Supporting Information

Li et al. 10.1073/pnas.1305987110

SI Materials and Methods

Construction and Testing of Zinc Finger Proteins. Several criteria have been put forth to define genomic safe harbors (1, 2). Ideally, a safe harbor should be distant from the 5′ end of a gene, and especially distant from any oncogene. Gene addition should be outside a transcriptional unit, including microRNAs, and outside ultraconserved regions of the human genome. The location must be accessible to allow the transposon and transposase to reach the target sequence, thereby promoting efficient integration. A transcriptionally active region would help to ensure that the DNA is accessible and may be required to ensure stable expression of the therapeutic inserted transgene. We chose human *ROSA26* and L-gulono-γ-lactone oxidase (*GULOP*) loci as two candidate safe harbors. We reasoned that, because mouse *Rosa26* is a target for many site-specific insertions of foreign DNA with no known adverse effects, the human *ROSA26* (3) also represents a safe harbor candidate. *GULOP* is a unitary pseudogene that is far distant from neighboring transcriptional units. In most nonhuman mammals, *GULOP* synthesizes the precursor of L-ascorbic acid (vitamin C); however, in humans, the majority of the gene has been deleted, and within the remaining sequence several anomalous nucleotide changes have occurred (4, 5). None of the genes flanking *GULOP* or *ROSA26* are known tumor suppressors or oncogenes. Neither candidate encodes a protein product, although *ROSA26* encodes a noncoding RNA.

To identify regions within these genes that are rich in *piggyBac* target sequence sites TTAA, we developed a scoring algorithm that analyzed TTAA density for indicated regions (Fig. S4). For each TTAA, the number of adjacent sites was determined within a given window. A 128-bp window on either side of each site was used; thus the score denotes the TTAA density within a 256-bp sliding window.

Six-finger zinc finger arrays were assembled using two-finger zinc finger units as previously described (6). Two-finger units, each expected to specify 6 bp of DNA, were chosen from three-finger zinc finger proteins (ZFPs) engineered by the oligomerized pool engineering method or used to practice the context-dependent assembly method (7, 8). Using these two-finger units, we assembled six-finger arrays targeted to TTAA-rich regions within the *ROSA26* and *GULOP* sites (Fig. S4).

Bacterial Two-Hybrid and Mammalian One-Hybrid Assays. The GULOP and human ROSA26 zinc finger proteins were assayed for activity using a bacterial two-hybrid–based reporter system (7, 8) (Fig. S5). β-Galactosidase assays for assessing the DNA-binding activities of zinc finger proteins in a bacterial two-hybrid assay were performed as described previously (8). Mammalian one-hybrid assays were performed as described previously (9) (Fig. S5). Briefly, the activation plasmids were constructed by inserting cDNA encoding a C-terminal fusion of the herpes simplex virus protein 16 activation domain and each of the engineered ZFPs into the BamHI/XhoI-digested pCAGGs backbone. The ZFP target reporter plasmid was constructed by annealing oligos containing four copies in tandem of the ZFP target sequence and cloning the annealed oligos upstream of a minimal human thymidine kinase promoter driving firefly luciferase in the pTATA vector (a kind gift from James Darnell, Laboratory of Molecular Cell Biology, The Rockefeller University, New York, NY). HeLa cells were transfected with 0.4 μg each of ZFP activator and target reporter plasmids using Lipofectamine 2000 (Invitrogen) as directed by the manufacturer. Transiently transfected cells were harvested in 1× Passive Lysis Buffer (Promega) after 48 h. Twenty-μl lysates were assayed using the Luciferase Reporter Assay System (Promega) according to the manufacturer's instructions. Based on the results of these assays, we selected the ZFPs termed ROSA3b and GULOP1b for further use in this study. The *ROSA26* target site is GATGCCTGGTAGGGATGCA (58% GC) and the *GULOP* target site is TGGGATGCAGCCAGATGAG (58% GC). The DNA sequences of the ZFPs are shown (Fig. S6).

Integration-Site Recovery for Illumina HiSeq2000 Sequencing. Integration sites were recovered as described (10). Briefly, HeLa cells (5×10^6) were transfected with 10 μg *pXL-BacII PB-GFP/Puro* transposon plasmid and 2 μg of each transposase plasmid, and then integrants were selected with puromycin (0.5 μg/mL) for 3 wk. Genomic DNA from three separate transfections was extracted from the integration library using the DNeasy tissue kit (Qiagen). Pooled DNA (2 μg) was digested overnight with ApoI or BstYI at 50 °C and 60 °C, respectively; DNA fragments were purified with the QIAquick PCR purification kit (Qiagen) and ligated to ApoI and BstYI linkers overnight at 16 °C. Nested PCR was carried out under stringent conditions using the transposon end-specific primers AAACCTCGATATACAGACCGATAAAACACATGCGTCAATTTTACGC (primary) and AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTXXXXCGTACGTCACAATATGATTATCTTTC (secondary; XXXX denotes bar code; underlined sequence indicates Illumina cluster-generation sequence) and linker-specific primers CGTAGGGAGCAAGCAGAAGACGG (primary) and CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT (secondary). DNA barcodes were included in the second-round PCR primers to track sample origin. The PCR products were gel-purified, pooled, and sequenced using the Illumina HiSeq2000 sequencing platform.

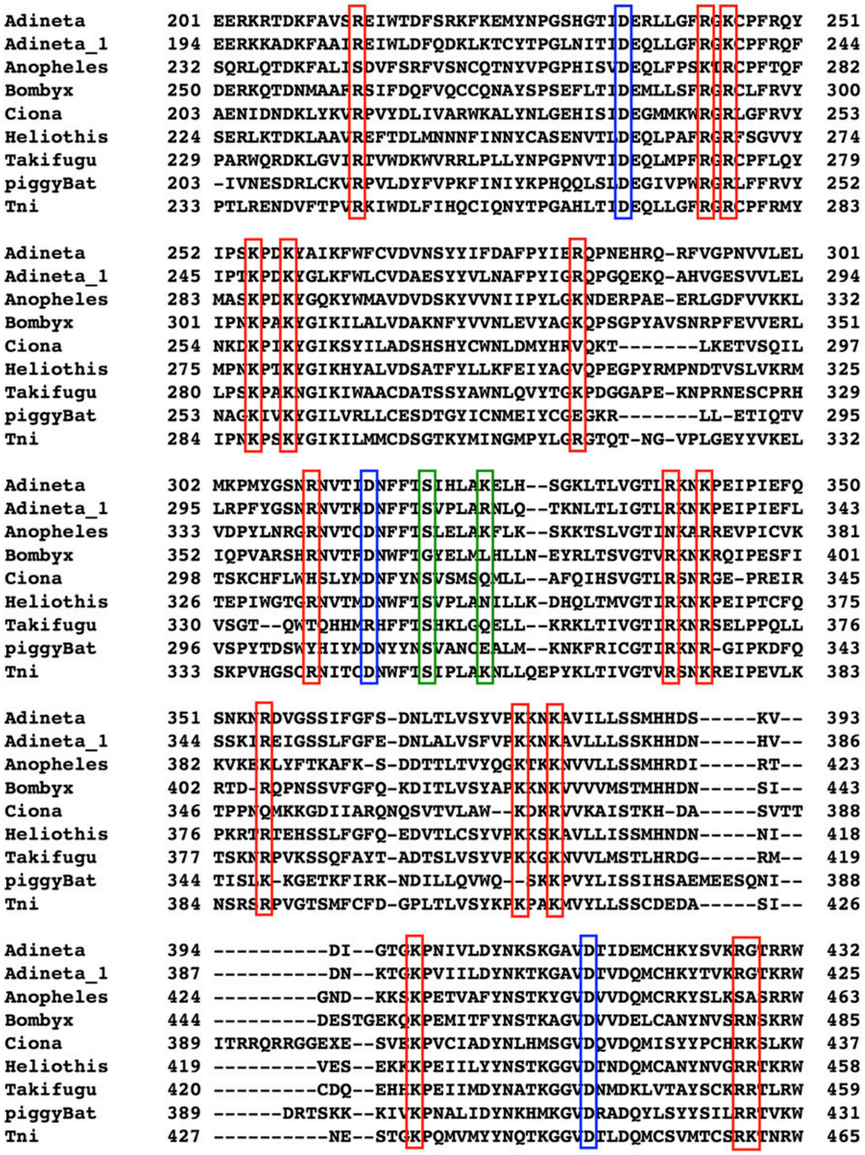
Reads from each flow cell lane were trimmed according to the barcodes and linkers expected, using a custom R wrapper for the BioStrings trimLRPatterns function (11) and allowing no mismatches in the barcode and up to two mismatches in the linker sequence. Trimmed reads were aligned to the hg18 human genome build using Bowtie (12), allowing two mismatches in each alignment and requiring the alignment to be unique.

Insertion-site coordinates were sorted and collapsed; multiple reads often mapped to a single site. Furthermore, many sites with large numbers of reads were immediately flanked by a few sites with one or two reads. Upon examination, these nearly always prove to be slight alignment errors. Thus, insertion counts in this configuration are collapsed into the site with the most counts, using a simple Perl script that scans for insertions mapping to adjacent positions. This leaves a set of sites, each associated with a number of mapped insertions. As we do not know whether multiple recovered insertions are real or are PCR artifacts, we proceed with the analysis using only the sites. For a subset of the sites, we have recovered insertions in both orientations (on the + and the − strand). These are necessarily independent events, and these "bidirectional" sites are noted separately.

For genome-wide feature correlation analysis, we could not include all sites, due to computational limitations. Thus, we included all bidirectional sites for each of the experiments in HeLa cells, as well as a randomly chosen subset of sites whose insertion counts were in the third quartile of the insertion counts for all sites, reasoning that these should be strong sites, yet representative of the insertion landscape for each experiment. After this process we had subsets of roughly 2,500–3,000 sites for each of the experiments.

Initial sites, insertion counts, and bidirectional status (0 if not bidirectional, 1 otherwise) are provided as supplemental -s; the files labeled "sites_analyzed" are those that were included in the genome-wide analysis and the others contain the full list of sites for each element. R and Perl scripts are available upon request.

1. Papapetrou EP, et al. (2011) Genomic safe harbors permit high β-globin transgene expression in thalassemia induced pluripotent stem cells. *Nat Biotechnol* 29(1):73–78.
2. DeKelver RC, et al. (2010) Functional genomics, proteomics, and regulatory DNA analysis in isogenic settings using zinc finger nuclease-driven transgenesis into a safe harbor locus in the human genome. *Genome Res* 20(8):1133–1142.
3. Irion S, et al. (2007) Identification and targeting of the ROSA26 locus in human embryonic stem cells. *Nat Biotechnol* 25(12):1477–1482.
4. Zhang ZD, Frankish A, Hunt T, Harrow J, Gerstein M (2010) Identification and analysis of unitary pseudogenes: Historic and contemporary gene losses in humans and other primates. *Genome Biol* 11(3):R26.
5. Inai Y, Ohta Y, Nishikimi M (2003) The whole structure of the human nonfunctional L-gulono-gamma-lactone oxidase gene—the gene responsible for scurvy—and the evolution of repetitive sequences thereon. *J Nutr Sci Vitaminol (Tokyo)* 49(5):315–319.
6. Moore M, Klug A, Choo Y (2001) Improved DNA binding specificity from polyzinc finger peptides by using strings of two-finger units. *Proc Natl Acad Sci USA* 98(4): 1437–1441.
7. Maeder ML, et al. (2008) Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification. *Mol Cell* 31(2):294–301.
8. Sander JD, et al. (2011) Selection-free zinc-finger-nuclease engineering by context-dependent assembly (CoDA). *Nat Methods* 8(1):67–69.
9. Yant SR, Huang Y, Akache B, Kay MA (2007) Site-directed transposon integration in human cells. *Nucleic Acids Res* 35(7):e50.
10. Burnight ER, et al. (2012) A hyperactive transposase promotes persistent gene transfer of a piggyBac DNA transposon. *Mol Ther Nucleic Acids* 1:e50.
11. Pages H, Aboyoun P, Gentleman R, DebRoy S (2006) Biostrings: String objects representing biological sequences, and matching algorithms. Bioconductor. Version 2.22.0.
12. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10(3):R25.



```
Adineta      201 EERKRTDKFAVSREIWTDFSRKFKEMYNPGSHGTIDERLLGFRGKCPFRQY  251
Adineta_1    194 EERKKADKFAAIREIWLDFQDKLKTCYTPGLNITIDEQLLGFRGKCPFRQF  244
Anopheles    232 SQRLQTDKFALISDVFSRFVSNCQTNYVPGPHISVDEQLFPSKTRCPFTQF  282
Bombyx       250 DERKQTDNMAAFRSIFDQFVQCCQNAYSPSEFLTIDEMLLSFRGRCLFRVY  300
Ciona        203 AENIDNDKLYKVRPVYDLIVARWKALYNLGEHISIDEGMMKWRGRLGFRVY  253
Heliothis    224 SERLKTDKLAAVREFTDLMNNNFINNYCASENVTLDEQLPAFRGRFSGVVY  274
Takifugu     229 PARWQRDKLGVIRTVWDKWVRRLPLLYNPGPNVTIDEQLMPFRGRCPFLQY  279
piggyBat     203 -IVNESDRLCKVRPVLDYFVPKFINIYKPHQQLSLDEGIVPWRGRLFFRVY  252
Tni          233 PTLRENDVFTPVRKIWDLFIHQCIQNYTPGAHLTIDEQLLGFRGRCPFRMY  283

Adineta      252 IPSKPDKYAIKFWFCVDVNSYYIFDAFPYIERQPNEHRQ-RFVGPNVVLEL  301
Adineta_1    245 IPTKPDKYGLKFWLCVDAESYYVLNAFPYIGRQPGQEKQ-AHVGESVVLEL  294
Anopheles    283 MASKPDKYGQKYWMAVDVDSKYVVNIIPYLGKNDERPAE-ERLGDFVVKKL  332
Bombyx       301 IPNKPAKYGIKILALVDAKNFYVVNLEVYAGKQPSGPYAVSNRPFEVVERL  351
Ciona        254 NKDKPIKYGIKSYILADSHSHYCWNLDMYHRVQKT-------LKETVSQIL  297
Heliothis    275 MPNKPTKYGIKHYALVDSATFYLLKFEIYAGVQPEGPYRMPNDTVSLVKRM  325
Takifugu     280 LPSKPAKNGIKIWAACDATSSYAWNLQVYTGKPDGGAPE-KNPRNESCPRH  329
piggyBat     253 NAGKIVKYGILVRLLCESDTGYICNMEIYCGEGKR-------LL-ETIQTV  295
Tni          284 IPNKPSKYGIKILMMCDSGTKYMINGMPYLGRGTQT-NG-VPLGEYYVKEL  332

Adineta      302 MKPMYGSNRNVTIDNFFTSIHLAKELH--SGKLTLVGTIRKNKPEIPIEFQ  350
Adineta_1    295 LRPFYGSNRNVTKDNFFTSVPLARNLQ--TKNLTLIGTIRKNKPEIPIEFL  343
Anopheles    333 VDPYLNRGRNVTCDNFFTSLELAKFLK--SKKTSLVGTINKARREVPICVK  381
Bombyx       352 IQPVARSHRNVTFDNWFTGYELMLHLLN-EYRLTSVGTVRKNKRQIPESFI  401
Ciona        298 TSKCHFLWHSLYMDNFYNSVSMSQMLL--AFQIHSVGTLRSNRGE-PREIR  345
Heliothis    326 TEPIWGTGRNVTMDNWFTSVPLANILLK-DHQLTMVGTIRKNKPEIPTCFQ  375
Takifugu     330 VSGT--QWTQHHMRHFFTSHKLGQELL--KRKLTIVGTIRKNRSELPPQLL  376
piggyBat     296 VSPYTDSWYHIYMDNYYNSVANCEALM--KNKFRICGTIRKNR-GIPKDFQ  343
Tni          333 SKPVHGSCRNITCDNWFTSIPLAKNLLQEPYKLTIVGTVRSNKREIPEVLK  383

Adineta      351 SNKNRDVGSSIFGFS-DNLTLVSYVPKKNKAVILLSSMHHDS-----KV--  393
Adineta_1    344 SSKIREIGSSLFGFE-DNLALVSFVPKKNKAVLLLSSKHHDN-----HV--  386
Anopheles    382 KVKEKLYFTKAFK-S-DDTTLTVYQGKTKKNVVLLSSMHRDI-----RT--  423
Bombyx       402 RTD-RQPNSSVFGFQ-KDITLVSYAPKKNKVVVVMSTMHHDN-----SI--  443
Ciona        346 TPPNQMKKGDIIARQNQSVTVLAW--KDKRVVKAISTKH-DA-----SVTT  388
Heliothis    376 PKRTRTEHSSLFGFQ-EDVTLCSYVPKKSKAVLLISSMHNDN-----NI--  418
Takifugu     377 TSKNRPVKSSQFAYT-ADTSLVSYVPKKGKNVVLMSTLHRDG-----RM--  419
piggyBat     344 TISLK-KGETKFIRK-NDILLQVWQ--SKKPVYLISSIHSAEMEESQNI--  388
Tni          384 NSRSRPVGTSMFCFD-GPLTLVSYKPKPAKMVYLLSSCDEDA-----SI--  426

Adineta      394 ----------DI--GTGKPNIVLDYNKSKGAVDTIDEMCHKYSVKRGTRRW  432
Adineta_1    387 ----------DN--KTGKPVIILDYNKTKGAVDTVDQMCHKYTVKRGTKRW  425
Anopheles    424 ---------GND--KKSKPETVAFYNSTKYGVDVVDQMCRKYSLKSASRRW  463
Bombyx       444 ---------DESTGEKQKPEMITFYNSTKAGVDVVDELCANYNVSRNSKRW  485
Ciona        389 ITRRQRRGGEXE--SVEKPVCIADYNLHMSGVDQVDQMISYYPCHRKSLKW  437
Heliothis    419 ---------VES--EKKKPEIILYYNSTKGGVDTNDQMCANYNVGRRTKRW  458
Takifugu     420 ---------CDQ--EHHKPEIIMDYNATKGGVDNMDKLVTAYSCKRRTLRW  459
piggyBat     389 ------DRTSKK--KIVKPNALIDYNKHMKGVDRADQYLSYYSILRRTVKW  431
Tni          427 ----------NE--STGKPQMVMYYNQTKGGVDTLDQMCSVMTCSRKTNRW  465
```

Fig. S1. Protein sequence alignment of *piggyBac* family members. The catalytic domain of eight *piggyBac* transposase family members were aligned to *Trichoplusia ni* (Tni) (1, 2). Blue boxes indicate the requisite catalytic amino acids, red boxes indicate conserved argenines and lysines, and green boxes indicate the positions of HIV integrase mutations with known altered target joining in HIV integrase (3).

1. Mitra R, Fain-Thornton J, Craig NL (2008) piggyBac can bypass DNA synthesis during cut and paste transposition. *EMBO J* 27(7):1097–1109.
2. Mitra R, et al. (2013) Functional characterization of piggyBat from the bat Myotis lucifugus unveils an active mammalian DNA transposon. *Proc Natl Acad Sci USA* 110(1):234–239.
3. Harper AL, Skinner LM, Sudol M, Katzman M (2001) Use of patient-derived human immunodeficiency virus type 1 integrases to identify a protein residue that affects target site selection. *J Virol* 75(16):7756–7762.

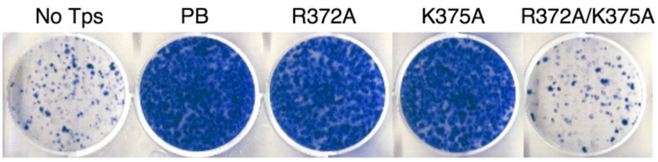


Fig. S2. Colony formation assay with PB^{R372A} and PB^{K375A} individual mutations. HeLa cells were transiently cotransfected with a transposon expressing blasticidin^{r} and the indicated mutant *piggyBac* (PB) transposase. Cells were selected for blasticidin resistance and stained with methylene blue to identify viable cell colonies. No transposase (–Tps), wild-type transposase (PB), and PB$^{R372A/K375A}$ transposase were included as controls.

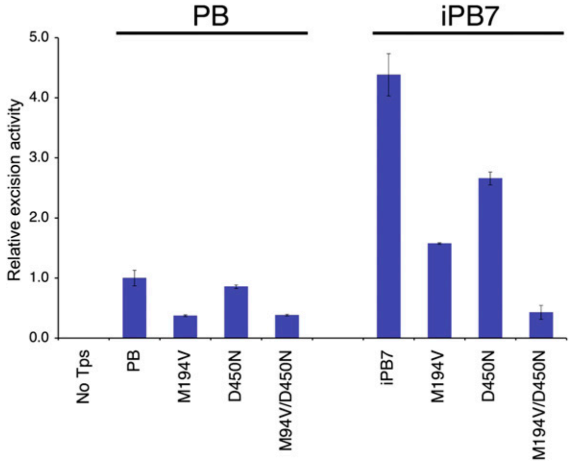


Fig. S3. Excision assays of individual M194V and D450N mutations on the PB or iPB7 transposase backbones. The indicated wild-type PB (*Left*) or insect-derived *piggyBac* transposase 7 (iPB7) (*Right*) mutants were transiently transfected into HEK293 *GFP::PB* cells. The frequency of excision is indicated by GFP fluorescence intensity, determined by FACS analysis, and normalized to the wild-type PB control. No transposase (−Tps) and unmodified iPB7 were included as additional controls.

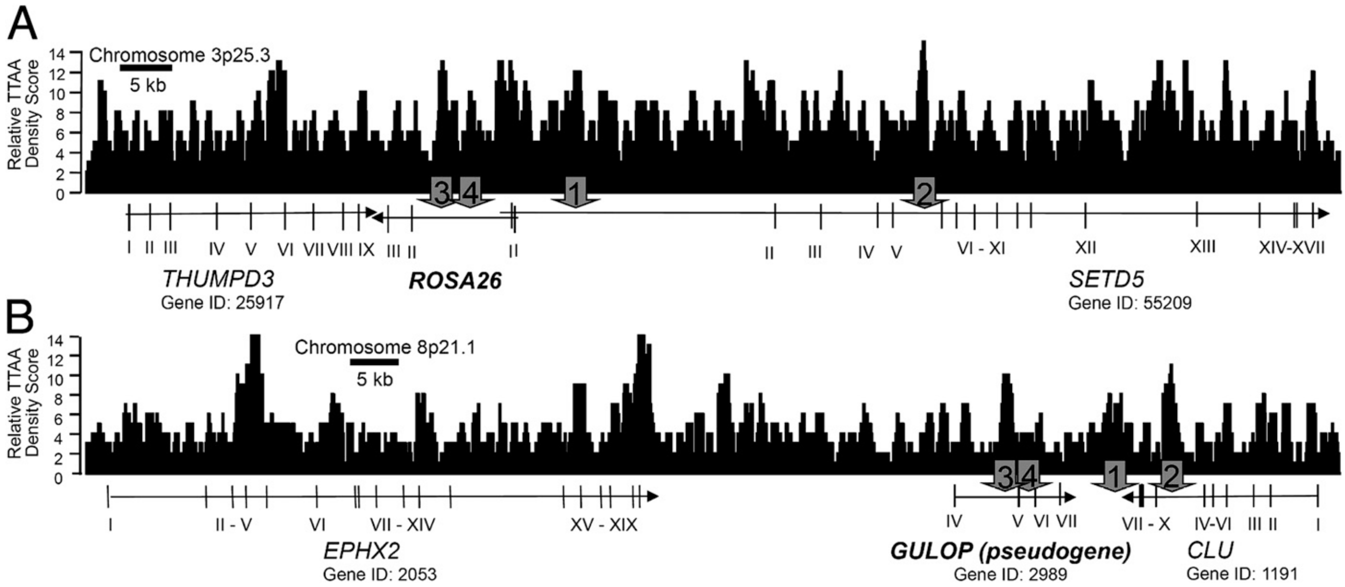


Fig. S4. Schematic representation of ZFP target genomic loci. (*A*) The human *ROSA26* locus is flanked by the *THUMPD3* and *SETD5* genes on chromosome 3p25.3. (*B*) The *GULOP* pseudogene is flanked by *EPHX2* and *CLU* in chromosome 8p21.1. Gray arrows indicate approximate ZFP target sites. The TTAA density score was determined for a given 256-bp window.

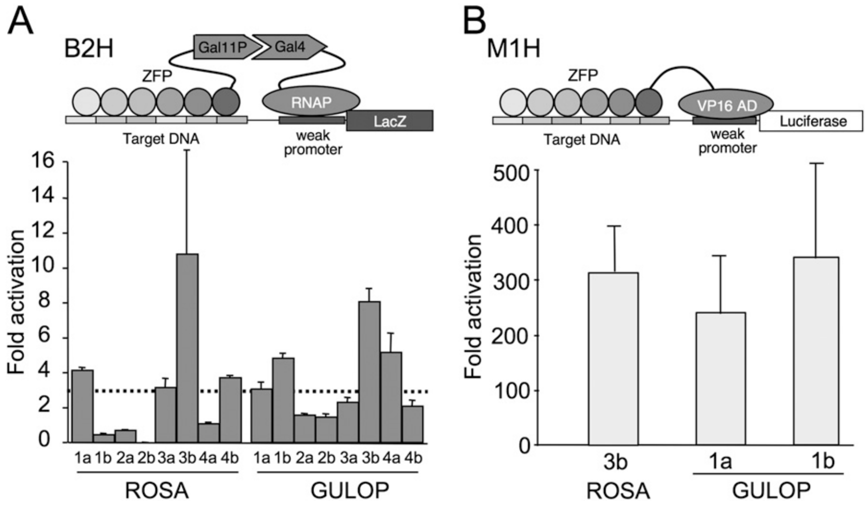


Fig. S5. Engineered ZFP activity in cells. (*A*) A bacterial two-hybrid (B2H) assay was used to assay activity of engineered ZFPs. A six-finger ZFP is fused to the Gal11P fragment, shown schematically. ZFP binding to its target recruits RNA polymerase to a weak promoter driving the reporter *lacZ* gene in bacteria through interaction of the *GAL4* domain fused to the RNAP. Eight ZFPs targeting four sites at or near the *ROSA26* locus or eight ZFPs targeting four sites at or near the *GULOP* locus were evaluated. Bars represent LacZ activity in bacteria transformed with the ZFP library. The dashed line represents an arbitrary threshold at which B2H activity is typically effective in mammalian cells. (*B*) For mammalian one-hybrid (M1H) assays, activator plasmids expressing the ZFPs fused to the VP16 activation domain from Herpes Simplex Virus 1 were cotransfected with plasmids containing four copies of the target sequence upstream of a minimal promoter driving firefly luciferase, shown schematically. ZFP binding to its target sequence activates luciferase transcription. ZFP activity is reported as a function of luciferase activity. Bars represent mean fold activation in cells transfected with activator and reporter plasmids relative to luciferase activity in cells transfected with reporter alone. $n = 3$.

GULOP

```
ATGTCTAGACCAGGAGAGCGACCATTCCAGTGCCGGATTTGCATGCGCAATTTTTCCAG
ACAGGCCAACCTCGTCAGACACACGAGGACACATACTGGTGAGAAGCCCTTCCAGTGTC
GCATCTGTATGCGCAATTTTTCAGTGGCGCATAATCTGACTAGGCATCTCAGGACTCAC
ACTGGGGGAGGAGGCTCCCAGAAGCCTTTTCAGTGCAGGATCTGTATGAGAAATTTTTC
AGATTCCTCTGTGCTGAGGAGGCACCTCAGGACGCATACCGGAGAAAAACCATTCCAGT
GTAGAATTTGCATGAGAAACTTTAGTCAAGGCGGGACCCTTAGGAGGCACTTGAAAACA
CATACAGGCTCCCAGAAGCCATTTCAGTGCCGCATCTGTATGCGCAACTTTTCTGTGCA
CCATAACCTCGTGAGACATCTGAGGACTCACACTGGAGAGAAGCCATTTCAGTGTAGGA
TTTGCATGAGGAATTTTAGTAGGTCCGACCATCTGAGTCTTCACCTGAAGACACATCTG
CGG
```

ROSA26

```
ATGTCTAGACCTGGCGAACGCCCATTTCAGTGCCGCATTTGCATGAGAAATTTCAGCCT
TAAGCATTCTCTGCTTCGCCACACGCGGACCCACACCGGAGAGAAGCCCTTCCAGTGCC
GGATTTGTATGCGAAATTTTTCTCTGCGCCACAATCTTAGGAGGCACTTGCGAACTCAC
ACCGGCAGCCAGAAACCTTTCCAGTGTCGAATCTGCATGCGCAATTTTAGTCGCAGAGC
ACATCTCTTGAGCCACCTGCGAACGCATACCGGCGAGAAGCCCTTCCAGTGCAGGATCT
GCATGCGGAACTTCAGCGAGGCACATCACCTGTCTCGCCATCTGAAGACCCATACAGGC
GGTGGAGGTAGTCAAAAGCCGTTTCAGTGCAGGATTTGTATGAGGAATTTCAGTGATAG
TCCAACACTTCGGCGACACCTGCGCACTCACACAGGCGAGAAGCCGTTCCAGTGCAGGA
TCTGCATGAGAAATTTTTCCGTAAGACACAATCTCACGCGGCACCTTAAAACACACCTG
AGA
```

Fig. S6. Sequences of the GULOP and ROSA26 ZFPs. The primary sequence of the DNAs encoding the GULOP or ROSA26 ZFPs are shown.

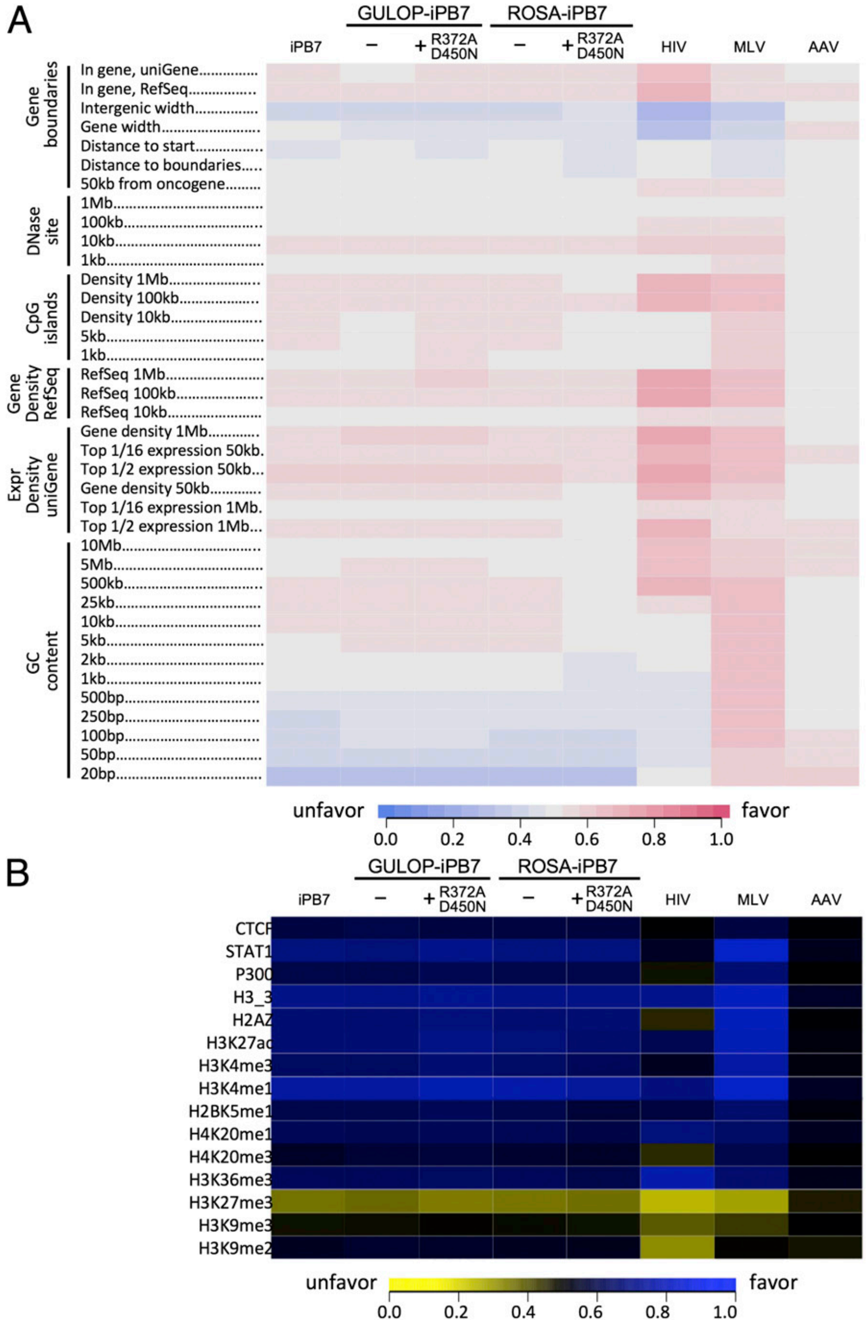


Fig. S7. Distribution of iPB7, GULOP-iPB7, GULOP-iPB7$^{R372A/D450N}$, ROSA26-iPB7, and ROSA26-iPB7$^{R372A/D450N}$–mediated insertions in the human genome. Integration-site datasets for ZFP–iPB7-mediated insertions are indicated by the columns, and genomic features or ChIP-Seq datasets are indicated by the rows (the latter were calculated over 10-kb windows). The departure from random distribution is indicated by colored tiles (key at bottom), and differences from random placement were scored using the Receiver Operator Characteristic (ROC) area method described previously (1). A detailed explanation of the variables studied can be found in Ocwieja et al. (2) or at http://microb230.med.upenn.edu/assets/doc/HeatMapGuide_v12_formatted.doc. (*A*) The integration frequency relative to selected genomic features is shown. Red shading indicates features where insertions are favored compared with random, whereas blue shading indicates unfavored integration events. Gray indicates random distribution. The distribution of HIV-, MLV-, and Adeno Associated Virus-mediated integrations are shown for comparison. (*B*) The integration frequency relative to bound proteins and modified histones was mapped using the ChIP-Seq method. Yellow and blue are used to indicate depletion or enrichment, respectively.

1. Berry C, Hannenhalli S, Leipzig J, Bushman FD (2006) Selection of target sites for mobile DNA integration in the human genome. *PLOS Comput Biol* 2(11):e157.
2. Ocwieja KE, et al. (2011) HIV integration targeting: A pathway involving Transportin-3 and the nuclear pore protein RanBP2. *PLoS Pathog* 7(3):e1001313.

Table S1. Imprecise excision by excision competent^{hyper}/integration defective mutant transposases

Transposase	Imprecise repair, %
PB	0.13*
PB$^{R372A/K375A}$	0.14
PB$^{M194V/R372A/K375A}$	0.21
PB$^{R372A/K375A/D450N}$	0.27
iPB7	0.24*

*Imprecise excision frequencies were determined as described in *Materials and Methods*. The imprecise excision frequency of Int^{+} transposases is underestimated by 40–60% because imprecise excisions that are accompanied by transposon reintegrations are not counted.

Table S2. Illumina sequencing of ZFP–iPB7-mediated genomic integrations

Element	Reads	Alignments	Initial sites	Collapsed sites	TTAA sites
iPB7	26,300,573	3,011,317	45,523	43,984	40,800
GLO-iPB7	79,825,963	9,897,552	61,914	58,900	54,803
GLO-iPB7$^{R372A/D450N}$	94,236,814	11,829,337	74,210	70,393	66,379
ROSA-iPB7	85,295,144	10,691,398	62,795	59,609	55,924
ROSA-iPB7$^{R372A/D450N}$	49,842,350	6,964,687	41,599	39,659	37,033

The number of total mapped integration reads and unique alignments for each ZFP–iPB7 chimera and unmodified iPB7 control are indicated and were determined as described in *Materials and Methods* and *SI Materials and Methods*. Collapsed sites, TTAA+ non-TTAA insertion sites; TTAA sites, only TTAA insertion sites.