candidate integration sites are shown (out of total 2637 sites for eeTRAC constructs and 1949 for ZF-fused eeTRAC constructs) with the on-target at the TRAC locus indicated between bolded lines. The entire datasets are provided in Supplementary Tables 18 and 19 and corresponding control datasets in Supplementary Tables 17 and 20. Average integration counts were determined using the average deduplicated read numbers across 2 replicates and 2 samples each using primers that amplify the attL and attR plasmid-genome junctions. **c.** Corresponding table format of data in b with candidate integration site information.



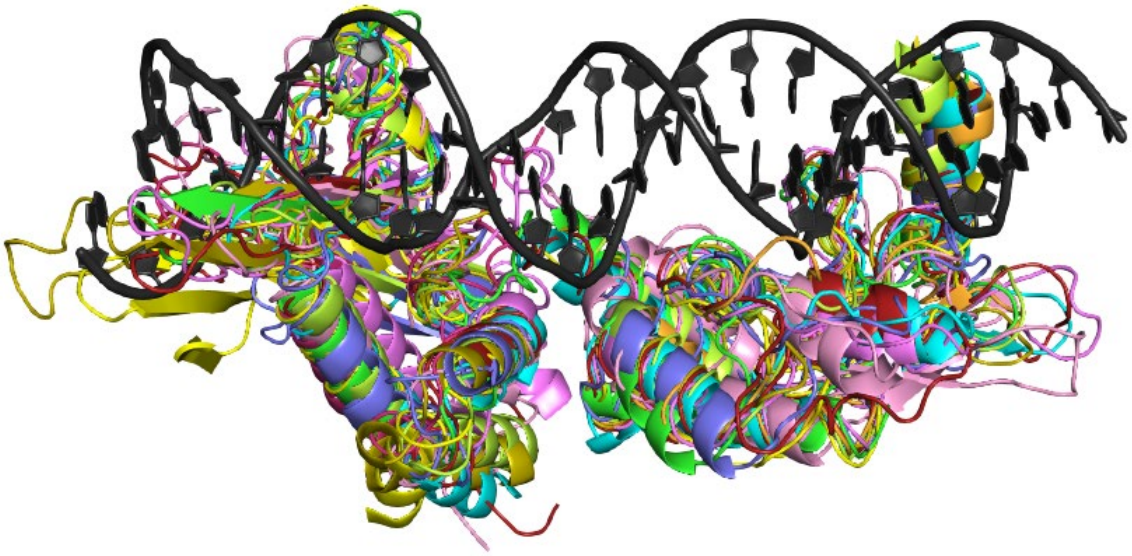
Extended Data Figure 10. Targeted integration characterization in GFP positive single-cell clones. a. Representative results of PCR analysis of different types of single-cell clones. Note, the expected PCR product with full length TI, including a single copy donor and adjunct genomic sequence, is 4,465 bp, and the wild-type amplicon is 723 bp. NTC, no template control. P-ctrl, positive control. **b**. Results for all GFP-positive clones analyzed in this experiment. Each clone was characterized using our NGS TI assay and the %TI values are consistent with the results of the PCR assay. **c.** Flow cytometry of GFP expression in a biallelic clone. **d.** Nextera sequencing of PCR amplicon from the biallelic clone from panel c with minimum 500X coverage confirms the full-length integrated donor in the eeBxb1-ZFP biallelic clone.



Extended Data Figure 11. Targeted integration percentages in MINT-treated T cells sorted based on GFP expression levels. Nucleofected cells were sorted based on GFP expression into the 4 fractions indicated on the flow cytometry plots and TI percentages in the sorted fractions were determined. The percentage of cells in the different GFP expressing subpopulations is listed in the plots, the TI numbers are listed below the flow cytometry plots, with the number determined for the unsorted sample listed at the bottom in parentheses. NA indicates samples with too low of a cell number to yield a reliable result.



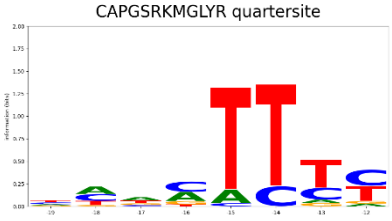
Supplementary Figure 1. Similar Bxb1 helix and ZF helix interaction patterns with DNA. Bxb1 data from **Figure 1f** and ZF data from Ichikawa *et al.* (2023) except for QHTGLNQ which is from Wolfe *et al.* (2001). ZF data shown with primary DNA contacting strand on bottom to match Bxb1 data (QSGTLRR would bind 5'-GCA is standard ZF notation). Similar residue-base correlations for the Bxb1 helix and the ZF helix are underlined. Bxb1 helix and ZF helix patterns usually do not match at Bxb1 residue 231 and ZF helix position -1.



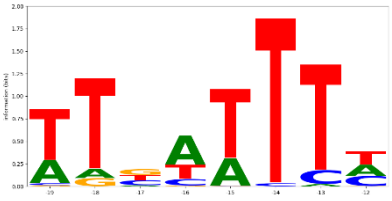
1364

1365 **Supplementary Figure 2. Structural alignment of predicted structures for RD and ZD**
1366 **domains of various LSIs.** Shown are Pa01, Kp03, Nm60, Si74, BcyInt, and SscInt with the
1367 predicted structure of Bxb1. The structures of the predicted domains align well enough to identify
1368 the loop, helix, and hairpin regions of these LSIs as well as any LSIs that have a similar or better
1369 structural alignment with the predicted Bxb1 structure. Directed evolution using a residue
1370 randomization scheme at the same residues of the loop, helix, and hairpin of any such LSR should
1371 enable the identification of LSR variants that can target desired DNA sequences.

CAPGSRKMGLYR quartersite

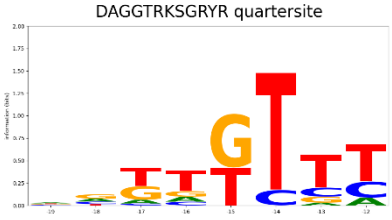


CAPGSRKMGLYR halfsite



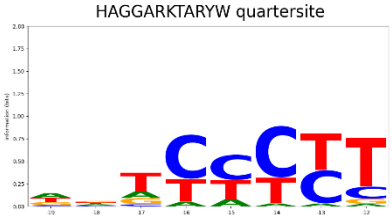
1372

DAGGTRKSGRYR quartersite

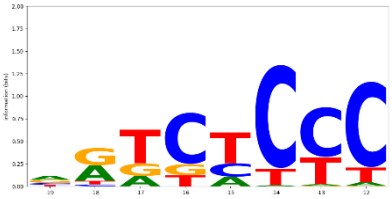


1373

HAGGARKTARYW quartersite

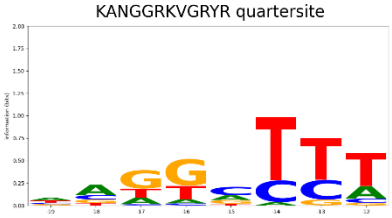


HAGGARKTARYW halfsite

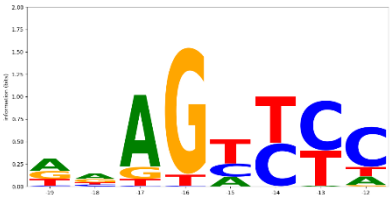


1374

KANGGRKVGRYR quartersite

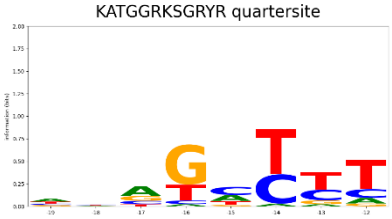


KANGGRKVGRYR halfsite

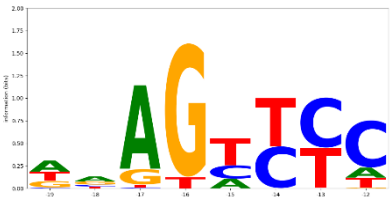


1375

KATGGRKSGRYR quartersite

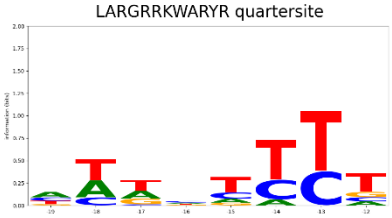


KATGGRKSGRYR halfsite

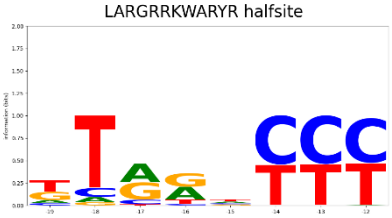


1376

LARGRRKWARYR quartersite

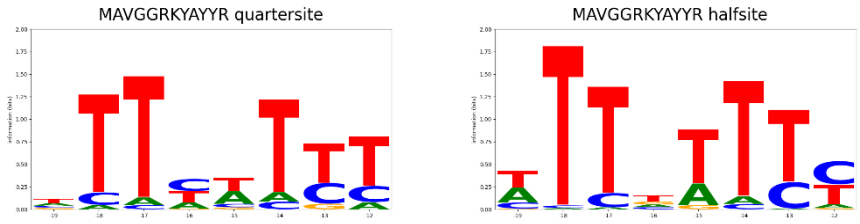


LARGRRKWARYR halfsite

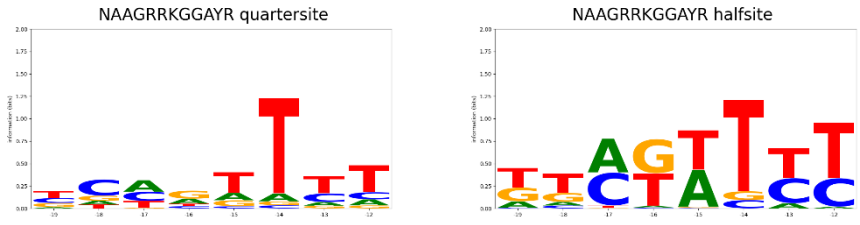


1377

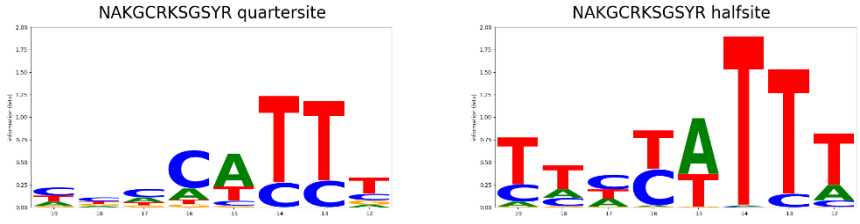
1378



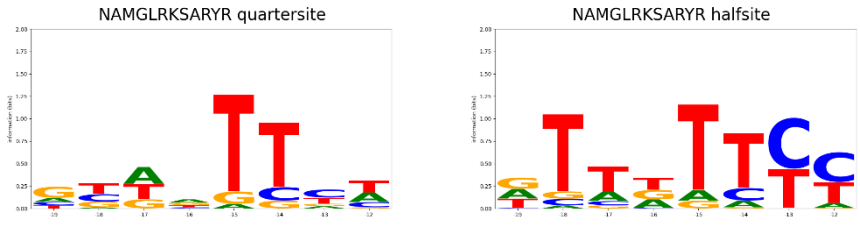
1379



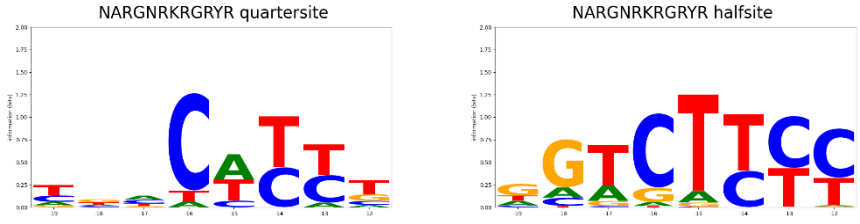
1380



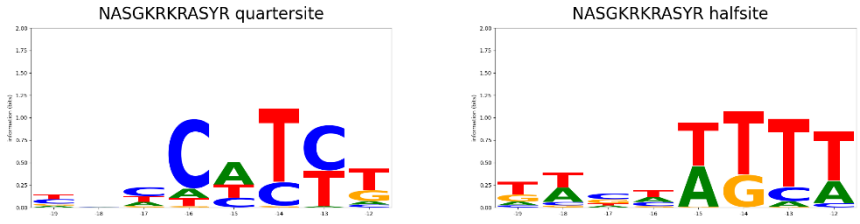
1381

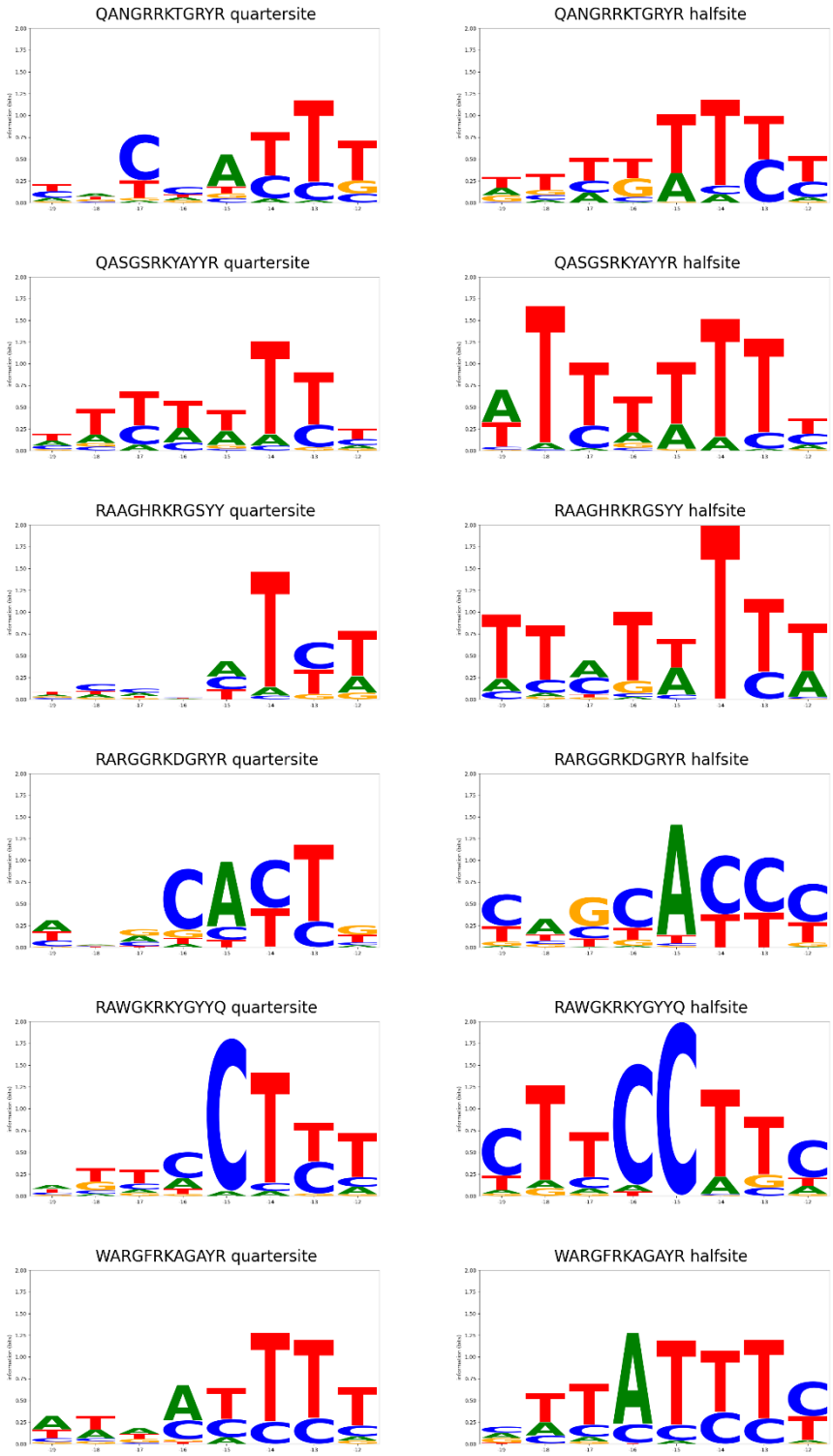


1382



1383

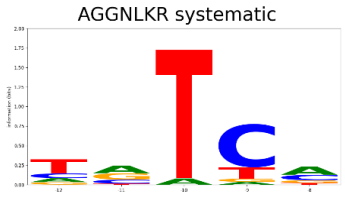




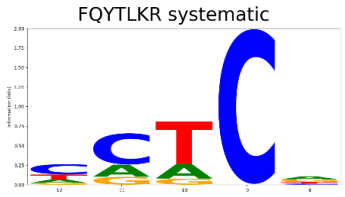
Supplementary Figure 3. DNA target site preferences of engineered Bxb1 hairpins. DNA target site preferences for 18 engineered Bxb1 hairpins as measured in human K562 cells for recombination of a donor plasmid with two different libraries of DNA target site plasmids are plotted. Note that binding site preference data was not obtained for 8 hairpins in the hairpin archive table because these hairpins were only tested in combination with helix variants to target the

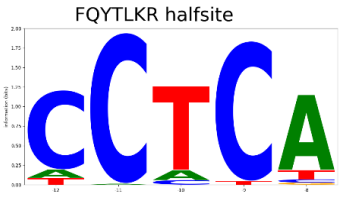
1395 endogenous sites indicated in the table. The "quarter-site" and "half-site" libraries are derived from
1396 500 potential human targets for Bxb1 variants within disease related genes. The hairpin quarter-
1397 site pool has bases at positions -19 to -12 in the left half-site of attB varied to match the
1398 corresponding portion of the relevant target half-site (the right half-site of attB bears corresponding
1399 mutations at +12 to +19). Wild-type Bxb1 was included in the experiment to enable use of a single
1400 donor plasmid bearing a wild-type Bxb1 attP target so only sequences with read fractions at least
1401 3 standard deviations above the read fraction for the wild-type Bxb1-only control at the same target
1402 sequence were analyzed. The sequence motif of the weighted average of the top 20 DNA targets
1403 at least 3 standard deviations above the wild-type-only control for positions -19 to -12 are plotted.
1404 If fewer than 20 sequences are above this cut-off then all successful sequences are plotted unless
1405 a single half-site represents more than half of the weighted average.

1406

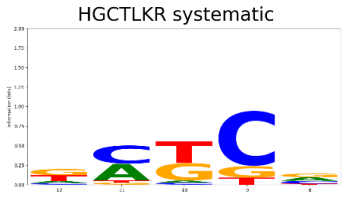


1407

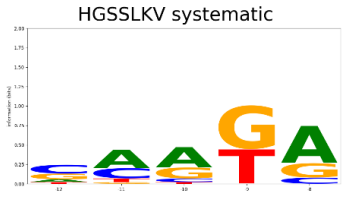




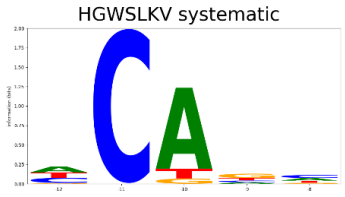
1408

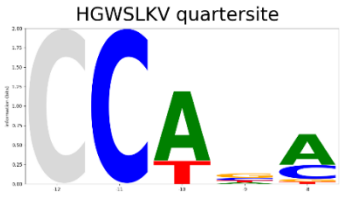


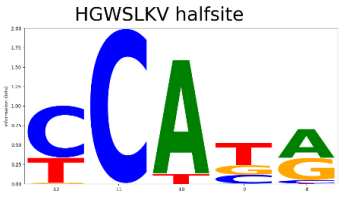
1409



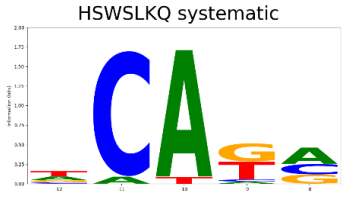
1410



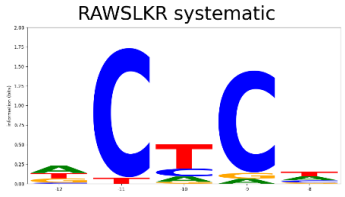




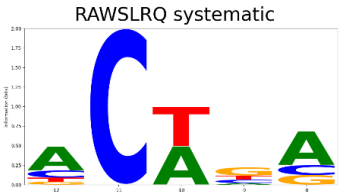
1411

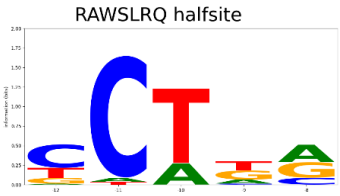


1412

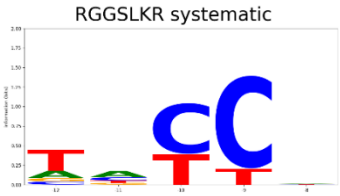


1413

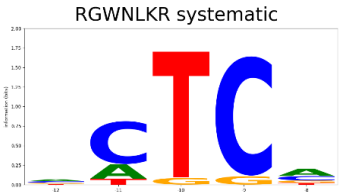




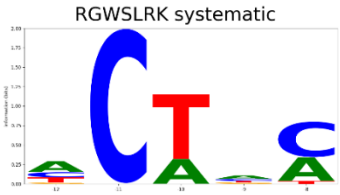
1414

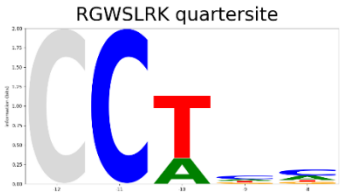


1415

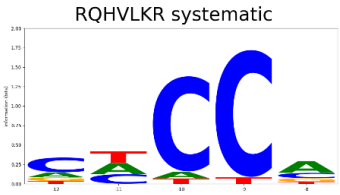


1416

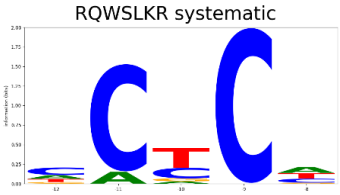


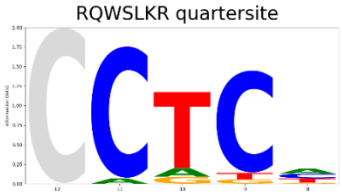


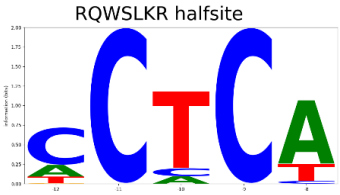
1417



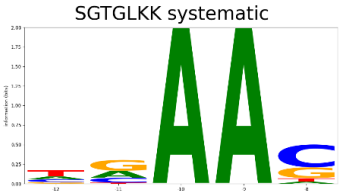
1418

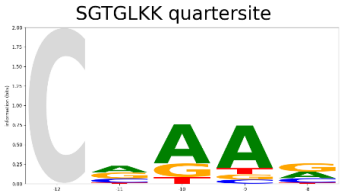


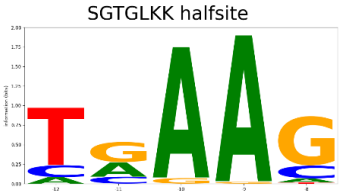




1419

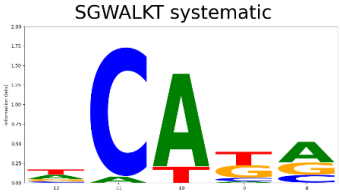






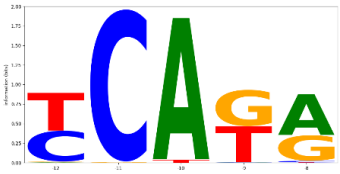
1420

SGWALKT systematic



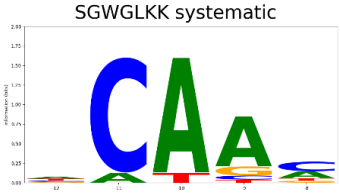
SGWALKT quartersite

SGWALKT halfsite



1421

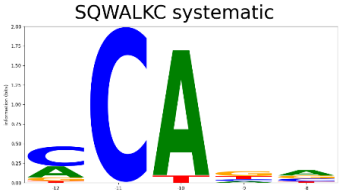
SGWGLKK systematic



SGWGLKK quartersite

1422

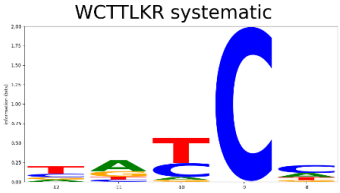
SQWALKC systematic



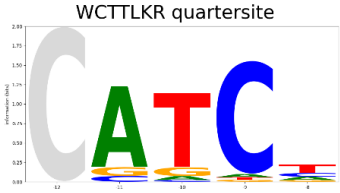
SQWALKC quartersite

1423

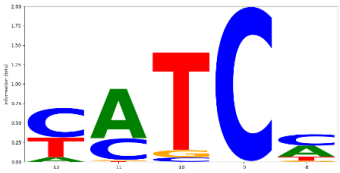
WCTTLKR systematic



WCTTLKR quartersite

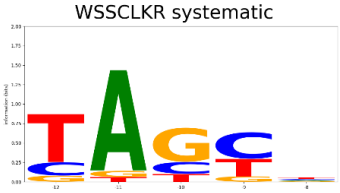


WCTTLKR halfsite



1424

WSSCLKR systematic



WSSCLKR quartersite

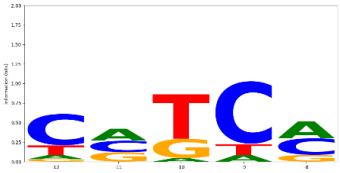
WSSCLKR halfsite

1425

WSSSLKR systematic

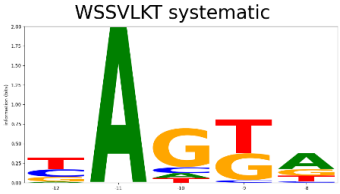
WSSSLKR quartersite

WSSSLKR halfsite

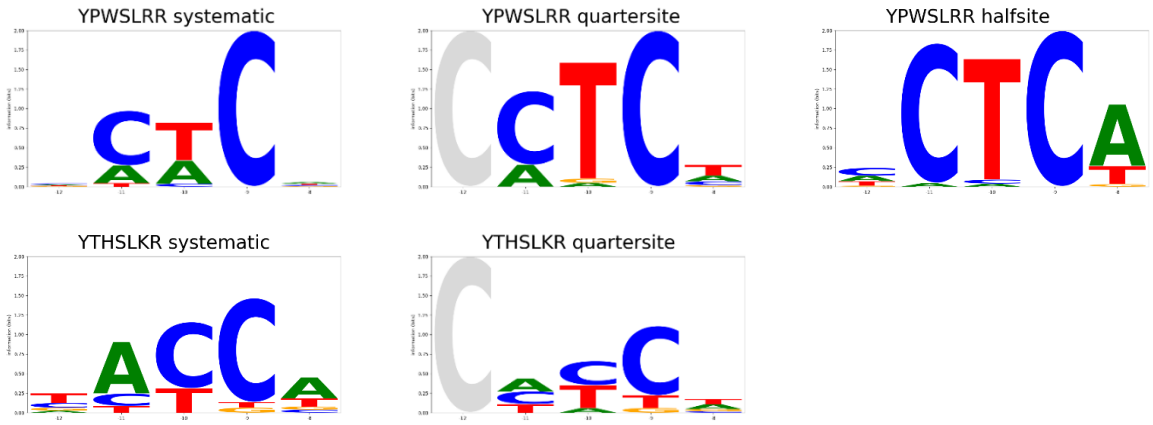


1426

WSSVLKT systematic



WSSVLKT quartersite



Supplementary Figure 4. DNA target site preferences of engineered Bxb1 helices. DNA target site preferences for 23 engineered Bxb1 helices as measured in human K562 cells for recombination of a donor plasmid with three different libraries of DNA target site plasmids are plotted. The "systematic" target site library has bases at positions -12, -11, -10, -9, and -8 in the left half-site of attB fully randomized (the right half-site of attB bore corresponding mutations at +8, +9, +10, +11, and +12). Note that YGSALKQ was not characterized with these DNA target pools because it was only tested in combination with hairpin variants to target the endogenous TRAC locus. The "quarter-site" and "half-site" libraries are derived from 500 potential human targets for Bxb1 variants within disease related genes. Each member of the helix quarter-site library contains the sequence from -11 to -2 from one half-site of the 500 potential target sites (1000 library members in total) and an inverted repeat of this sequence at positions +2 to +11. The gray C indicates that the C at position -12 wasn't varied in this library. The half-site libraries contain entire half-sites (-19 to -2) from these 500 potential target sites (1000 half-sites in total) with an inverted repeat of the same half-site from +2 to +19. All target sites in the library have the GT central dinucleotide (positions -1 and +1) from the wild-type attB target site in order to successfully recombine with the wild-type Bxb1 attP site in the donor. Wild-type Bxb1 was included in the experiment to enable use of a single donor plasmid bearing a wild-type Bxb1 attP target so only sequences with read fractions at least 3 standard deviations above the read fraction for the wild-type Bxb1-only control at the same target sequence were analyzed. The sequence motif of the weighted average of the top 20 DNA targets at least 3 standard deviations above the wild-type only control for positions -12, -11, -10, -9, and -8 are plotted. If fewer than 20 sequences are above this cut-off then all successful sequences are plotted unless a single half-site represents more than half of the weighted average.